Petra Perner (Ed.)

Advances in Data Mining

Applications and Theoretical Aspects

20th Industrial Conference, ICDM 2020
Amsterdam, The Netherlands
July 20 – July 21 2020
Proceedings

ibai Publishing
www.ibai-publishing.org

Volume Editor

Petra Perner
FutureLab Artificial Intelligence_IBaI_II
Institute of Computer Vision and Applied Computer Sciences,
IBaI PF 30 11 14
04251 Leipzig
E-mail: pperner@ibai-institut.de

P-ISSN 1864-9734
E-ISSN 2699-5220
ISBN 978-3-942952-76-7

The German National Library listed this publication in the German National Bibliography.
Detailed bibliographical data can be downloaded from http://dnb.ddb.de.

ibai-publishing
Prof. Dr. Petra Perner
PF 30 11 38
04251 Leipzig, Germany
E-mail: info@ibai-publishing.org
http://www.ibai-publishing.org

P-ISSN 1864-9734
E-ISSN 2699-5220
ISBN 978-3-942952-76-7

20th Industrial Conference on Data Mining ICDM 2020

www.data-mining-forum.de
July 20 – 21, 2020
Amsterdam, The Netherlands

Chair

Prof. Dr. Petra Perner
Institute of Computer Vision and applied Computer Sciences, IBaI

Program Committee

Ajith Abraham	Machine Intelligence Research Labs (MIR Labs), USA
Mohamed, Bourguessa	Universite du Quebec a Montreal - UQAM, Canada
Bernard Chen	University of Central Arkansas, USA
Jeroen de Bruin	University of Applied Sciences JOANNEUM, Austria
Antonio Dourado	University of Coimbra, Portugal
Stefano Ferilli	University of Bari, Italy
Geert Gins	Glaxo Smith Kline, Belgium
Warwick Graco	Australian Tax Office ATO, Australia
Aleksandra Gruca	Silesian University of Technology, Poland
Pedro Isaias	The University of Queensland, Australia
Piotr Jedrzejowicz	Gdynia Maritime University, Poland
Martti Juhola	University of Tampere, Finland
Janusz Kacprzyk	Polish Academy of Sciences, Poland
Mehmed Kantardzic	University of Louisville, USA
Lui Xiaobing	Google Inc., USA
Eduardo F. Morales	National Institute of Astrophysics, Optics, and Electronics, Mexico
Samuel Noriega	Universitat de Barcelona, Spain
Wieslaw Paja	University of Rzeszow, Poland
Juliane Perner	Novartis Institutes for BioMedical Research (NIBR), Switzerland
Rainer Schmidt	University of Rostock, Germany
Moti Schneider	PCCW Global, Greece
Victor Sheng	University of Central Arkansas, USA
Kaoru Shimada	Fukuoka Dental College, Japan
Gero Szepannek	University of Applied Sciences Stralsund, Germany
Joao Miguel Costa Sousa	Technical University of Lisbon, Portugal
Markus Vattulainen	Tampere University, Finnland
Zhu Bing	Sichuan University, China

Preface

The pandemic "Corona" has put us this year before a difficult time. With care we have kept to the hygiene rules not to get an infection with the virus Covid-19. With mask we have got into coaches and trains, have made our purchases or on work worked. Home office was the catchword of these days. The universities and research facilities have maintained only a small emergency company and lectures were held as online lectures. From home we have tried to do our scientific works. In 1-to-1 telephone calls or phone conferences we have organized with our colleagues the work and have discussed important results of the research. Under it the efficiency suffers what is easy to understand.

In the beginning of the pandemic fell the deadline of our conference. Insecurity spread. The figures of the infected persons increased rapidly. The virus spreads out in more and more countries and was further carried by continent to continent. Soon stood the whole world in the spell of Corona. A conference was the last to this in this situation most thought.

In this situation appeared once again which high demands for a scientist are made. It belongs to the job of a scientist that he presents his scientific results in conferences and makes thus his results of a wide public immediately available. A scientist should have well organized his research, should be able to do his scientific tasks and duties in a flexible way, and should have financed his research with suitable financial means. Only those who were meeting these rules could successfully continue in their professional research work.

The best of the best of us are represented with their papers in this volume. They presented themselves personal or in online presentations in the conference. The acceptance rate for the submitted paper of our conference was 33% percent for long paper as well as short papers. Because of many refusals because of missing financial means or other reasons the acceptance rate decreased to few percent. This shows once more the excellent quality of these scientists. Their papers are of most excellent quality and expand the state-of-the-art in an excellent way. The topics of the long papers range from event log file analysis, predictive maintenance, medical application, telecom application, fraud detection to a paper on how we should present the results to stakeholders so that they accept the findings of the data mining methods. The new arising topic we see here is predictive maintenance. All other topics follow the main topics of ICDM but present new excellent results and go over the recent questions to be solved with data mining for the specific applications. The short paper is a fine theoretical paper on optimal kernel density estimation.

The proceedings will be freely accessible as an OPEN-ACCESS Proceedings of a wide public so that, the new acquired knowledge on the different subjects is able to spread around quickly worldwide. You can find the proceedings for long papers and the poster proceedings for short papers at http://www.ibai-publishing.org/html/proceeding2020.php.

In this time, flexibility was a must Because the situation in the USA was still difficult, we have moved the conference to Amsterdam in the Netherlands. Here a variety

of the participants was able to do outward journeys. The ones who could not travel, were online present.

Extended versions of selected papers will appear in the international journal Transactions on Machine Learning and Data Mining (www.ibai-publishing.org/journal/mldm).

We hope to see you in 2021 in New York at the 21th Industrial Conference on Data Mining ICDM (www.data-mining-forum.de) again.

The conference runs under the umbrella of the World Congress on "The Frontiers in Intelligent Data and Signal Analysis, DSA 2021" (www.worldcongressdsa.com), which combines under its roof the following three events: International Conferences Machine Learning and Data Mining MLDM (www.mldm.de) , the Industrial Conference on Data Mining ICDM (www.data-mining-forum.de), and the International Conference on Mass Data Analysis of Signals and Images in Artificial Intelligence and Pattern Recognition with Applications in Medicine, Biotechnology, Chemistry and Food Industry, MDA-AI&PR (www.mda-signals.de).

We will give then the tutorials on Data Mining, Case-Based Reasoning, and Intelligent Image Analysis again (http://www.data-mining-forum.de/tutorials.php) again. The workshops running in connection with ICDM will also be given (http://www.data-mining-forum.de/workshops.php).

We would warmly invite you with pleasure to contribute to this conference. Please come and join us. We are awaiting you.

July, 2020 Petra Perner

Table of Content

Unified Expression Ripple Down Rules based Fraud Detection Technique for Scalable Data

Ikram Ul Haq[1], Iqbal Gondal[1], Peter Vamplew[1]

[1]ICSL, School of Science, Engineering and Information Technology, Australia
PO Box 663, Ballarat 3353, Victoria
ikramulhaq@students.federation.edu.au
{iqbal.gondal, p.vamplew}@federation.edu.au

Abstract. Fraud detection for online banking is an important research area and higher accuracy is highly desirable. The main challenges in fraud analysis are due to the presence of heterogeneous transactions data, large and distributed data. Among existing rule-based techniques for fraud detection, Ripple Down Rules (RDR) is ideal due to its less maintenance and incremental learning. However, banking data sets contains billions of transactions, so the performance of RDR on distributed and Big data platforms need to be studied for fraud detection applications. A Unified Expression RDR fraud deduction technique for Big data has been proposed and evaluated in this paper. By incorporating the Unified Expressions into the RDR and evaluating the expressions using the Lift score, the compactness of the ruleset can be achieved and the accuracy of the classification improved. In addition, the paper presents a compact model that fuses Majority and Minority classes for RDR-based classifiers. Classification accuracy is compared with the two existing RDR implementations RIDOR and Integrated Prudence Analysis technique and a non-RDR classifier as well. Empirical evaluations on various datasets have shown that not only the ruleset size of training and prediction dataset is reduced, but the accuracy of classification is also improved. The results showed an improvement in the classification accuracy when compared to two RDR and non-RDR based classifiers. The proposed technique is used for experimental validation and the development of fraud analysis, but it can also be used in other domains, in particular for scalable and distributed systems.

Keywords: Classification, Fraud Detection, Spark, MapReduce, Hadoop, Ruleset, RDR, Naïve Bayes, RIDOR, IPA, Unified Expressions.

1 Introduction

Fraud detection for online banking is vital as frauds can affect the core business of the financial industry in terms of loss of confidence of the public in the industry. Online banking frauds are resulting in billions of dollars of loss to banks around the world

[1]. As per the Microsoft Computing Safety Index survey (2014), the annual global-impact of phishing and various forms of identity theft is about US$5 billion. Internet Crime Complaint Centre has reported a 161% increase in the loses in 2018 [2].

Various fraud detection techniques have been developed over the last decade. In view of the importance of fraud detection in the banking sector, higher accuracy of fraud detection techniques is critical. One of the major challenges faced by fraud analysis research is the heterogeneous nature of transactions [3]. Typically, datasets can have both numeric and alphabetical attributes, but numeric data is known to provide better performance for machine learning algorithms. Large-scale data in online banking also requires algorithms to show better performance with scalable and distributed data. In [4, 5] authors highlight that Apache Spark is a popular open-source platform for large scale data processing and iterative machine learning tasks.

Section 2 describes the background of UE-RDR methodology and previous work. Section 3 is the methodology of proposed technique, while experimental setup is explained in section 4. Section 5 shows the results and section 6 concludes the work.

2 Prior Work on Fraud Detection Using Machine Learning and Background to UE-RDR Methodology

Kou et al. [6] believe that fraud detection research mostly uses data mining, statistics, and artificial intelligence; and fraud is identified from anomalies in data and patterns. Phua et al. [7] have surveyed fraud detection research to categorize the research using four main approaches including supervised, hybrid, semi-supervised and unsupervised and; also identified the relationship of fraud detection with other domains. Melo-Acosta et al. [8] have presented a credit card fraud detection technique using Big data, but their technique is more specific to imbalance and unlabelled data.

In [9], authors presented a fraud detection approach for Medicare fraud using three medicare and medicaid services datasets. They use the combined dataset for training with three learning methods: Random Forest, Gradient Tree Boosting and Logistic Regression models and used the Area Under the ROC Curve metric to measure the performance of fraud detection. They claim that best fraud detection performance is with the use of the combined dataset. Dataset size is not mentioned, but this technique is not ideal for large datasets, e.g. Synthetic data generation based on original seed datasets.

Integrated Prudence Analysis (IPA) is developed by [10] which uses prudence analysis in Ripple Down Rules (RDR) and has combined two of the previously developed Multiple Classification RDR (RM) and Ripple Down Models (RDM) [11, 12] techniques. A fundamental difference in these techniques is that RM is structural while RDM is attribute-based. The difference in these methods is well explained by [13]. IPA is a multi-class labels classifier. RDR is one of the well-known rule-based classification technique and was developed as an alternative to the traditional knowledge-based system [11, 14]. Maruatona [10] acknowledges that RDR is ideal due to its less maintenance and incremental learning capabilities. RDR significantly

reduces the time and effort required to make the alteration and ensure the consistency of the rulesets. Authors in [11, 15-17] have highlighted that RDR systems have been used in many applications and classification domains. RIDOR is an RDR implementation in WEKA and [18] also acknowledges that RIDOR is most widely used RDR machine learner. Below table shows an Iris ruleset generated from RIDOR.

Table 1. Iris RIDOR ruleset.

RIDOR Rule
class = setosa (150.0/100.0) Except (petal_len > 2.45) => class = virginica (66.0/0.0) [34.0/0.0] Except (petal_len <= 4.95) and (petal_wid <= 1.55) => class = versicolor (29.0/0.0) [16.0/0.0] Except (petal_wid <= 1.75) => class = versicolor (8.0/5.0) [1.0/0.0]

One of RDR implementation is RIDOR, which also has MapReduce [19] based implementation in WEKA for Apache Hadoop [19] wrapper, which can be used for the classification of large data. However, [4, 20] highlight that Spark is better as compared to conventional MapReduce. Spark retains the linear scalability and fault tolerance of MapReduce and is nearly 100 times more efficient than MapReduce. Mahout is another machine learning platform for Big data. [4] highlights that Mahout is also based on MapReduce and they observed that Spark's performance and scalability are better than Mahout.

Unified Expressions Language (UEL) is capable of evaluating a number of additional operators that are missing in RDR expressions. Unified Expressions (UE) can also replace existing operators with more efficient operators of IN and LIKE. Using UE, we can prepare compressed rule with a revised Lift [21, 22] score (explained with more detail in section 3.5) which is the ratio of target response divided by the average response. UEL supports contextual expressions and can also retrieve geocoding and demographics information from fraud datasets [23], that help to filter suspected cases. UE application in the proposed technique is explained in section 3.6. UE can offer a variety of operators that can help with the compactness of ruleset and evaluation of the expression based on Lift score. Furthermore, the UE can help in choosing the best rules with higher confidence; therefore, the more accurate class label is chosen, which improves accuracy. UE-RDR is implemented on Big data Spark platform by overcoming the limitation of mixed datasets. Apache Spark performance is known to be better than conventional Apache Hadoop MapReduce [4, 20] so UE-RDR on Spark will be more efficient than RDR MapReduce based implementation in WEKA and will also have iterative machine learning capability.

UE-RDR fraud detection technique for large scale mixed data has been developed and evaluated in this paper to improve detection accuracy and reduce computation costs. The technique has three models: the minority (UE-RDR-MIN) class, the majority (UE-RDR-MAJ) class-based models and combined model (UE-RDR-MIX). The combined and distinct rules in UE-RDR-MIX model gives better accuracy than

the other two models. UE-RDR-MIX is an innovative model and to the best of our knowledge, no study has been on in RDR based classifiers. UE-RDR performance is compared with RDR. The proposed technique is applied to various data datasets (Table 5), including Synthetic Bank datasets and three publicly available datasets from the UCI machine learning repository. Performance is evaluated and compared with two RDR based implementations (RIDOR and IPA) and a non-RDR classifier (Naïve Bayes [24] as well. The empirical evaluation has shown that the model's performance in terms of classification accuracy and ruleset size is better than RIDOR. Classification accuracy with UE-RDR-MIX is better than IPA and Naïve Bayes classifiers.

Considering these shortcomings, main contributions of the paper are listed below:

- UE implementation for RDR and development of a threshold-based approach for ruleset compression with the use of Lift score.
- Development of a single classification Unified Expressions RDR (UE-RDR) technique with three UE-RDR sub-models. UE-RDR-MIX is an innovative model, which makes use of majority and minority classes and multi-level compactness.
- Implementation of the developed technique on distributed and Big data machine learning platform, Spark.

With these contributions, we have proposed an innovative technique for fraud detection for large scale data and with rule-based classifiers using a supervised approach on labelled datasets. The developed technique can be used on mixed datasets. The developed algorithm is implemented on big and distributed data platform Spark and has shown better accuracy as compared with two of the existing RDR based classifiers and a non-RDR classifier. UE-RDR can process huge datasets, but upto 100,000 instances of the dataset were used in the evaluations.

3 Methodology

Knowledge-based systems are a major application for concept descriptions. Littin [25] mentions that rules and decision-trees are two of the common forms of concept descriptions in machine learning. Maruatona [10] indicates that banks and financial institutions use rule-based approaches in their Internet banking fraud detection systems.

3.1 UE-RDR Models

Fraud detection data is a single classification data and UE-RDR is also a single classification model, with UE based on RDR. In UE-RDR technique, three models are developed, UE-RDR-MIN, UE-RDR-MAJ and UE-RDR-MIX. Littin [25] highlights that the inclusion of RDR top-level empty rule is used generally with a default class. Gaines et al. [26] have used the class that occurs most frequently (Majority) as default in the training data, however in RIDOR the least frequently used class (Minority) is used as default class. UE-RDR technique is also graphically illustrated as a multi-step

process in Figure 1. Table 2 shows Iris ruleset for a UE-RDR-MIN model. But a typical ruleset and a rule structure of UE-RDR model is shown below.

{"defaultclass":"CLASS-LABEL", "model":"MODEL-NAME","count":TOTAL-POPULATION,"rulses":[RULES-COLLECTION]

RULE#{"number":#,"isParent":true,"level":#,"description":"UE-EXPRESSION", "lift":#,"cover":#,"ok":# "class":"CLASS-LABEL","parentid":#, "childrenNodes":#}

UE-RDR-MIN

In this model, least frequently occurring (Minority) class is the default class (like RIDOR), and the rules are for the remaining class labels. i.e. majority class label and other classes. In most of the cases ruleset set for this model is supposed to be larger than the ruleset for UE-RDR-MAJ, as least frequently used class is default class and rules are for the remaining class labels (including majority class).

UE-RDR-MAJ

In this model, most frequently occurring (Majority) class is the default class (as used by [26]), and the rules are for the remaining classes. In terms of ruleset size, this model would have a similar size of ruleset as UE-RDR-MIN model.

UE-RDR-MIX

This model is a union of the rules for the minority & majority class models and distinct rules for the remaining class-labels. Rules expressions are further compressed with revised Lift score outlined in sections 3.5 and 3.7. This model is our innovation and does not exist in RIDOR implementation. Algorithms 2a explains this model. In RDR ruleset, one class is the default class and ruleset contain rules for the remaining class labels. We claim that this model gives the best classification accuracy, as shown in Figure 3. Unlike RDR, it contains rules for all class-labels instead of using a default-class. In terms of ruleset compactness, Figure 4 shows that for some datasets, UE-RDR-MIN and UE-RDR-MAJ have good performance as well.

If there are more than two class labels in a dataset, this model also provides better accuracy for class labels that belong to neither majority nor minority classes. Considering Bank dataset (Dataset 1 & 2) example, there are three class labels: Fraud, Anon and None, where Anon as anonymous and Non as not a fraud. In this dataset Fraud class label does not fall into the majority or minority class, so UE-RDR-MIX model will give better accuracy for Fraud class labels in this dataset. Apart from the overall higher classification accuracy, classification accuracy is also sometimes important for a specific class label. For example, Fraud cases are more important for improved accuracy in the Bank dataset. A wrong prediction of a Fraud case would result in a greater loss compared to the mistake of None or Anon cases. Accuracy results from the confusion matrix are shown in Figure 6.

3.2 Algorithms

The developed technique is based on three algorithms. UE-RDR ruleset construction is explained in Algorithm-1, while ruleset compactness is explained in Algorithm-2 and prediction flow with Spark is explained in Algorithm-3. Algorithm-2a is for UE-RDR-MIX model only, which is further compactness of Majority and Minority class models (UE-RDR-MIN and UE-RDR-MAJ). Figure 1 illustrates UE-RDR process flow and glues three algorithms to demonstrate the three-stages. In Algorithm-3, when a data file is stored in Hadoop [19] Distributed File System (HDFS), the system breaks it down into individual blocks set and stores these blocks in multiple worker-nodes in the cluster. Rows division in each data block can be determined with Eq. (1).

$$Rows^{Block} = \Sigma Rows/SparkNodes/BlockSize/RowDataSize \quad (1)$$

The mentioned algorithms are given below:

ALGORITHM 1: Building Training Model

Input: Ruleset from a RIDOR.
Output: Training model for a UE-RDR.
Begin
1. Process RIDOR ruleset.
2. Process each expression in the ruleset.
3. Get Ok and Cover values of each expression.
4. Calculate Lift score of the expression from Ok and Cover values using Eq. (4).
5. Prepare the expression in UE format using funcUEL Eq. (5).
6. Convert the expression in JSON format with attributes (See Table 2).
7. IF (more expressions in the ruleset) Goto step-2

ELSE FINISH
End

ALGORITHM 2: Compactness

Input: Training model for a UE-RDR.
Output: Compact UE-RDR Training model.
Begin
1. Process each rule in the ruleset of the training model.
2. Traverse Ruleset & Get Lift score of the rule
2.1. Find the merging rule (using the custom thresholds approach listed in Table 4).
2.2. Merge UE rule.
3. Traverse rule to compact UE (See UE operators Table 4)
3.1. Calculate and update the revised Lift score, from updated Ok and Cover values of merging rule – see Eq. 4.
3.1 Update UE rule.
3.2 IF (more expressions to process) Goto step-3
3.3 Process all expressions from complete rule from Step 3 – 3.2
4 IF (more rules) Goto Step-1 ELSE FINISH
End

ALGORITHM 2a: UE-RDR-MIX Compactness

Input: Training model for a UE-RDR-MIN and UE-RDR-MAJ.
Output: Compact UE-RDR Training model for UE-RDR-MIX.
Begin
1. Repeat Algorithm-2 with the input of two UE-RDR Training Models.
2. Repeat Steps 1 to 3.2 from Algorithm-2.
End

ALGORITHM 3: Prediction Process

Input: Training model from UE-RDR and dataset. **Output**: Accuracy for dataset.
Begin
1. Load Dataset
1.1. Process each instance.
1.2. Transform instance to RDD double Vector, including categorical attributes using funcTransRDD Eq. (2).
1.3. Split data on Spark nodes based on the data block size using Eq. (1)
2. Load UE-RDR training model.
3. Load RDD vector collection from data locality.
3.1. Process each rule from the Training Model.
3.2. Transform categorical attributes in expression with funcTransCat function (3).
3.3. Evaluate UE rule expression and pick the predicted class.
3.4. If multiple rules are true, then pick predicted class of better Lift score rule.
3.5. IF (more rules in the ruleset) Goto step-3.1
IF (more instances to process) Goto step-3 ELSE FINISH
End

UE-RDR Process Flow

Figure 1 links three algorithms to illustrate the flow of the three-step algorithms. The dependency in each step and the main and the sub-tasks in each step are clarified there. Loading and Prediction are the two steps in the Prediction process.

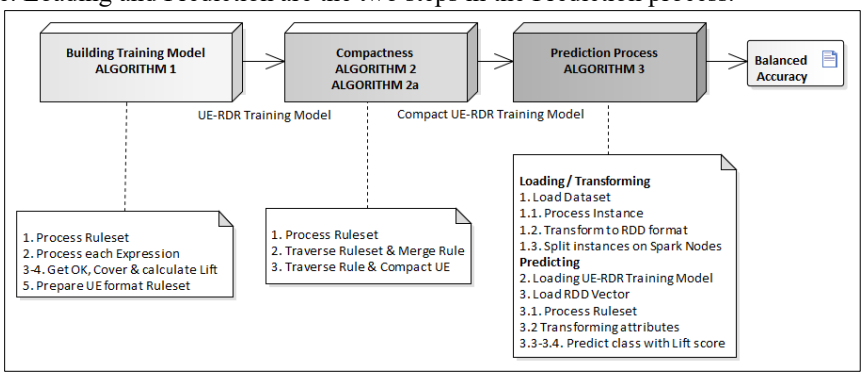


Fig. 1. UE-RDR process flow

3.3 Transformations

Due to the large datasets, the developed technique was implemented on Spark. The core of Spark is a concept called the Resilient Distributed Dataset (RDD), which is a collection of records. The default data-format for Spark platform is numeric; however the Bank dataset and many real-life datasets contain mixed attributes. Two transformation functions were developed, which are explained below. The function in Eq. (2) transforms mixed data to numeric RDD format at loading time.

$$\text{Transformation}^{\text{RDD}}=\text{funcTransRDD}\int_{i}^{n_Y} att \neq numeric \quad (2)$$

where $\text{Transformation}^{\text{RDD}}$ is the RDD format and funcTransRDD is a function to convert a row y with only categorical attributes from 1 to n on ith index. While function Eq. (3) transforms the categorical value of the attribute to numerical value at the expression evaluation time.

$$\text{Transformation}^{\text{CAT}} = \text{funcTransCat}\int_{i}^{n_Y}(att \text{ in exp}) \quad (3)$$

where $\text{Transformation}^{\text{CAT}}$ is the RDD format and funcTransCat is a function to convert a row y with only categorical attributes from 1 to n on the ith index and which exist in an expression. These transformations are necessary in order to evaluate expressions from the original ruleset.

3.4 UE-RDR Ruleset

Table below shows an iris ruleset generated from UE-RDR.

Table 2. Iris UE-RDR ruleset.

UE-RDR Rule
{"defaultclass":"setosa", "model":"UE-RDR-MIN","count":3, "rules":[{"number":1,"isParent":true,"level":1,"description":"(petal_len > 2.45)","lift":1.5, "cover":100.0, "ok":100.0, "class":"virginica","parentid":0, "childrenNodes":2}, {"number":2,"isParent":false,"level":2,"description":"(petal_len > 2.45) && (pet-al_len <= 4.95) && (petal_wid <= 1.55)","lift":3.333333, "cover":45.0, "ok":45.0,"class":"versicolor","parentid":1,"isChild":true}, {"number":3,"isParent":false,"level":2,"description":"(petal_len > 2.45) && (pet-al_wid <= 1.75)", "lift":7.4074, "cover":9.0,"ok":4.0,"class":"versicolor", "parentid":1}]}

where “Cover” is the number of instances a rule expression correctly identifies and “Ok” is how many instances (out of the Cover) are correctly classified by this rule. While the Lift is the score for Cover, Ok values and the “count” (total population), determined in Eq. (4). While "description" is the rule expression in UEL format.

3.5 Lift

Association rules are used to identify associations between variables. Analyses based on association rules in many fields that are particularly useful in large datasets [22]. In data mining and association rule learning, the Lift [21, 22] is a measure of the performance of a model (association rule) for prediction or classification as having an enhanced response (with respect to total population), measured against a random choice of model. So, Lift is ratio of target response divided by the average response.

For example, if the average response rate of a population is 4%, but a segment in a model or rule has a response rate of 12%. Then the Lift score of the segment would be 12% / 4% = 3.0. Let us consider Dataset 1 (Bank dataset) with a distribution of transactions from UK with 4 Fraud and 2 None cases, while 4 Fraud cases from AU. Consider the following rule:

Rule: UK implies Fraud, i.e. IF Country is UK THEN Class = Fraud

$$\text{Lift} = (\text{Ok} / \text{Cover}) / (\text{Cover} / \text{Total}) \tag{4}$$

The Lift for the rule using Eq. (4) is $(4/6)/(6/10) \approx 1.11$

When Country is UK and Class is Fraud = 4 (OK)

When Country is UK = 6 (COVER)

Total population(instances) = 10 (TOTAL)

While evaluating the expressions of the rules, when multiple rules are true, choosing the predicted class of better Lift score (higher confidence) rule will increase accuracy.

3.6 Unified Expressions (UE)

UEL can evaluate mathematical expressions with various operators. It enables dynamic scripting feature. Some of the advantages of UEL is that it supports more than 30 different operators; Rule-based classifiers use only limited operators but using UEL many more operators can be used which are not available in rule-based classifiers, e.g. IN and LIKE Operators. Authors in [3] have highlighted the importance of compactness of the prediction model and demonstrated that a compact prediction model is more efficient. The UE will help in ruleset compactness along-with revised Lift score and hence will improve performance in terms of the time taken for model prediction.

$$\text{Expression}^{\text{UE}} = \text{funcUEL}(\text{Expr}^{\text{RDR}}) \tag{5}$$

where Expression is a UE format and ExprRDR is RDR format expression. funcUEL is a function to convert RDR format expression to UEL format. Main function of funcUEL is to transform RDR operators and operands to UEL operators and operands. Few of the transformation are:

Transform “and” to “&&” operator, “=” to “==” operand.

To make the transformation more generic, profiles are used for transformation operators and operands. Table below shows the transformation detail.

Table 3. RDR and UEL transformation.

RDR	UEL	Category
And	&&	Operator
=	==	Operand

3.7 Compactness

The compactness of ruleset can improve the performance of the algorithms and has been proposed in this paper. One of the challenges was deciding which rules to compact. One of the approaches considered was the nearest neighbour technique using Euclidian based similarity of the instances of two rules. This approach determines [25] distances using Eq. (6) and Eq. (7):

$$Dp = \sqrt{0.2^2 + 0.3^2} = 0.36 \tag{6}$$

$$Dn = \sqrt{0.4^2 + 0.3^2} = 0.5 \tag{7}$$

where Dp and Dn are the distances of class p and n respectively. However, this technique is computationally expensive, so instead, a customized threshold-based approach is used. The measures and threshold used in the technique are listed in Table 4.

Table 4. RDR and UEL transformation.

Measure	Threshold
Nearest Lift score	<= 0.05
Same parent rule	
Smaller expression rule	<=2
IN / BETWEEN operators	> 2

New values of Ok, Cover and Lift score are calculated for merging rules of the customized scheme.

4 Experimental Setup and Data

A multimode Hadoop cluster including Spark nodes was set up on a NECTAR [27] research cloud to develop and evaluate this technique for large datasets. Spark is ideal for iterative machine learning tasks and is much faster than conventional MapReduce. Figure 2 is a typical diagram of Spark [28] internal execution on a Hadoop cluster, which makes it iterative and more efficient than MapReduce.

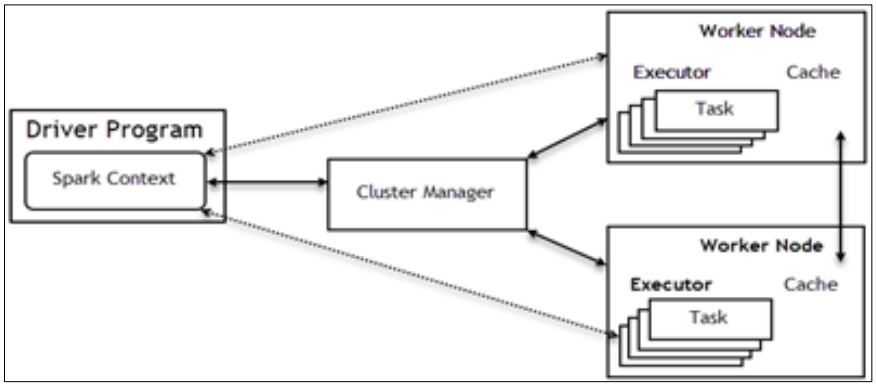


Fig. 2. Spark execution flow.

Characteristics of multiple datasets used for evaluation are listed in below table.

Table 5. Dataset characteristics.

Dataset	Description	Instances	Features
Dataset 1	Reference Bank Data [29]	1,756	14
Dataset 2	Synthetic Bank Data [29]	100,000	14
Dataset 3	German Credit Data [30]	1,000	11
Dataset 4	Credit Approval [31]	691	16
Dataset 5	Adult (Census Income) [6]	32,562	8

Synthetic Bank data was generated from reference Bank data using HCRUD [29] technique. This technique can produce huge dataset on the Hadoop cluster, which is similar to the original reference dataset. The dataset is produced with uniform distribution of class labels, individual and combination of attributes as well. RMSE of the difference of distributions in individual attributes is between 0.00 to .78, while the combination of attributes is between .80 to 1.85. Spark can use huge datasets, but for evaluation purpose, 100,000 instances of the dataset were used.

5 Results

Classification accuracy of UE-RDR technique is compared with existing RDR implementation in WEKA (RIDOR). An empirical evaluation was performed with various datasets listed in Table 5, with 30% / 70% split for training and testing datasets respectively. Average measurements were taken for various small to large dataset sizes and with multiple simulation executions. Vertical axes in Figure 3 - Figure 5 are the percentage of performance improvement of UE-RDR models over the other classifiers. Performance comparison for the accuracy is shown in Figure 3 and Figure 5, the accuracy is ratio of correctly predicted observations to total observations.

$$\text{Accuracy}= \frac{TP+TN}{\text{TP+FP+FN+TN}} \tag{8}$$

where true positives (TP) are the correctly predicted positive values and true negatives (TN) are the correctly predicted negative values, false positives (FP) when actual class is no and predicted class is yes and false negatives (FN) when actual class is yes but predicted class is no.

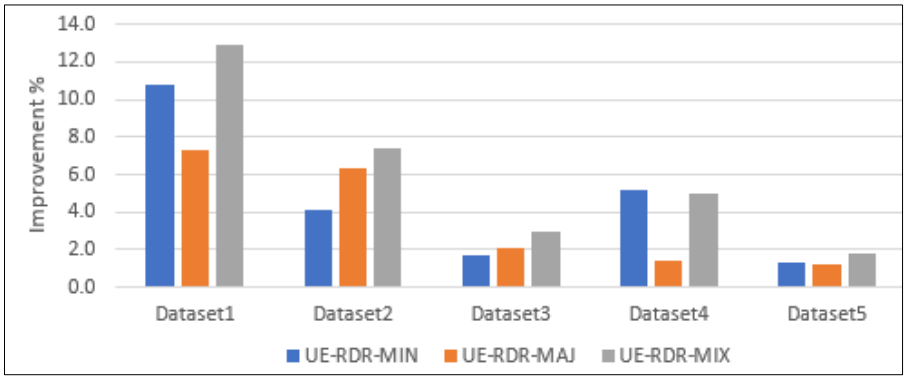


Fig. 3. Improvement in Classification Accuracy over RIDOR.

The results show that classification accuracy with all the datasets is improved. Out of the three UE-RDR models, UE-RDR-MIX performance is best among all datasets other than Dataset 4 (Credit Application dataset) where UE-RDR-MIX and UE-RDR-MIN accuracy is almost the same.

Similarly, ruleset compactness results are displayed in Figure 4.

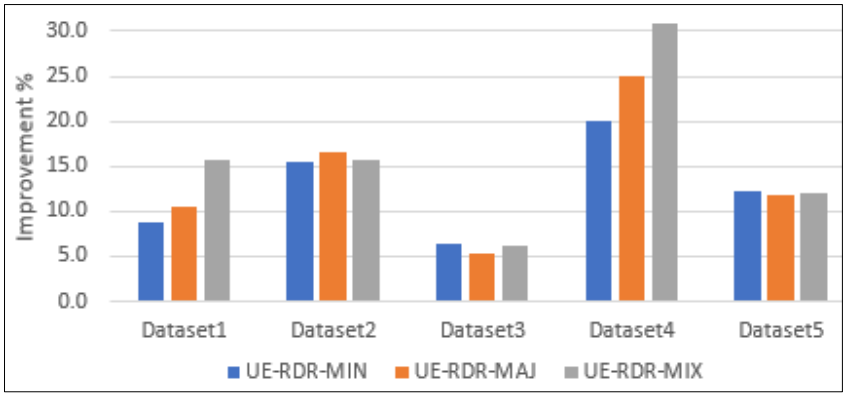


Fig. 4. Improvement in Ruleset Compactness over RIDOR.

The results show that compactness with all datasets is improved. However, UE-RDR-MIX compactness is better in Dataset 1 (Bank dataset) and Dataset 2 (Synthetic Bank dataset). For the remaining three datasets, either UE-RDR-MIN or UE-RDR-MAJ

model performance is better. Figure 5 shows the improvement in classification accuracy, while Figure 4 shows the improvement in ruleset compactness with UE-RDR as compared to RIDOR. IPA accuracy for mixed Bank data is compared with UE-RDR-MIX model. Table 6 shows that UE-RDR accuracy is higher than IPA classifier.

Table 6. Accuracy comparison with IPA.

Technique	Accuracy
UE-RDR-MIX	83.76%
IPA[10]	73.90%

For further verification, the UE-RDR-MIX classification accuracy is also compared to a non-RDR classifier: Naïve Bayes. Figure 3 shows that UE-RDR accuracy is higher than Naïve Bayes accuracy for all datasets, with substantial improvements in accuracy for Datasets 1 and 4.

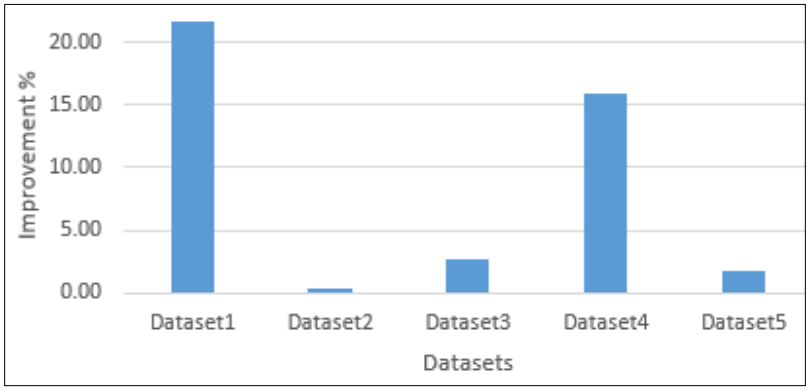


Fig. 5. Improvement in Classification Accuracy over Naïve Bayes.

Classification accuracy is compared among the three UE-RDR-models for a specific class label for mixed Bank data. Figure 6 shows that classification accuracy is higher with UE-RDR-MIX model.

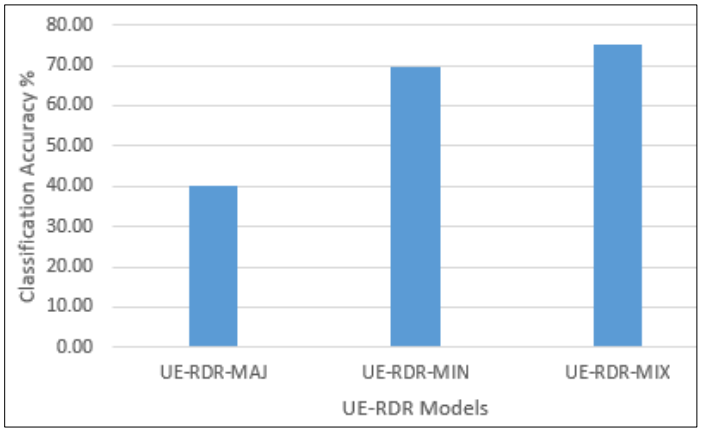


Fig. 6. Classification accuracy in Fraud Class among UE-RDR models

Results in Figure 3, Figure 4 and Table 6 show that UE-RDR-MIX model gives best classification accuracy. Figure 6 shows that a specific class label which is neither majority class nor minority class, also has a higher classification accuracy with UE-RDR-MIX model. Reason for higher accuracy is due to combined and compact rules in UE-RDR-MIX model for that class from majority and minority training models.

6 Conclusion

Fraud detection for online banking requires higher classification accuracy for the detection to enhance the confidence of its customers. Out of the available rule-based techniques for fraud detection, RDR is ideal due to its lower maintenance and incremental learning. However, testing and evaluating RDR on distributed and Big data platform is a challenging task, as the RDR classifier has not yet been implemented on Spark. Paper has shown that the challenge in fraud analysis due to the heterogeneous nature of transactions data (mixed attributes) and Big data can be overcome with UE-RDR. Introducing Unified Expressions in the RDR and evaluating the expressions based on Lift score helped to achieve ruleset compactness and higher accuracy. Further three models, including UE-RDR-MIN, UE-RDR-MAJ and UE-RDR-MIX are also developed in this paper. UE-RDR-MIX is the most innovative model, which does not exist in RIDOR. It combines and further compacts Majority and Minority class models with least usage of default class and unlike RDR it contains rules of all class labels, so it gives better accuracy from RDR based classifiers.

Classification accuracy is compared with existing RDR implementation: RIDOR. This technique is applied on various datasets including fraud analysis Bank & Synthetic Bank datasets and three publicly available German Credit, Adult (Census Income) and Credit Approval datasets. The empirical evaluation has shown that not only the ruleset size of training and prediction dataset is reduced, but classification accuracy is also improved. Classification accuracy with UE-RDR for Bank dataset is

also compared with another RDR based IPA technique and a non-RDR classifier (Naïve Bayes). Results have shown improvement in classification accuracy when compared with these classifiers as well. In this paper, the developed technique is used for the experimental validation and development of fraud analysis, but it can be used in other domains as well, especially for scalable and distributed systems. Further, this technique can be enhanced for other data formats (libsvm and arff) and a multi-classification system.

References

1. S. McCombie, "Trouble in Florida, The Genesis of Phishing attacks on Australian Banks," in 6th Australian Digital Forensics Conference., Perth, 2008, p. 16.
2. FBI. (2018, Internet Crime Complaint Center, 2018. 2019(20/08/2019). Available: https://pdf.ic3.gov/2018_IC3Report.pdf
3. I. Ul Haq, I. Gondal, P. Vamplew, and S. Brown, "Categorical Features Transformation with Compact One-hot Encoder for Fraud Detection in Distributed Environment," in The 16th Australasian Data Mining Conference, Bathurst NSW, Australia, 2018, pp. 69-80.
4. X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, et al., "Mllib: Machine learning in apache spark," Journal of Machine Learning Research, vol. 17, 2016.
5. N. Pentreath, Machine learning with spark: Packt Publishing Ltd, 2015.
6. Y. Kou, C.-T. Lu, S. Sirwongwattana, and Y.-P. Huang, "Survey of fraud detection techniques," in IEEE International Conference on Networking, Sensing and Control, 2004, 2004, pp. 749-754.
7. C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," arXiv preprint arXiv:1009.6119, 2010.
8. G. E. Melo-Acosta, F. Duitama-Munoz, and J. D. Arias-Londono, "Fraud detection in big data using supervised and semi-supervised learning techniques," in 2017 IEEE Colombian Conference on Communications and Computing (COLCOM, Cartagena, Colombia, 2017.
9. M. Herland, T. Khoshgoftaar, and R. Bauder, "Big Data fraud detection using multiple medicare data sources," Journal of Big Data, vol. 5, pp. 1-21, 2018.
10. O. Maruatona, "Internet banking fraud detection using prudent analysis," PHD PHD, School of Science, Information Technology and Engineering (SITE), UOB, 2013.
11. B. H. Kang, P. Compton, and P. Preston, "Multiple Classification Ripple Down Rules: Evaluation and Possibilities," in 9th Banff Knowledge Acquisition for Knowledge Based Systems Workshop, Banff, 1995, pp. 17-26.
12. A. Prayote, "Knowledge Based Anomaly Detection," PHD PHD, The School of Computer Science and Engineering, University of NSW, 2007.
13. O. Maruatona, P. Vamplew, and R. Dazeley, "RM and RDM, a Preliminary Evaluation of Two Prudent RDR Techniques," in Pacific Rim Knowledge Acquisition Workshop, 2012.
14. P. Compton and R. Jansen, "Knowledge in context: A strategy for expert system maintenance," in Australian Joint Conference on Artificial Intelligence, Adelaide, Australia, 1988, pp. 292-306.
15. D. Richards, "Knowledge-based system explanation: The ripple-down rules alternative," Knowledge and Information Systems, vol. 5, pp. 2-25, 2003.
16. A. Kelarev, R. Dazeley, A. Stranieri, J. Yearwood, and H. Jelinek, "Detection of CAN by ensemble classifiers based on ripple down rules," in Pacific Rim Knowledge Acquisition Workshop, 2012, pp. 147-159.

17. Y. S. Kim, P. Compton, and B. H. Kang, "Ripple-down rules with censored production rules," in Pacific Rim Knowledge Acquisition Workshop, 2012, pp. 175-187.
18. P. Compton, "Pacific Knowledge Systems: Challenges with Rules," University of New South Wales, Sydney, White Paper2011.
19. 1ASF. (2015, 26/04/2016). Apache Hadoop. Available: http://hadoop.apache.org/
20. J. G. Shanahan and L. Dai, "Large scale distributed data science using apache spark," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 2015, pp. 2323-2324.
21. G. Martinez. (2019, 12/08/2019). Lift (data mining). Available: https://en.wikipedia.org/wiki/Lift_(data_mining)
22. P. D. McNicholas, T. B. Murphy, and M. O'Regan, "Standardising the lift of an association rule," Computational Statistics & Data Analysis, vol. 52, pp. 4712-4721, 2008.
23. I. Ul Haq, I. Gondal, and P. Vamplew, "Enhancing Model Performance for Fraud Detection by Feature Engineering and Compact Unified Expressions," in 19th International Conference on Algorithms and Architectures for Parallel Processing, Australia, 2019.
24. S. R. Swain and S. S. Sarangi, "Study of Various Classification Algorithms using Data Mining," International Journal of Advanced Research in Science and Technology (IJARST), vol. 2, pp. 110-114, 2013.
25. J. N. Littin, "Learning relational ripple-down rules," PHD PHD, Computer Science Department, University of Waikato, Hamilton New Zealand, 1996.
26. B. R. Gaines and P. Compton, "Induction of ripple-down rules applied to modeling large databases," Journal of Intelligent Information Systems, vol. 5, pp. 211-228, 1995.
27. G. Moloney, M. Barker, P. Coddington, and K. Mecoles. (2011). NECTAR. Available: https://nectar.org.au
28. J. Aven, Data analytics with Spark using Python: Addison-Wesley Professional, 2018.
29. I. Ul Haq, I. Gondal, P. Vamplew, and R. Layton, "Generating Synthetic Datasets for Experimental Validation of Fraud Detection," in Fourteenth Australasian Data Mining Conference, Canberra, Australia, 2016, pp. 1-9.
30. H. Hofmann, "Statlog (German Credit Data) Data Set," U. a. Hamburg, Ed., ed. UCI Machine Learning Repository, 1994.
31. J. R. Quinlan, C4.5: Programs for Machine Learning, Ist ed.: Morgan Kaufmann, 1992.

Root causes labelling of industrial assets via relevancy estimation of event logs

Pierre Dagnely[1], Tom Tourwé[1], and Elena Tsiporkova[1]

Sirris - Elucidata Innovation Lab, A. Reyerslaan 80, 1030 Brussels, Belgium,
(pierre.dagnely, tom.tourwe, elena.tsiporkova)@sirris.be

Abstract. Assessing the performance of industrial assets typically requires exploring and combining sensor data, event logs, asset characteristics and domain expert knowledge. This process can be very time and resource consuming. Being able to extrapolate the asset performances solely from the event logs could be a valuable shortcut enabling optimal and pro-active planing of maintenance. In previous work, we have demonstrated that event logs could be numerically encoded into event profiles accurately representing characteristic asset event behavior. However, this methodology could be rather computationally expensive. We have also demonstrated previously that event logs can be used to classify operating cycles as faulty or not, although this methodology does not provide detailed performance profiles. We propose an integrated workflow enabling the rapid performance quantification of the operating cycles of an industrial asset. Our workflow extracts the event profile of a new operating cycle and links it with the similar event profiles of past operating cycles for which the performance is known. The performance of a new asset operating cycle can then be assessed with negligible computational time. The validation of this workflow on real industrial use case data have demonstrated that the performance labelling of new operating cycles using the previously observed ones can be very accurate.

keywords: Performance assessment · Classification · Event logs numerical encoding · Photovoltaic plants

1 Introduction

Rapid labelling of the performance of running operating cycles of any industrial asset is important for reliable monitoring and maintenance purposes. If the performance of an asset can be quickly assessed, maintenance activities can be planned in advance, and the asset unavailability can be potentially minimised. In addition, if the performance can be associated (labelled) with a representative profile indicating the root cause, then the maintenance team will already have useful insights about the current problem and its potential solution before arriving on site. It would lead to more optimal planning of maintenance activities.

However, extracting performance profiles indicating the potential root causes is a complex and time-consuming task. For instance, we have shown in [1], that

in the photovoltaic (PV) domain, it requires to combine irradiation data, yield sensor data, event logs, plant characteristics and domain expert knowledge. Such process is time and resource consuming and can not be performed every night to assess the performance of the thousands of plants in a portfolio.

Therefore, methods that could quickly label asset performances would be very valuable in industrial domains. One solution is to rely on the valuable information provided by the event logs. Meaningful event profiles can be extracted from the event logs. For instance, one profile would be mainly characterized by the occurrences of the event "over-temperature" while the other would be characterized by the occurrences of the events "under-temperature" and "sensor test". These event profiles represent various internal behaviors, e.g. the profile characterized by the event "over-temperature" reflects under-performance behaviour while the one characterized by "under-temperature" and "sensor test" reflects a regular behaviour. Assuming that these event profiles have been characterized, a classifier can then be trained on them. New operating cycles can then be classified by that model solely based on their event logs. For instance, if the asset reports "under-temperature" and "sensor test", the maintenance team will know that the asset is performing as expected.

The main challenges of this approach are: 1) the extraction of the relevant events for the event profile construction and 2) the characterization of these event profiles. A single event is often not enough to characterize an asset behaviour and the combination of the events needs to be considered since it ensures a correct representation of the asset behaviour. For instance, an asset reporting "under-temperature" would be under-performing while one reporting "under-temperature" and "sensor test" would perform as expected. Moreover, some events could be irrelevant, e.g. as they have been defined for debugging purpose by the manufacturer. Therefore, the combination of relevant events needs to be identified first. The second challenge is linked to the difficulty to have a precise understanding of an asset behaviour. As mentioned above for the PV domain, but also in other industrial domains, it often relies on the combination of multiple data sources and domain knowledge in order to ensure a correct labelling (with root causes) of the asset behaviour.

In this paper, we propose a methodology for root causes labelling of industrial assets based on their event logs. Our methodology first detects the relevant combination of events required to represent the asset event profiles. Then, it proposes a framework to guide domain expert in the labelling of these past event profiles. Finally, a classifier is used to link new operating cycles (i.e. event profiles) to the past labelled event profiles. Our methodology can be applied to any industrial asset generating events describing its behaviour and for which additional sensor data (describing its performance) is available.

This methodology is then showcased in the PV domain using real-life data from a Belgian plant. We have demonstrated that this methodology allows to extract profiles such as Profile A: "Inverter-days with small outages due to *Riso low*, mainly occurring in the "west" orientation" or profile B: "Inverter-days with high outages due to *Internal error*, mainly occurring at the end of the summer".

New operating cycles (days) of the plant can then be classified within one of these profiles with negligible computation time. Hence, maintenance teams can have a detailed overview of the plant performance (and root causes) at the end of the day and decide if a maintenance is needed without having to manually check the multiple sensor data streams sent by the plant.

The paper is organised as follows. First, the relevant literature is explained in Sections 2. Then, in Section 3, we explain our methodology for fast labelling of new operating cycles. In Section 4, we validate our methodology in the PV domain. Finally, we conclude the paper with a discussion in Section 5.

2 Literature Review

The leveraging of event logs for root cause labelling is still in its infancy and few researches have been conducted. The main challenge is the textual nature of the event logs while most of the existing machine learning algorithms are optimized for numerical values. The typical approach in text classification is then to extract numerical features from the text and apply the classification on these features. For instance, the features could be the total number of words in the documents, the average length of the words used in the documents or the total number of punctuation marks in the documents. However, these methodologies can be complex to deploy. As stated by Dalal et al. [2], they would require to be tailored to the event behavior, e.g. "do all events have the same impact?" or "is the repetitions of the same event relevant?", and would likely have to include meta-data. Therefore, textual classification poses a challenge to develop an agnostic methodology. On the other hand, numerical classification methodologies are well defined, more agnostic and validated in various domains. These methods can be applied to various numerical data, if the training dataset has been well constructed. For instance, SVMs have been applied to detect oil spill in ocean through the classification of radar images [3] or to predict the electricity price using historical data from the electricity market [4].

Therefore, another approach is to numerically encode the event logs. Fronza et al. [5] have applied such method using random indexing (RI) to numerically encode the events logs. RI is a data reduction method from the text mining field proposed by Sahlgren in [6]. This method is used to store in a condensed way the "context" of a word, i.e. the surrounding words. Fronza et al. have applied this methodology on event logs generated by software by considering each event as a word and each event log representing a software run as a textual document They have trained an SVM classifier to assess the performance of software runs as faulty or not, solely based on their event logs. They were able to classify the software runs as faulty or not faulty with a high accuracy.

In [1], we have proposed another approach based on TF-IDF (term frequency - inverse document frequency) and compared the performance with the RI approach of Fronza et al. for various classifiers. It appeared that both methodologies seem to have different accuracy performance depending on the application domain. We suspect that RI is more suited for procedural event data with strong

context-awareness while the relevancy score methodology is more adapted for variable, less contextually dependent, noisy log data, i.e. data with many irrelevant events. However, both studies focused on labelling asset runs as faulty or not, without providing any information of the root causes of these performances. It decreases the industrial applicability as domain experts rely on the root causes to decide if a maintenance team need to be send.

3 Methodology

We have developed a novel methodology to label performance of asset operating cycles based on their event logs. The steps of our methodology are the following:

1. Segment event log data into periods representing different operating cycles e.g. a full production cycle, a 24 hour operation, a segment between a certain start and end events
2. Convert the segmented operating cycle logs to numerical standardized profiles, i.e. event relevancy score vectors
3. Derive performance profiles (annotations/labels) by combining various sources e.g. production performance, energy consumption, exogenous conditions, etc.
4. Train a classifier to associate performance labels to different relevancy score vectors of operating cycles
5. Extract relevancy score vectors of new operating cycles and label/quantify their performance on the fly using the classifier

3.1 Relevance Score Extraction

The main challenge faced by our methodology is the textual nature of the event logs which hinders their processing. Extracting typical event logs, i.e. event profiles, from textual event logs would require domain knowledge. However, we intend to minimize the need of domain expert inputs as their time is valuable. We solved this problem by numerically encoding the event logs as relevancy score vectors. It allows to build agnostic clustering from the event logs. The relevancy score methodology that we defined in [1] is applied on the event logs. Our methodology follows 2 steps: 1) The event logs are segmented into representative operating cycles; 2) The relevancy scores are computed based on the event frequency.

Segmenting Event Sequences Into Atomic Event Logs The first step is to divide the event logs into atomic pieces, i.e. into "traces or meaningful periods" of the asset, called atomic event logs (AEL). For instance, in case of a car, the event logs could be divided into operating cycles, from the start of the travel to its end. The definition of these atomic event logs is therefore domain and goal oriented. The main interest is to transform the continuous stream of events into a meaningful finite set of event logs. These AELs will be easier to analyze and interpret. In addition, they contain all the event correlations. For example, the

interpretation of the event "temperature error" is modified in case it is preceded by the event "temperature sensor broken". The goal of the segmentation into AELs is to have the events that could interact all stored together into one file.

Computing Relevancy Scores We have used a method inspired by the widely exploited in text mining TF-IDF score, where for each event type of each AEL, its relevancy score is computed. The goal is to attribute a score reflecting the "abnormality" of the event, i.e. determine degree to which the event represents a deviating performance from the regular asset behavior. For example, the critical event "temperature error" that occurred 2 times in the atomic logs should have a high relevancy score as it indicates a failure, while the event "start" (representing the usual behavior of the device) that occurred 17 times should have a relevancy score of 0. Therefore, the events' frequencies need to be carefully exploited.

By considering the AELs as a text, text mining methods such as TF-IDF can be adapted for this purpose. Therefore, our methodology relies on the computation of two frequencies: 1) the frequency of the event (type) in the AEL, and 2) the frequency of the event (type) in well selected corpus of AELs aligned with the analysis goal in mind.

First, the term frequency (TF) is computed, i.e. for each event type that can be reported by the asset, its frequency in the AEL is computed. The formula below is used.

$$TF_{e_i,a_i,l_i} = \frac{\#\text{ occ. of events } e_i \text{ in logs of asset } a_i \text{ for AEL } l_i}{\#\text{ of event in AEL } l_i \text{ for asset } a_i}$$

The inverse document frequency (IDF) needs to be adapted to the industrial event logs context as the text corpus on which it relies does not apply here. Therefore, the corpus definition needs to be adapted. Three approaches are possible and need to be carefully selected:

- The corpus consists of all available AELs. It allows to compare asset behavior over time and across assets.

$$IDF_{e_i} = \log \frac{\#\text{ of AEL in all assets and all days}}{\#\text{ of AEL where event } e_i \text{ occured}}$$

- The corpus consists of all the AELs of one asset. It allows to focus on one asset behavior and monitor the evolution of performance over time.

$$IDF_{e_i,a_i} = \log \frac{\#\text{ of AEL for asset } a_i}{\#\text{ of AEL for asset } a_i \text{ with event } e_i}$$

- The corpus is composed of AELs of all assets for the same trace (e.g. the same day). It allows objective comparison of performance across assets for the same operating cycle. However, as events occurring in all AELs of the corpus are considered less relevant, a failure occurring in all assets would be masked by this case.

$$IDF_{e_i,p_i} = \log \frac{\#\text{ of AEL occurring at the period } p_i}{\#\text{ of AEL for period } p_i \text{ with event } e_i}$$

Subsequently, the relevancy score is computed by multiplying TF and IDF: Relevancy score $= TF_{e_i,a_i,l_i} * IDF$

In this way, the relevancy score uses the frequency of the event (more frequent events have higher scores) corrected by the IDF that will decrease the score of events frequent in the corpus (if an event occurs in all AELs of the corpus, its IDF is $\log(1) = 0$, which leads to a relevancy score of zero).

By computing the relevancy scores over all events for each operating cycle, a numerical vector representing the event relevancy, i.e. event profile, for the asset operating cycle is obtained. For instance, if the following three events occur in an AEL "over-temperature", "under-temperature", and "sensor test", than the event profile of this AEL will be represented by the vector [0,4,1] indicating that for that AEL, the event "over-temperature" is irrelevant (score 0), the event "under-temperature" is occurring and relevant (score 4) and the event "sensor test" is occurring but not very relevant (score 1). The textual representation of the events has then been transformed into a numerical feature vector. In this way, the textual event log sequences of variable length have been standardized by converting them into numerical event profile vectors of the same length (the number of different events).

3.2 Deriving Performance Labels

The approach is composed of 3 steps:

1. The relevancy scores are computed for each AEL as described in the foregoing section, using the approach where the reference corpus is composed of AELs of all assets for the same trace (operating cycle)
2. The (numerical) relevancy score vectors per AEL (across assets) are clustered to find typical profiles, i.e. each cluster corresponds to a "characteristic" event relevancy score profile and, hence, represents some typical asset behaviour in terms of events.
3. The resulting from the clustering characteristic relevance profiles are subsequently associated (annotated) with performance labels by using domain knowledge and additional (sensor) datasets i.e. the aim is to link (label) the profiles to some specific (critical) phenomenon e.g. failure or underperformance. In this way, the profiles can be used as indicator/summary of the asset performance for the considered operating cycle (including all the information present in the event logs and sensor data).

Clustering Relevancy Score Vectors As it was demonstrated in the foregoing section, by computing the relevancy scores over all events for each AEL, a numerical vector representing the asset event behaviour/profile for the trace represented by the AEL is obtained. Subsequently, this numerical encoding can easily be used as a basis for more advanced machine learning/data mining methods. For instance, relevancy score vectors can be used to extract representative profiles as they represent the event behavior of the asset. They are then clustered to extract clusters of relevancy score vectors, i.e. group of relevancy score

vectors with similar behavior. Relevancy score vectors are sparse vectors. During one operating cycle, usually, only few of all the possible events actually occurred, and of those events, only few of them are relevant, i.e. do not have a null score. Therefore, the vectors are mainly populated with zeros. Spherical k-means clustering has been selected due to its ability to deal with sparse data. The number of clusters is evaluated using traditional silhouette, Calinski-Harabaz and connectivity metrics. The clusters of similar event relevancy scores are clusters of similar asset behaviour for the period represented with the AEL corpus, i.e. each cluster represents an event profile of the assets.

Profile Characterization Once the event profiles are extracted by spherical k-means, they need to be labelled using additional data sources. The clustering has indicated what are the typical event profiles encountered in the assets. This step tries to characterize further the event profiles by associating them with some performance indicators, i.e. is a certain event profile linked to under or over performing asset? It requires to combine all the various data streams generated by (most) industrial assets nowadays.

The data that can be used to label event profiles can come from: 1) sensors embedded in the assets, 2) asset characteristics as described by the manufacturer, 3) asset configuration, 4) domain experts, 5) data reported by additional monitoring systems, e.g. a UAV monitoring a factory, or 6) data generated by other assets under the same operating conditions. The labelling is the only domain dependent part of our methodology. Therefore, a thorough presentation of this step is impossible here, but some of the labelling alternatives are demonstrated in the validation section. Nonetheless, there are different strategies to label the event profiles that can be listed e.g. consider whether the operating cycles represented by a certain event profile

- produced as expected based on a physics-based model of the asset behavior or compared to similar assets (production-based)
- experienced any failure or other critical operation disruption (failure occurrence)
- have all a particular configuration, e.g. share the same asset type or installation (technical characteristics)
- belong to the same time-windows, e.g. month or season (time-dependency or seasonality)
- only occur in one or a few assets (asset specificity)
- exclusively contain certain specific events (event occurrence)

A more detailed description of possible labelling methods is provided in the validation section.

3.3 Classifier Training

The goal is to build a model classifying any relevancy score vector generated from the event logs into their corresponding performance profile, e.g. classify the

operating cycle represented by the relevance score vector as healthy or as underperforming. Any general purpose classifier can be used due to the numeric nature of the relevancy score vectors. In addition, the relevancy score methodology preprocesses the event logs in such a way that it hides the irrelevant events and pinpoints the most important ones, which also has a positive impact on the quality of such classification model, which can be easily constructed using any widespread classifier (e.g. kNN, SVM, ...) given the availability of representative datasets.

3.4 Labelling New Operating Cycles

The classification model derived in the foregoing subsection can then be used to classify the performance of newly incoming operating cycles, solely based on their event log sequences. The relevancy score vector of a new operating cycle can be extracted from its event log. The classifier can then detect to which performance label it belongs and hence quantify rapidly its performance. For new operating cycles, the relevancy score vectors is extracted and fed to the classifier model which output the operating cycle labelled profile. This process has negligible computation time and resources. The only complex task is the labelling of the past profiles, when domain experts need to manually label them. However, this task only needs to be performed once, when the model is built.

4 Evaluation and Discussion

The critical performance aspect of our methodology is the ability of the classification algorithm to correctly label the new incoming operating cycles. We have first extracted the performance profiles from the data. As there is no ground truth, i.e. labelled data, it is not possible to statistically assess the correctness of the performance labels. The latter has been evaluated via a visual inspection by a domain expert from our industrial partner 3E, which is active, through its Software-as-a-Service SynaptiQ, in the PV plant monitoring domain. Subsequently, we have used the data validated by the domain expert to evaluate the accuracy of the classification model.

4.1 Data Understanding

We have used one year of event logs from one - often faulty - PV plant. The data has been provided by our industrial partner. PV plants are composed of several PV modules (that convert the irradiation into direct current) connected to one or several inverter(s) (that convert the direct current to alternative current) which send the current to the grid. These systems are now continuously monitored. In addition, various sensors (measuring the irradiation reaching the plant, electricity production, ... usually at a 15 min granularity) are present in the plant. Meteorological data are also either measured on site or inferred from

nearby meteorological stations and satellite measurements. These meteorological data usually cover irradiation, temperature, rainfall, snowfall and wind speed measurements. An inverter reports status, i.e. its current state like start, stop or running, but also other events that can represent e.g. an outage (such as grid failure or string disconnected) or other phenomena (such as over-temperature or DC current under threshold).

In the PV case, an AEL corresponds to the event logs of one inverter for one day. As the plant is only active during the day, it can be considered that it "reboots" at night (often, small problems disappear the next morning).Therefore, each day corresponds to an operating cycle. In addition, as the events are monitored and reported at the inverter level, AELs are at the inverter level. Hence, for our plant with 26 inverters, 9490 AELs have been obtained (26 inverters-logs times 365 days). An AEL typically contains around 5 distinct event types. Over our one year dataset, 54 distinct event types have been reported. Therefore, our relevancy score vectors have a length of 54 but only around 5 non zero values.

4.2 Profile Extraction

Based on the above findings, we display in this section the labelling of the data. We have extracted 12 clusters, based on the silhouette, Calinski-Harabaz and connectivity scores. In the following we describe their labelling using 7 distinct characterization processes.

Model-based Characterization The performance of an inverter can be compared to the expected yield. This expected yield can be modelled based on the amount of irradiation reaching the plant. If the amount of irradiation reaching the inverter and the inverter capacity are known, the amount of electricity that should be produced can be computed. Therefore, it is possible to compare the real and expected yields to assess if the inverter-days of one profile are over or under-performing.

However, current losses naturally occur in PV plant between the PV modules and the inverters (due to some physical properties of the system). Those losses are complex to estimate and are, therefore, not included in the computation of the expected yield. It implies that an inverter will never be able to reach its expected yield. By definition, the closer inverters are to the expected yield, the better they behave.

The difference between the daily electricity production of the inverter-days of each profile and their expected electricity production is computed and aggregated per profile. The results are shown in Figure 1. Inverter-days in profile 0 and 6 are labelled as behaving well as they are close to their expected yield, i.e. close to the red line. The profile 2 and 9 are identified as under-performing.

Portfolio-based Characterization However, in practice, the expected amount of electricity that should be produced is not always reliable. For instance, as the amount of irradiation reaching the plant is not correctly known due to that

the irradiation sensors on site not correctly oriented or regularly cleaned. The expected yield provides an indication but is not always trustworthy and other metrics should be used to confirm this analysis. For instance, by comparing the production to the one of the other assets of the portfolio at the same time, i.e. by analyzing if the inverter performed better or worse than the other inverters in the plant for that day.

It is achieved by comparing the yield of the inverter-days in one profile to the yield of the other inverters (of same capacity and orientation) for the same day, i.e. for each inverter-day of a profile. By comparing the production of inverters that share the same capacity and PV module orientation, the over or under-performing inverter-days can indirectly be assessed. A boxplot of this comparison can be seen in Figure 2, profiles below the red line correspond to inverter-days under-performing. It appears that profiles 3, 7 and 9 behave above average while inverters-days in profiles 2 and 8 are slightly below average. Most of the other profiles are in the average.

By comparing with the previous electricity production characterization, profile 2 can be labelled as under-performing as it is labelled as such by both approaches. However, the other profiles are characterized as over-performing by one analysis and as under-performing by the other. Therefore, no consistent conclusions could be drawn for these profiles. It demonstrates the difficulty to characterize the profiles without domain knowledge and ground truth.

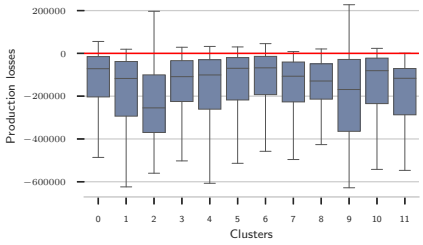


Fig. 1: Inverter-day electricity production compared to the expected electricity production of the inverter-day

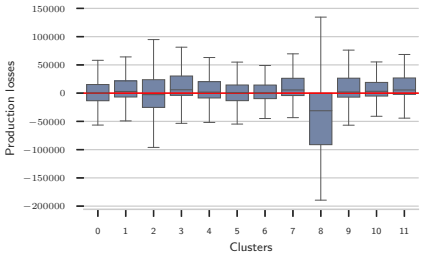


Fig. 2: Comparison of the inverter-day electricity production of each profile with the mean electricity production of the inverter-day of all profiles

Failure-based Characterization The links between failures and profiles can be analyzed. It can be done by checking for each inverter-day of each profile if an outage/failure occurred during that day. Then the mean amount of production lost due to that outage is computed (i.e. the amount of sun reaching the inverter that has not been converted into electricity due to this failure.). It is depicted in Figure 3, where inverter-days in profile 9 are associated with high losses.

Inverter-days in profile 2 are associated with low losses while the other profiles are not associated with any significant losses.

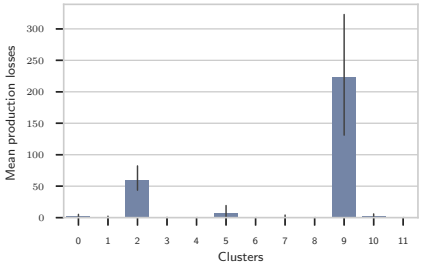


Fig. 3: Mean amount of production lost due to outage occurring during the inverter-day of each profile

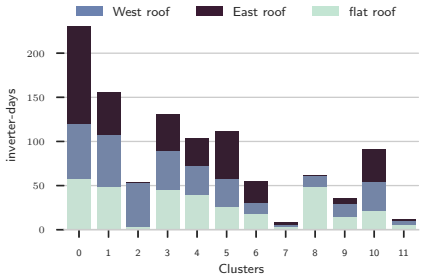


Fig. 4: Orientation distribution of the modules associated to each inverter-day of each profile

Configuration-based Characterization The orientation of the modules associated to the inverter can be analyzed. In a PV plant, the inverters are not always oriented in the same direction. For example, in plant A (which is on a factory roof), some modules are oriented to the east while others are oriented to the west. Some modules are also "flat", i.e. simply put on the roof, due to space limitation. Similarly to the other visualizations, the (normalized) orientation distribution of the modules associated to each inverter-days of each modules can be compared, as shown in Figure 4. It appears that the orientation "Flat" is preponderant in profile 8 and the orientation "West" in profile 2, indicating that these profile behavior might be orientation dependent.

Temporal-based Characterization The time of occurrence of each profile can be also considered as a discriminative feature. For instance, the monthly distribution of each profile can be displayed. In Figure 5, for each profile, the amount of inverter-days of that profile occurring each month is shown. It clearly appears that profile 2 mainly occurs during the end of the summer and profile 8 during the end of the year.

Specificity-based Characterization The distribution of the inverter-days across the portfolio can also be analogously explored, e.g. are certain profiles only occurring in one or few inverters? In Figure 6, for each profile, the amount of inverter-days of that profile occurring in each inverter is shown. The profile 8 only occurs in three inverters 10. Profile 2 mainly occur in 15 inverters.

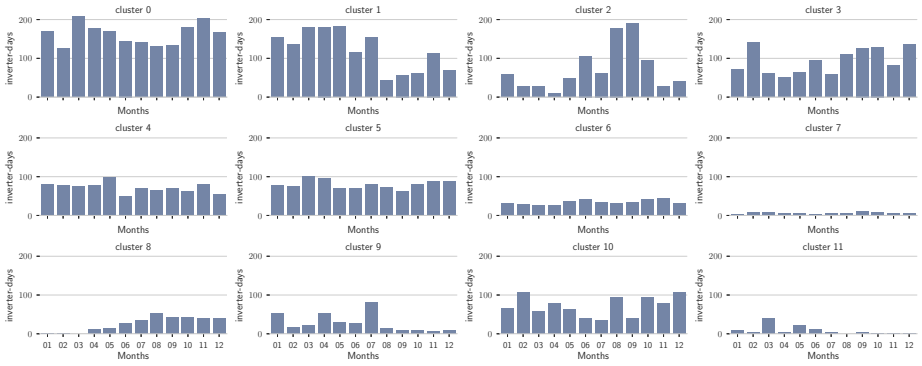


Fig. 5: Monthly distribution of inverter-days associated to each profile

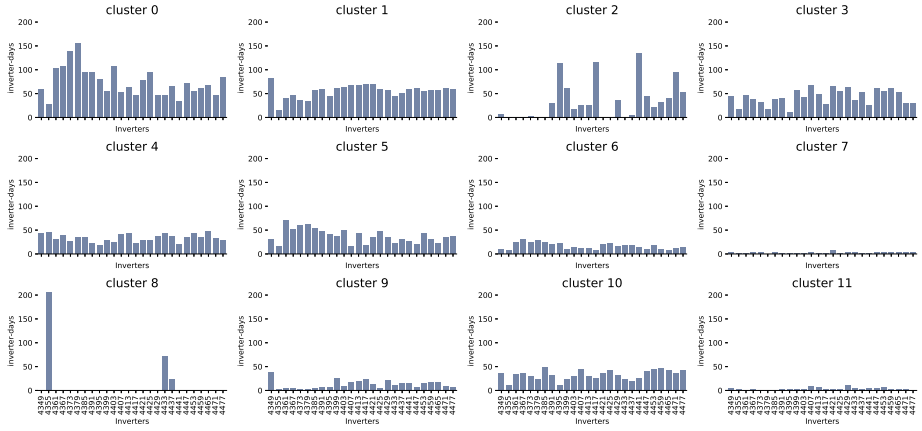


Fig. 6: Inverter distribution of inverter-days associated to each profile

Event-based Characterization The profile labelling can also take the relevant events into account. They should confirm the labelling extracted from the previous exploration. Due to space limitation, the event profile of each inverter-day can not be displayed in the paper. However, we have observed that profile 2 corresponds to inverter-days where the event *Riso low*, a failure, is important. Profile 9 correspond to event *Internal error* and profile 8 to event *Bulk over-voltage*.

Performance Profile Labelling The explorations and characterisations of the clusters presented above can be combined and used to assign performance label to each of the 12 event profiles. Majority of profiles are representing “regular" performance, i.e. no deviation in terms of production performance, outage, etc. as it was expected (as most of the inverter-days have varying regular behaviors). However, some profiles are more interesting:

– Profile 2: Inverter-days with small outages due to *Riso low*, mainly occurring in the "west" orientation
– Profile 8: Low production, only occurring in 3 inverters during the end of the year, due to *Bulk over-voltage*
– Profile 9: Inverter-days with high outages due to *Internal error*, mainly occurring in the end of the summer.

The nine remaining profiles represent the variability in regular operating behaviour. Domain experts and maintenance teams are less interested in those but rather expect to know whether the plant is behaving regularly opr there is some anomaly. However, it is important to adequately capture the variability of regular performance in order to be able to reliably detect any significant deviations.

The plant history can then be visualized through the performance indicators. It is illustrated in Figure 7, which shows the performance indicators of each inverter (in ordinate) of plant A for each day (in abscissa) for August to November. For readability purpose, the nine regular profiles have been merged into one, only displaying the irregular profile in Figure 8. It easily appears that the occurrences of profile 9 slightly decrease through the month of August.

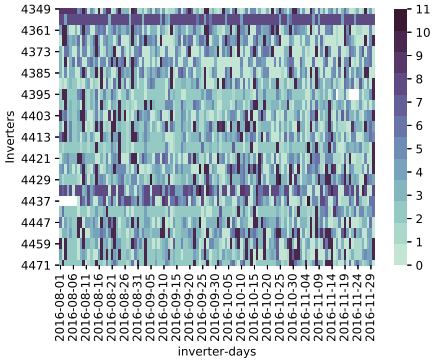


Fig. 7: Plant performance for 4 months, visualized through the 12 performance profiles

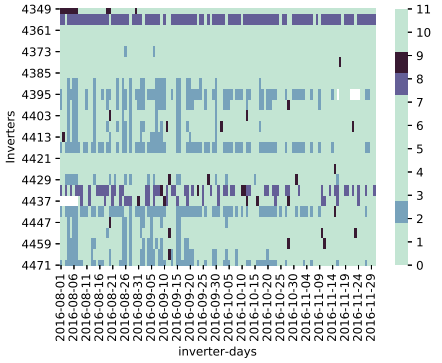


Fig. 8: Plant performance for 4 months only showing the 3 irregular profiles and merging the 9 regular profiles

4.3 Evaluation of Performance Classification

We have applied a standard kNN, as it has been shown to deliver more accurate results in the given application domain [1]. We have applied 10-fold cross validation on the available dataset. Applied on the 12 performance profiles, kNN was able to label the new operating cycles accurately, as shown with the confusion

matrix in Figure 9. The figure displays, for the 12 performance profiles, the percentage of instances in a predicted class versus the instances in an actual class. For most of the profiles, the classification was correct for 96% to 100% of the test dataset.

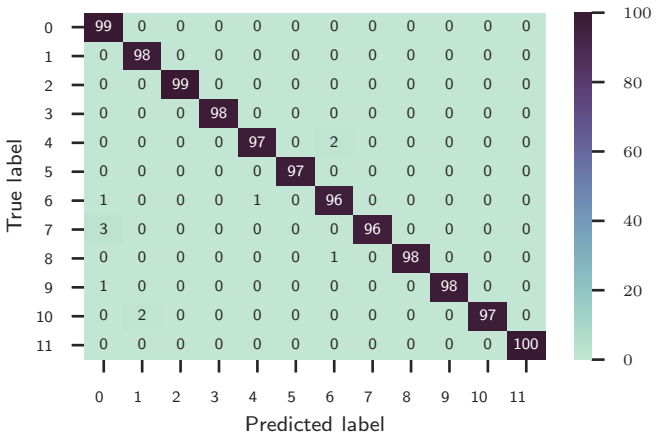


Fig. 9: Confusion matrix of the labelling of new inverter-days in the 12 existing profiles

In addition, not all performance profiles have the same importance. The analysis of these 12 profiles has shown that only profiles 2, 8 and 9 were relevant for detecting non-regular behaviors, while the other nine profiles represent diverse regular behavior. Therefore, only the labelling of these three profiles is crucial, as the difference between the other profiles is less relevant for maintenance purposes. The classification of profiles 2, 8 and 9 delivered an accuracy of respectively 99%, 98% and 98%. The incorrect labelling of the 3 profiles is as follows:

- Profile 2 is wrongly labelled as profile 7, a regular profile, in 1% of the cases
- Profile 8 is wrongly labelled as profile 6, a regular profile, in 2% of the cases
- Profile 9 is wrongly labelled as profile 0, a regular profile, in 2% of the cases

Overall, the three relevant profiles are correctly labelled with a high accuracy of at least 98%. The accuracy of the labelling of the other profiles never go below 96%. Moreover, the other labelling errors do not impact the accuracy of the method as the other profiles represent regular behaviors, i.e. operating cycles where the inverter behaved as expected. For example, the 3% of the operating cycles with profile 7 that have been labelled as profile 0 have no impact on any decision that a maintenance partner could take. Both profiles indicate that the inverter had a regular behavior and that no action should be taken.

As a conclusion, we have proposed an integrated workflow enabling the rapid performance quantification of the operating cycles of an industrial asset. The

validation of this workflow on real industrial use case data have demonstrated that the performance labelling of new operating cycles using the previously observed ones can be very accurate. It demonstrated further that the generated performance profiles are significantly distinct leading to robust classification performance using kNN.

5 Conclusion

Our methodology allows rapid annotation of new asset operating cycles with a known performance label or profile, only based on event logs. Assessing the performance of an asset is time and resource consuming as it implies analyzing the event logs, the various sensor data, the asset characteristics and requires domain experts' knowledge. We have shown that our methodology was able to label performance of new operating cycles with a mean accuracy of 98% using a kNN classifier, solely based on the event logs and with negligible computation time.

Acknowledgements

This work was subsidised by the Region of Bruxelles-Capitale - Innoviris.

References

1. P. Dagnely, T. Tourwé, and E. Tsiporkova, "Annotating the Performance of Industrial Assets via Relevancy Estimation of Event Logs," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 1261–1268.
2. M. K. Dalal and M. A. Zaveri, "Automatic Text Classification: A Technical Review," *International Journal of Computer Applications*, vol. 28, no. 2, pp. 37–40, Aug. 2011. [Online]. Available: http://www.ijcaonline.org/volume28/number2/pxc3874633.pdf
3. V. V. Lakshmi, "Oil Spill Detection in Oceans using Threshold Segmentation and SVM classification," p. 4.
4. Z. Shao, S. Yang, F. Gao, K. Zhou, and P. Lin, "A new electricity price prediction strategy using mutual information-based SVM-RFE classification," *Renewable and Sustainable Energy Reviews*, vol. 70, pp. 330–341, 2017.
5. I. Fronza, A. Sillitti, G. Succi, M. Terho, and J. Vlasenko, "Failure prediction based on log files using Random Indexing and Support Vector Machines," *Journal of Systems and Software*, vol. 86, no. 1, pp. 2–11, Jan. 2013. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0164121212001732
6. M. Sahlgren, "An introduction to random indexing," 2005.

Application of Machine Learning to Predictive Maintenance

Feranmi Akanni[1] and Martti Juhola[2]

[1] Tampere University, Tampere, Finland.
[2] Tampere University, Tampere, Finland.
feranmitimothyakanni@gmail.com

Abstract. The advancement in technologies has contributed largely to almost every areas of life, and importantly, most organizations use technology in various dimensions to drive optimization of their business values and build competitive advantages. Maintenance in industry is one of the tools that an organization can use to achieve optimization of the business value of a functional unit and this optimization can only be achieved through predictive maintenance which prevents an occurrence of breakdown of functional units because of missed maintenance activities and at the same time ensures that maintenance activities are not carried out before due time to avoid unnecessary cost on the functional units.

We studied maintenance engineering focusing on the use of classification methods to predict failures of a functional units. We started by exploring a real-life data from a functional unit and use missing data techniques to handle missing values in the dataset, which resulted in a complete dataset. We explored various feature selection techniques to extract important features and reduce dimensionality of the dataset. Then, we explored the uses of the following machine learning methods: logistic regression, naïve Bayes, support vector machine, k-nearest neighbor searching and ensemble learning techniques which are bagging and boosting methods.

We use evaluation metrics to compare the performance of different machine learning methods. The results of this experiment indicated that ensemble learning techniques performed better than other machine learning methods because the predictions from ensemble learning techniques produced better evaluation metrics.

Keywords: predictive maintenance, machine learning, evaluation metrics

1 Introduction

Maintenance can be described as the set of activities and actions which involve functional checking, servicing, testing, measurement, repairing or replacing of devices, equipment, machineries, and supporting utilities in industrial, business, governmental and residential environment [1]. Maintenance can also be defined as the combination of all technical and associated administrative actions intended to retain an item in or

restore it to a state in which it can perform its required function (British standard glossary of terms used in terotechnology, 1993) [2].

Maintenance ensures that the functional units are effective in their performance, preserves the life span of the functional unit and contributes to the sustainability and availability of the functional units. The lack or ineffectiveness of maintenance activities can contribute negative effects to the overall business performance through their impact on quality, the availability of the equipment, the organization competitiveness and the organization environment.

There are three main types of maintenance, which are corrective, preventive and predictive maintenance. Corrective maintenance is a type of maintenance where maintenance activities are carried out on the equipment mainly after the breakdown of the equipment. Preventive maintenance is also referred to as predetermined preventive maintenance and is a type of maintenance where maintenance activities are carried out on the equipment at fixed interval to avoid malfunctioning or breakdown of the equipment. These two types of maintenance are referred to as traditional maintenance strategies. Predictive maintenance is also referred to as condition-based maintenance (CBM). CBM is a set of maintenance actions based on the real-time or near real-time assessment of equipment condition, which is obtained from embedded sensors and/or external tests and measurements, taken by portable equipment and/or subjective condition monitoring [3]. Predictive maintenance is maintenance carried out following a forecast derived from repeated analysis or known characteristics and evaluation of the significant parameters of the degradation of the equipment [4].

Jantunen et al. [5] suggest that the concept of maintenance has evolved over the last few decades from a corrective approach (maintenance actions after a failure) to a preventive approach (maintenance actions to prevent the failure). Notably, the path of evolution of the maintenance activities has been from non-issue to business strategic concern. Initially, maintenance was majorly seen as an inevitable part of production where the maintenance activities were carried out after the breakdown of the equipment because downtime was not a critical issue and it was adequate to carry out maintenance after breakdown.

Later, it was conceived that maintenance was a technical matter and this did not only include optimizing technical maintenance solutions, but it also included the attention of the organization on the maintenance work [6]. Going forward, maintenance was separated from being a subfunction of production and was considered as a functional unit which represents one of the profit contributors to the organization. At this stage, the downtime from equipment breakdown was a critical issue and maintenance activities were carried out to prevent equipment breakdown.

The major impact of technology advancement in the area of maintenance can be observed in predictive (condition-based) maintenance where sensors are used to measure relatively huge amounts of data about the conditions of the equipment and this data is used to create models using different methods such as machine learning methods to determine the optimal time to carry out maintenance activities on the equipment just before the equipment failure or breakdown. The new technology such as IoT promotes the instantaneous availability and accessibility of the data about the conditions of the machines or products.

Riccardo et al. [7] exploited three classification models which were decision trees, random forests and neural network to a complex high-speed packing machine for making decision related to predictive maintenance and the result of their study revealed that random forest classifier performed better than other two classifier in terms of accuracies of the models. C. Gondek, D. Hafner & O. Sampson [8] used combination of feature engineering and one classification method which was random forest to predict the failure of Air Pressure System of Scania Trucks. This research study is different from the mentioned studies because it explored combination of feature engineering techniques with different classification methods. This research paper entirely concentrated on a novel application topic for industrial machine learning. Therefore, no novel methodological research was presented.

This paper is arranged in different sections: Section 2 describes the data used, Section 3 involves the methods used and how the methods are applied in achieving the obtained results of the experiment. Section 4 consists of the obtained results of machine learning methods and the description of the results in comparison to other results of the experiments. Finally, Section 5 presents the conclusion of the experiments and the discussion on the future recommendation on area of improvement.

2 Data Used

The experiment focuses on using different classification methods of supervised machine learning on data collected from a heavy Scania truck of Scania AB organization which is a major Swedish manufacturer of commercial vehicles. The public dataset which was discovered from UCI machine learning repository website https://archive.ics.uci.edu/ml/datasets/APS+Failure+at+Scania+Trucks, consists of data generated from everyday utilization of a heavy Scania truck and the main component of focus is Air Pressure System (APS) which generates pressurized air for effective operation of various components of the truck such as brake and gear components. The dataset consists of one response variable which is named class and 170 independent variables which have been anonymized for security purpose and to reduce the risk of unintended usages of the dataset.

The dataset includes the training set which consists of 60,000 instances and the test set which consists of 16,000 instances. The class label of the response variable for the training set consists 1,000 cases with the positive class and 59,000 cases with the negative class while the class label for the test set consists of 375 cases with the positive class and 15,625 cases with the negative class. The positive class of the dataset indicates a truck with failures which are related to APS and requires that the maintenance should be carried out on the APS just before breakdown. The negative class of the dataset indicates a truck with failures which are not related to APS. Tabel 1 represents a small section of the dataset.

Table 1. Excerpt of the data set

Filter

	class	aa_000	ab_000	ac_000	ad_000	ae_000	af_000	ag_000	ag_001	ag_002	ag_003	ag_004
1	neg	76698	NA	2130706438	280	0	0	0	0	0	0	37250
2	neg	33058	NA	0	NA	0	0	0	0	0	0	18254
3	neg	41040	NA	228	100	0	0	0	0	0	0	1648
4	neg	12	0	70	66	0	10	0	0	0	318	2212
5	neg	60874	NA	1368	458	0	0	0	0	0	0	43752
6	neg	38312	NA	2130706432	218	0	0	0	0	0	0	9128
7	neg	14	0	6	NA	0	0	0	0	0	0	1202
8	neg	102960	NA	2130706432	116	0	0	0	0	0	0	2130
9	neg	78696	NA	0	NA	0	0	0	0	0	0	458
10	pos	153204	0	182	NA	0	0	0	0	0	11804	684444
11	neg	39196	NA	204	170	0	0	0	0	0	0	4352
12	neg	45912	NA	0	454	0	0	0	0	0	0	2106
13	neg	2104	NA	36	26	0	0	0	0	9744	13148	98310
14	neg	118950	NA	1390	1298	0	0	0	0	0	0	40932
15	neg	24416	NA	0	NA	0	0	0	0	0	0	556

3 Methods

The methods used in achieving the obtained results of the experiments made can be categorized into three categories: missing data handling methods, feature selection methods and machine learning methods.

3.1 Handling of Missing Data

Firstly, exploratory data analysis was carried out on the dataset the class label is a categorical variable and consists of two classes: positive and negative classes. It was observed that the dataset is skewed towards negative class. Figure 1 represents the skewness of the dataset.

All the independent variables are numeric variables. The dataset contains significant amount of missing values, out of 60,000 cases in the training set, there are 591 cases without the missing values, and hence the method of deleting cases with missing values is not used for this dataset at this stage. There is only one feature that is without missing value out of 170 independent features in the dataset. Figure 2 represents the missing value percentage in each feature.

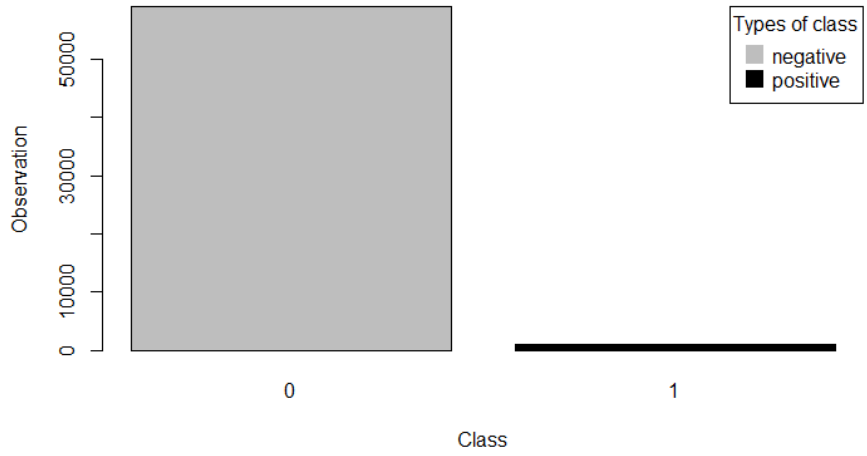


Fig. 1. A plot showing the histogram of the target feature (number of observations of each target value)

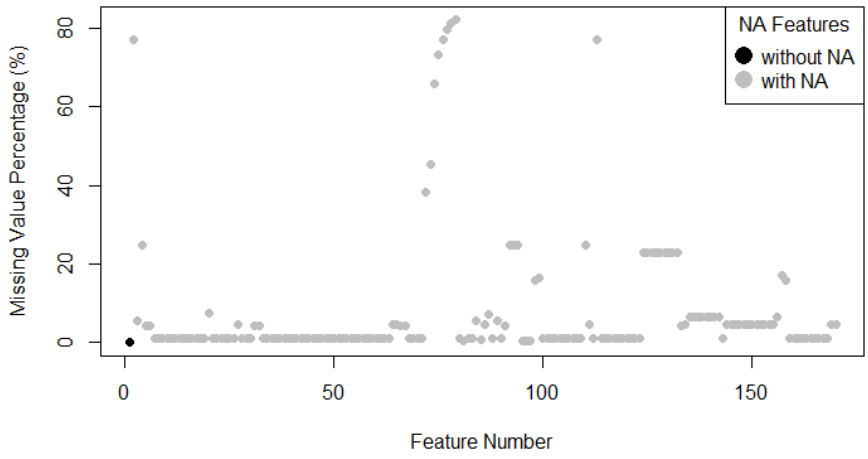


Fig. 2. A plot showing percentage of NA (missing value) in each feature before Imputation

The missing values were handled through missing at random (MAR) method. Missing at random (MAR) is one of the types of missing data mechanism and data is missing at random when the probability of the missing data on a feature Y depends on the other

observed feature(s), but not to the value of Y that should have been observed [9]. The MICE (Multivariate Imputation via Chain Equations) package in R is used to perform missing value imputation and MICE assumes that the data are missing at random (MAR) [10]. The method of MICE was set to classification and regression tree (CART) and Figure 3 represents the missing value percentage in each feature after imputation of missing values.

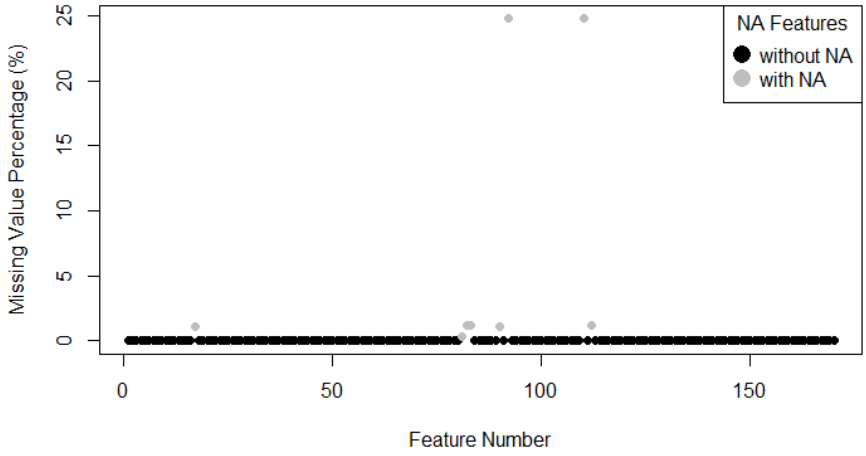


Fig. 3. A plot showing percentage of NA in each feature after Imputation

Figure 4 shows that 8 features are still having missing values after imputation of missing values were carried out, and at this stage, the deletion of cases with missing value was executed which resulted into the dataset with total of 44,667 cases for building the models. There are no missing values in the version of the dataset of this project work because cases with missing values have been removed.

3.2 Feature Selection Method

Feature selection involves choosing a k-dimensional important and relevant feature subspace from the initial d-dimensional feature space by selecting k of the original features where k is less than d and ignoring the remaining (d-k) features which are assumed to be irrelevant features or too noisy to benefit the performance of the models. We used three feature selection methods, namely: Information Gain (I.G), Random Forest (R.F) and lasso regression (L.R).

Information Gain (I.G). which is also referred to as Mutual Information (MI) measures the dependency between two variables. It can be defined as the amount of information obtained about one random variable from observing the other random

variable[11]. Feature selection is carried out on the dataset using selection of top ranking features having the highest mutual information with target variable of the dataset, and Figure 4 represents selected 94 features which are significantly better than other features for prediction of target variables.

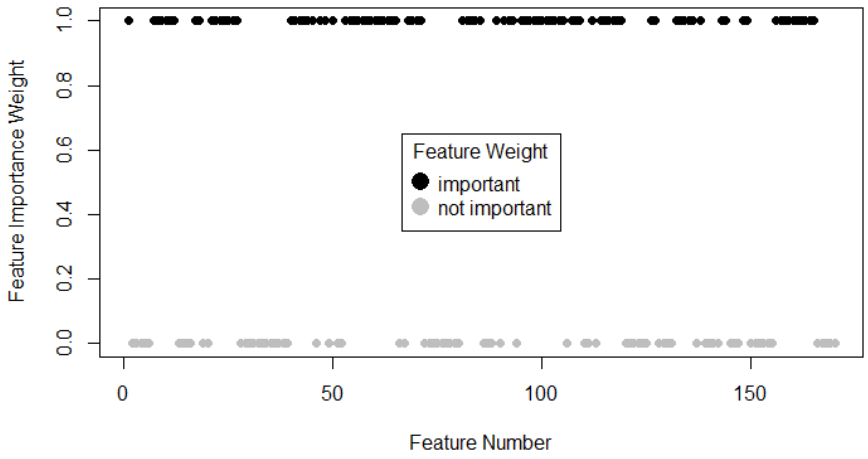


Fig. 4. A plot showing weight of features using Information Gain

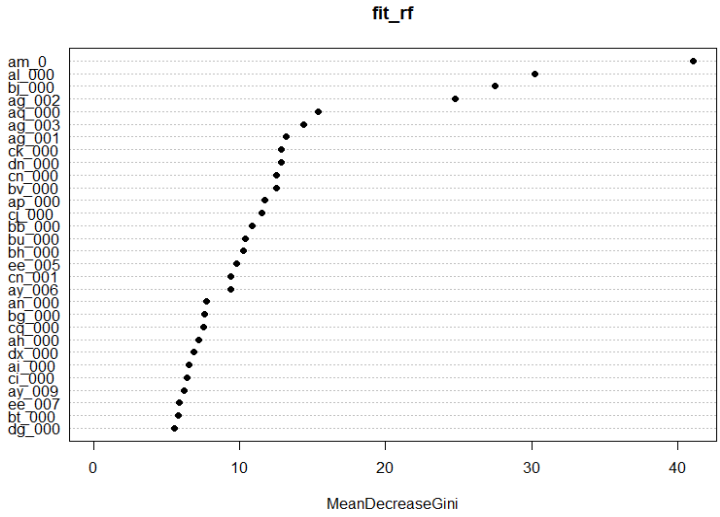


Fig. 5. A plot of important features vs level of importance using Random Forest

Random forest (R.F). can be used for a feature selection purpose and in this experiment, Gini impurity was used as the measure of impurity for choosing the best features. A threshold was applied to select features with mean decrease Gini of at least 5 and Figure 5 indicates that 30 features are important out of the total 170 features in the sense that their mean decrease Gini were the threshold.

LASSO. (Least Absolute Shrinkage and Selection Operator) is a regularization method which is used to reduce the model complexity and a powerful technique for feature selection by selecting the significant features to predict the dependent variables while shrinking the coefficients of unimportant features to zero. The Lasso features selection method produced 68 features which balance accuracy with model simplicity out of total 170 features in the dataset.

3.3 Machine Learning Methods

The following machine learning methods for classification purpose were applied: Logistic regression, Naïve Bayes classifier, K-Nearest Neighbor Searching, Support Vector Machine and Ensemble learning (Bagging and Boosting methods). For each method, four different types of models were created which are models with all the features in the dataset, and models with features selected from information gain, random forest and Lasso regression feature selection techniques. The models were built from the total sample sizes which were randomly divided into training set and test set. The total sample size was 44,667 where 70% were selected randomly as training set and remaining 30% selected as test set. The training set is used for creating the model, while the test set is used for creating model evaluation metrics which are used to determine the model predictive performance. The results obtained during the experiment are shown in Section 4.

Logistic regression. is an extension version of linear regression which is used to solve classification problems where the dependent variable is categorical variable such as pass/fail, default/not default, and win/lose [12]. Binary logistic regression model is used in this experiment because the target label is categorical variable with labelled 1 and 0. Method of glm from "caTools" library with binomial family and link value of logit was used to build Logistics regression models in this experiment.

Naïve Bayes classifier. is one of the practical Bayesians learning methods where calculation for a hypothesis is explicitly based on probabilities through the application of Bayes theorem with the fundamental assumption that each feature makes an independent contribution to the result of the outcome. Bayes theorem provides a way of calculating the probability of occurrence of an event based on the probability of another event that has already occurred. In this experiment, the method of naiveBayes from "e1071" library was used to build Naïve Bayes models with the target variable in a factor representation.

K - Nearest Neighbors (KNN). is one of the instance-based learning methods where the training dataset is stored and learning of the discriminative function is delayed and carried out until there is a new instance to be classified. When there is a new instance, a set of instances that are like the new instance are retrieved from the stored training dataset and they are used to classify the new instance. The value of K, which represents

the number of neighbors, is crucial in finding balance between overfitting and underfitting of KNN classifiers.

In this research, the centering and scaling method was used to preprocess the variables because KNN requires normalized/scaled variables. 10-fold cross validation was used and repeated 3 times, and the accuracies of resampling results against the numbers of neighbors is used to determine the optimal value of K with the highest accuracy. Cross validation method uses a small portion of the training set as validation set which is used to evaluate the performance of the KNN model under different values of K. The value of K that produces the best performance on the validation dataset is selected, and the best value of K in this experiment was 5.

Support Vectors Machine (SVM). can be used for classification task. It creates different hyperplanes that separate the data samples and amongst these different hyperplanes, it locates optimal hyperplane with maximum margin between the data samples that can accurately distinguish one class from the other class depending where the data sample is positioned on the side of the hyperplane [13].

There are basically two different categories of linearly separable SVM, which are Hard-SVM and Soft-SVM. In this part of the experiment, we implemented linear SVM and Radial SVM methods. However, we assumed that the dataset is fully or partially linearly separable because the results of the Radial SVM models were not better than the results of the Linear SVM models and the computational time of Radial SVM methods were higher than the computational time of Linear SVM methods. The resampling method of repeated cross validation, with 10-fold cross validation repeated 3 times, was used in building each Linear SVM models. The centering and scaling method were used to preprocess the variables, with tune length value of 10.

The Ensemble method produces a classifier with reduction in variance, bias and improved predictive power. There are various methods of ensemble learning and the most common two of these methods which are Bagging and Boosting methods are considered in this experiment.

Bagging technique. of ensemble learning which involves building N number of classifiers. The training samples used for each classification model is a subset of the initial training set and each subset is drawn at random with replacement from the initial training set, because of this the bagging technique is also called bootstrap aggregating. In this experiment, we used bootstrap samples (nbagg) = 25, which represents 25 subsets which are drawn randomly with replacement from the initial training set to build 25 different classifiers. The results of predictions from each of the 25 classifiers are used to provide the result of the final prediction, and this was achieved through majority voting scheme of the predictions from the 25 classifiers.

Boosting technique. of ensemble learning converts weak learners to strong learners by focusing on training of samples that are difficult to classify, and this is achieved by giving more weight to samples that were previously misclassified and reducing the weight of correctly classified samples. Weak learners, such as decision trees, are the learners that have slightly better performance evaluation metrics than random guessing. In this experiment, gradient boosting method (gbm) method was used to build boosting models with parameters such as distribution, which was Bernoulli, number of trees = 200, interaction.depth = 4, shrinkage = 0.01, and 10-fold cross validation. After

building each boosting model, cross validation method together with gbm.perf function was used to determine the optimal number of trees for predicting the accuracy of the model. Boosting technique can be an effective method of reducing bias of a model [13].

4 Results

This section involves model evaluation metrics used for comparing the machine learning models built, the machine learning methods applied, how they are applied and the results of each methods.

Model evaluation metrics [21] are used to determine the prediction performance of a model to new unseen observations. The following model evaluation metrics are considered in this experiment; accuracy, precision, recall, F1 score, and area under curve of receiver operating characteristics (AUCROC). The classification models for the experiments are binary classification where the target variable has only two classes to be predicted and straightforward explanations of evaluation metrics [21] such as accuracy, precision, recall and F1 score (see Formula 1-4) can be achieved using confusion matrix.

$$Accuracy, = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$\text{Precision} = TP/(TP+FP) \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1\ score = \frac{2*Recall*Precision}{Recall+Precision} \tag{4}.$$

Table 2 represents the performances of the different logistics regression models with their evaluation metrics, and it shows that models constructed with features from Lasso regression feature selection techniques had the highest values in terms of accuracy, precision, recall and F1score, while model constructed with features from random forest feature selection technique had the highest value in terms of AUCROC. Thus, logistic regression model demonstrated a high level of performance with models built through Lasso regression features.

Table 3 represents the performances of the different Naïve Bayes models with their evaluation metrics, and it shows that the model constructed with features from random forest feature selection techniques had the highest values in terms of accuracy, precision, and F1-score, while the model constructed with features from information gain feature selection technique had the highest value in terms of recall and AUCROC. Thus, Naïve Bayes model demonstrated a high level of performance with models built through random forest features.

Table 2. Logistics regression models

	All features (171)	Information Gain features (94)	Random Forest features (30)	Lasso Regression features (68)
Accuracy	0.9947	0.9952	0.9951	**0.9962**
Precision	0.6667	0.7176	0.7500	**0.8400**
Recall	0.5941	0.5980	0.5347	**0.6238**
F1 score	0.6283	0.6524	0.6243	**0.7159**
AUCROC	0.7959	0.8010	**0.9659**	0.9443

However, it is important to highlight that generally the precisions and F1 scores from Naïve Bayes models were low compared to the ones from logistics regression classifiers as shown in Table 2, because of a relatively great false positive number. Thus, although recalls from Naïve Bayes models were much better than that from logistics regression classifiers as shown in Table 2, precisions were clearly worse than in Table1

Table 3. Naïve Bayes models

	All features (171)	Information Gain features (94)	Random Forest features (30)	Lasso Regression features (68)
Accuracy	0.9694	0.9735	**0.9811**	0.9775
Precision	0.1828	0.2092	**0.2683**	0.2340
Recall	0.8812	**0.9010**	0.8713	0.8713
F1 score	0.3028	0.3396	**0.4103**	0.3689
AUCROC	0.9256	**0.9375**	0.9266	0.9248

.

Table 4 shows the performances of the different KNN models with their evaluation metrics, and it shows that the model constructed with features from random forest feature selection technique had the highest values in all the evaluation metrics, and in term of precision, KNN models demonstrated high level of performance compared to models from logistic regression and Naïve Bayes models.

Table 3. Results of KNN models

	All features (171)	Information Gain features (94)	Random Forest features (30)	Lasso Regression features (68)
Accuracy	0.9948	0.9957	**0.9967**	0.9963
Precision	0.8076	0.8333	**0.9014**	0.8714
Recall	0.4158	0.5445	**0.6336**	0.6039
F1 score	0.5489	0.6586	**0.7441**	0.7133
AUCROC	0.7075	0.7718	**0.8165**	0.8016

Table 5 presents the performances of different linear SVM models with their evaluation metrics, and it shows that the model constructed with features from Lasso regression feature selection technique had the highest values in terms of evaluation metrics compared to the models constructed from other features.

Table 4. Linear SVM models

	All features (171)	Information Gain features (94)	Random Forest features (30)	Lasso Regression features (68)
Accuracy	0.9962	0.9959	0.9959	**0.9965**
Precision	0.8493	0.8219	0.8405	**0.9130**
Recall	0.6138	0.5940	0.5742	**0.6237**
F1 score	0.7125	0.6896	0.6822	**0.7411**
AUCROC	0.8065	0.7965	0.7867	**0.8115**

Table 6 presents the performances of the different bagging models with their evaluation metrics, and it shows that the model constructed with features from information gain feature selection technique had the highest values in terms of evaluation metrics.

Table 5. Results of bagging models

	All features (171)	Information Gain features (94)	Random Forest features (30)	Lasso Regression features (68)
Accuracy	0.9970	**0.9974**	0.9969	0.9969
Precision	0.8604	**0.8764**	0.8488	0.8409
Recall	0.7326	**0.7722**	0.7227	0.7326
F1 score	0.7913	**0.8210**	0.7806	0.7830
AUCROC	0.8658	**0.8857**	0.8608	0.8658

Table 7 presents the performances of the different boosting models with their evaluation metrics and it shows that there is no strong difference in the evaluation metrics of boosting models constructed with all the features and with the features from the feature selection techniques.

Table 6. Results of boosting models

	All features (171)	Information Gain features (94)	Random Forest features (30)	Lasso Regression features (68)
Accuracy	0.9957	0.9956	**0.9957**	0.9954
Precision	0.8437	0.8115	**0.8666**	0.8030
Recall	0.5346	**0.5544**	0.5148	0.5247
F1 score	0.6544	**0.6587**	0.6459	0.6346
AUCROC	0.9712	0.9709	0.9715	**0.9802**

5 Conclusions

This research work involved examination of six different machine learning methods for the classification task, and these methods were logistic regression, Naïve Bayes classifier, KNN, Linear SVM, Bagging and Boosting methods of Ensemble learning.

In terms of accuracy, all the considered models performed very well, but the bagging method of ensemble learning had the highest performance in term of accuracy with the accuracy of 99.74% and followed by KNN model with the accuracy of 99.67%. In terms of precision and recall which are the focus of this research work, Linear SVM model

had the highest performance with the precision value of 91.30% and followed by KNN model with the precision value of 90.14%.

The Naïve Bayes model had the highest performance in terms of recall with the recall value of 90.10% and followed by Bagging model with the recall value of 77.22%. However, the precision of Naïve Bayes models was low because of relatively great false positive number compared to true positive number and the model did poorly in terms of precision and F1 score as the evaluation metrics, thereby making the model the least performing model out of the six models that were considered.

In terms of AUCROC, all the considered models performed relatively well, but the boosting method of ensemble learning had the highest performance in term of AUCROC with the value of 0.9802 and followed by Logistics regression model with the AUCROC value of 0.9659. The F1 score evaluation metric is the harmonic mean of both precision and recall and it was used to select the best performing model out of the six models because it contributes to achieving minimum type I and type II errors where high precision and recall are contributing factors respectively. The Bagging method had the highest performance in term of F1 score with the F1 score of 82.10% and followed by KNN models.

The results of this study demonstrated the importance of feature engineering in improving the performance of the machine learning models, and the results also suggested that Ensemble learning methods are efficient in reducing variance and bias in the dataset. Based on the current results, these methods could be practically applied to the current data and industrial application. This work can be improved in the future by investigating the behavior of the other latest machine learning technique and most especially the artificial neural networks techniques.

References

1. Maintenance Technical , https://en.wikipedia.org/wiki/Maintenance_(technical), last accessed 2019/05/15.
2. British Standard Institution, “Glossary of Terms Used in Terotechnology.”, BS 3811, United Kingdom. (1993).
3. MARCUS BENGTSSON, ERIK OLSSON, PETER FUNK, and MATS JACKSON, “Technical Design of Condition Based Maintenance System: A Case Study Using Sound Analysis and Case-Based Reasoning.”, Maintenance and Reliability Conference, Knoxville, USA, (2004).
4. British Standards Institution, “Maintenance - Maintenance terminology.”, BS-EN-13306, United Kingdom. (2010).
5. ERKKI JANTUNEN, AITOR ARNAIZ, DAVID BAGLEE and LUCAS FUMAGALLI, “Identification of wear statistics to determine the need for a new approach to maintenance.”, Euro Maintenance Conference, Helsinki, Finland, (2014).
6. KOBBACY KHAIRY A.H. and MURTHY D.N. PRABHAKAR, “Complex System Maintenance Handbook.”, Springer-Verlag London Limited, (2008).
7. RICCARDO ACCORSI, RICCARDO MANZINI, PIETRO PASCARELLA, MARCO PATELLA, and SIMONE SASSI, “Data Mining and Machine Learning for Condition-based Maintenance.”, Journal of Procedia Manufacturing, Vol. 11, (2017): p.1153-1161.

8. CHRISTOPHER GONDEK, DANIEL HAFNER, and OLIVER R. SAMPSON, “Prediction of Failures in the Air Pressure System of Scania Trucks using Random Forest and Feature Engineering.”, Conference: The 15th International Symposium on IDA, (2016).
9. CRAIG K. ENDERS, “Applied Missing Data Analysis.”, The Guilford Press., USA (2010).
10. Mice Vignettes, https://www.gerkovink.com/miceVignettes/, last accessed 10-June-2019/06/10.
11. Wikipedia; Mutual Information. https://en.wikipedia.org/wiki/Mutual_information, last accessed 2019/05/12.
12. Wikipedia; Logistic regression. https://en.wikipedia.org/wiki/Logistic_regression, last accessed 2019/07/05.
13. SEBASTIAN RASCHKA, and VAHID MIRJALILI, “Python Machine Learning.”, Packt Publishing Ltd., UK (2017).
14. MICHELE ALBANO, ERKKI JANTUNEN, GREGOR PAPA, and URKO ZURUTUZA, “The Mantis Book: Cyber Physical System Based Proactive Collaborative Maintenance.”, River Publishers., DENMARK (2019).
15. Medium; Dealing with Missing Data using R. https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17, last accessed 2019/06/10.
16. TowardsDataScience; Feature Selection Using Random Forest. https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f, last accessed 2019/05/21.
17. Interpretable Machine Learning; Logistic regression. https://christophm.github.io/interpretable-ml-book/logistic.html, last accessed 2019/07/10.
18. TOM M. MITCHELL, “Machine Learning.”, McGraw-Hill Science/Engineering/Math, USA (1997).
19. JUDITH HURWITZ, and DANIEL KIRSCH, “Machine Learning for dummies.”, John Wiley & Sons, Inc., USA (2018).
20. SHAI SHALEV-SHWARTZ AND SHAI BEN-DAVID, “Understanding Machine Learning: From Theory to Algorithms.”, Cambridge University Press., USA (2014).
21. P. PERNER, U. ZSCHERPEL, C. JACOBSEN, A Comparision between Neural Networks and Decision Trees based on Data from Industrial Radiographic Testing, Pattern Recognition Letters 22 (2001), pp. 47-54.

What Do Stakeholders Expect from Analytics?

Warwick Graco

Analytics Shed, Urana Street, Jindera, New South Wales 2642, Australia

warwick.graco@analyticsshed.com

Abstract. This paper discusses what stakeholders expect from analytics. These stakeholders include owners, users, candidates, and citizens. Fifteen expectations are identified and addressed. These expectations range from privacy, security, bias, through to actuarial prediction and trust. The key messages include: trust rather than transparency is the key to having the results of analytics accepted; greater productivity would be gained from data scientists if they focused their energies on mining and modelling tasks rather the responsibilities for data extraction and curation; a variety of skill sets are required to deliver analytics capabilities; analytics is about providing results that help people to understand issues and to reach decisions; and if you provide stakeholders with what they require to perform their duties and responsibilities, they will take ownership of what is given to them and they will come back wanting improvements and enhancements to what is supplied.

Keywords: takeholder expectations, Sense making, Decision making, Analytics professionals.

1 Introduction

Analytics involves using advanced computational techniques, such as statistical and machine learning methods, to extract new knowledge from knowledge, information and data (KID). These tasks are to assist stakeholders to perform their duties and responsibilities. Data are stimuli that people perceive through their senses. Information is data that have been processed into a form that is meaningful to the recipient. Knowledge is what has been understood and evaluated by the recipient [1]. An example of analytics is using modelling and mining techniques to identify profitable customers for a supermarket chain and offering these customers marquee services to entice them to continue to purchase products from this chain.

There are a number of stakeholders, also called 'interested parties', involved in analytics. They include owners, candidates, citizens, and users.

Owners are those who either own or lease analytics capabilities. They determine whether the capabilities will be developed and employed to meet their business requirements for purposes such as to raise profits, to lower costs and/or to improve business outcomes.

Candidates can be clients, fraudsters, patients, criminals, students, or any other party who are detected by modelling solutions and/or are discovered using mining algorithms. They are the outputs of the analytics process.

Another important party is the citizens of the country. They constitute the broader community and judge how analytics are used. It is suggested that their approval and support are required for an analytics capability to progress and prosper. If the public does not like what an analytics capability delivers and how it impacts the community, then its future is likely to be tenuous. It is the citizenry of a country; usually through their elected representatives in parliament, their government regulators, and their judiciary; that exercise authority over the production and appropriate use of analytics.

Users use the results of analytics to make decisions about how the results will be employed to achieve business and other incomes. Users can be, for example, business analysts, auditors, investigators, customer service officers and nudge experts who craft treatments to guide and reinforce in small steps desirable behaviors from citizens. Users decide based on the analytics results produced, which candidates will be targeted and what treatments they will be given to attain business objectives. The objectives are usually determined by senior management of the enterprise and approved by the owners.

These stakeholders also have many expectations of analytics. This subject is addressed in Section 2. This is followed by a discussion of the requirements that arise from the expectations in Section 3. Lastly, conclusions are provided in Section 4.

2 Expectations

There is an increasing discussion in blogs, articles and other publications about issues to do with privacy, security and ethical use of KID and analytics. There is also a focus on what owners, users, candidates and citizens want from analytics in terms of results and how the results are applied. Many of these expectations are listed in Table 1 below. The remainder of this paper focuses on what stakeholders expect from analytics and how these expectations are being met.

Table 1. Expectations with KID and Analytics.

Serial	Issue	Explanation	Comments
1	Privacy	Confidentiality of candidate details and results that candidates do not want shared with other parties	There is an exception here where there is a legal obligation on all citizens to report breaches of the law to authorized parties
2	Security	Preventing unauthorized parties gaining access to candidate details and results. It covers physical, personnel, and data protection, as well as precautions against user	This issue is becoming an increasingly challenging task with the sophisticated techniques and tactics used by cyber crimi-

		error. Disruption of or interference with the physical infrastructure affects perceptions of trust. Access to personal details is a privacy issue while access to input information, algorithms, or results is a protective security issue	nals to penetrate computer systems to steal information and funds
3	Ethical Use	Using KID and the results of analytics in a manner that does not unfairly harm stakeholders	Includes detecting those who intend to harm society and its citizens
4	Bias	Using KID and analytics that does not discriminate unfairly and disadvantage different segments of a population	Includes discriminating unfairly against candidates based on criteria such as race, ethnicity, religion, education, occupation, age, income, health and geography
5	Transparency of Results	Producing results that are understandable to those who interpret and use KID and the results of analytics	Includes issues to do with explicability, interpretability and comprehensibility of results. Ideally the results should be transparent to all parties, but this can be difficult for those who do not have specialist knowledge to understand the technicalities of analytics and the results produced
6	Acceptability of Results	The results meet the needs and expectations of different parties	They do not have to be acceptable to some parties such as those who commit fraud
7	Usefulness of Results	The results assist relevant parties such as owners and users to understand candidates and reach decisions	The term 'usefulness' can have many meanings such as the reasons why candidates were identified by an analytics solution, understand the context of the candidates and how they might act if targeted for treatment and the intelligence the results pro-

			vide on what types of candidates are detected and the threats they pose
8	Validity of Results	The results measure what they purport to measure	An example is a solution that detects customers who change providers. The solution needs to demonstrate that it detects customers who churn in this manner
9	Reliability of Results	The detection/discovery solution consistently produces valid results over time	Stakeholders will lose confidence in solutions that produce variable results
10	Discrimination of the Results	The discovery/detection solution should clearly distinguish between candidates in different classifications. This is an indication of the resolution power of the solution	For example, the solution should clearly distinguish between candidates who are high risk and those who are low risk of not repaying a loan to buy a residential property
11	Accuracy of Results	The discovery/detection solution should minimize the incidence of false positives and false negatives with candidates	A false positive is a candidate who appears to be a true positive, such as a patient who has cancer when he/she does not have this disease, and a false negative is a true positive that is misclassified as a negative (i.e. classified as a patient who does not have cancer when he/she has the disease)
12	Utility of Results	The discovery/detection solution should show both the benefits and costs of the analytics solution -i.e. Utility = Benefits – Costs	Assist users to identify the optimal selection decision, based on trade-offs between costs and benefits of targeting and treating candidates and the costs of developing, maintaining, and supporting the analytics solution

13	Actuarial Prediction	Contribution that statistical and machine-learning models make to producing accurate classifications and predictions compared to those based on human judgment	It is assumed by many that human beings are superior in their judgments of candidates compared to the classifications and predictions of statistical and machine-learning models
14	Knowledge, Information and Data (KID)	KID is curated and fit for purpose: to ensure the KID is useful, accurate and valid	KID should be clean, complete and relevant to the issue being analysed
15	Trust	The ultimate test of any solution is whether stakeholders have trust in the results produced by the analytics solution. It is suggested that trust leads to confidence	If an analytics solution does not have the trust of stakeholders, it is unlikely to be supported regardless of its scientific merits

There are fifteen expectations listed in Table 1. There are likely to be others that should be added to this list. Equally there maybe expectations in the table that need further explanation to explain more fully what they entail and what their implications are. For example, the ethical aspects [2-3] of analytics (see Serial 3 in Table 1 above) covers issues such as ownership of KID, consent, and access to the algorithms that generate the KID and algorithms that extract knowledge and insights from this resource.

3 Requirements

The expectations listed in Table 1 indicate that there are many requirements that those who provide analytics solutions need to consider when determining what they will deliver to stakeholders. It is considered that only a few of these requirements are discretionary.

By discretionary it is meant that it is up to the developer to provide them if the owner does not request them. Examples of discretionary requirements include transparency and the utility of the results (see Serials 5 and 12 of Table 1). More is said about these issues shortly.

The remaining expectations are seen to be non-discretionary in that those developing analytics solutions are putting their solutions at risk if they ignore these expectations. For example, if the requirements for security are not considered, this can have obvious consequence of compromising the privacy of candidates (see Serial 1 of Table 1).

3.1 Violating Privacy

When it comes to this issue, one does not need to look further than the Cambridge Analytica scandal [4] where data on an estimated 87m Facebook users was acquired without their approval and was used by this analytics firm to identify and target voters to try to persuade them to vote for Donald Trump in the 2016 Presidential elections. The adverse publicity from this incident led to Cambridge Analytica going out of business.

3.2 Bias

Another requirement that is not always given sufficient attention by those who perform the analytic function is bias of the results. One outspoken critic of this problem with statistical and machine-learning models is Cathy O'Neil [5].

This issue is illustrated by the tendency of model developers to use available cases of, for example, fraud to develop modelling solutions. Bias is evident here because model developers do not check to see if the available fraud cases adequately and fairly represent those found in the target population. This oversight means that models can be biased towards detecting certain types of fraudsters but not all fraudsters in the target population. To illustrate, the cases detected maybe white-collar criminals defrauding publicly listed companies while those not detected can be operators of not-for-profit organizations such as charities.

This problem of bias can be reduced (note not eliminated) by stratifying the target population using criteria such as ethnicity, age, education, income, occupation and geographic location to see in which strata known fraud cases are found. If certain strata have sprinklings of these cases while others do not, it could indicate that the modelling solution is biased. This suggests that those strata with very low/zero sprinklings should be analysed more thoroughly to see if fraud cases have been missed and should be included in the examples used to develop the modelling solution.

Not only can the cases used to develop models be biased but so too can be the techniques used to analyse data. For example, clustering algorithms can be biased towards recovering different shaped clusters from data. The k-means clustering algorithms is biased towards recovering hyper-spherical shaped clusters [6] from the data analyzed.

3.3 Effectiveness of the Predictions

There is an issue in how analytics results are judged as effective by stakeholders (see Serial 13 of Table 1). The author has seen various responses by users to analytics results including those who checked the results to see if they agreed with their judgments. If there was agreement, action was taken with the candidates. The modelling results gave users reassurance that their judgments were right.

Another response was where users cherry-picked the results by selecting candidates that agreed with their deeply held beliefs or alternatively were supportive of a policy or course of action they wanted to pursue. In other words, the selected results were used for political purposes.

Unfortunately, the author has seen the results of analytics rejected by users because they either felt threatened by them or because there were political implications that necessitated that the results be shelved. Those who occupy senior positions in organizations that use analytics have to weigh both the technical merits of the results produced versus their broader economic, social and political implications. This is illustrated with the current debate about climate change in the world where there are those who argue for the need to reduce the amount of carbon dioxide in the atmosphere versus those who oppose such measures because of the increased costs to industry in making the required reductions in this gas.

A fourth reaction was where users believed in the superiority of their own judgments and experience and rejected the results provided by analytics solutions. The results of empirical research provide interesting insights into the comparative advantages of using human judgment versus those provided by actuarial prediction - i.e. the results of statistical and machine-learning models.

There are four broad discernible trends with the research into this issue. The first is that the results of actuarial prediction have been found to consistently outperform the judgments of experts in most of the reported studies [7-9] that were reviewed. The second is that the research [10-13] has also indicated that better classifications and predictions are obtained if human judgment is combined with actuarial prediction. The third [14] is that the pooled results of a collection of experts have been shown to outperform those of statistical and machine-learning models. The fourth [15] is that it has been found from recent research that the decision making involved in making classifications and predictions is more intricate and involved than has been previously assumed and that more research is required to understand what human beings and what statistical and machine-learning models contribute to these outcomes. What the results do reveal is that the judgments of individuals have not consistently outperformed those of statistical and machine-learning models.

3.4 Transparency and Trust

Many will argue that transparency [16-17] is paramount with analytics results because if stakeholders do not understand the reasons for classifications and predictions, they will not trust them and therefore will not accept them. Furthermore, they will have little or no confidence in the results (see Serials 5 and 15 of Table 1). This is a challenge because opaque models [18] are produced by those who do analytics. These are models where stakeholders are not provided the reasons for the outputs of models. An example is artificial neural networks (ANNs) [19-20]. These are viewed as ‘black boxes’ because it is difficult to work out why candidates are given their classifications. ANNs can be converted into what are called ‘grey boxes’ where, for example, a decision-tree model is ‘bolted on’ as a backend to an ANN to identify the reasons why candidates were given their classifications. Progress is also being made with image data where ANNS are being developed that perform human complex reasoning tasks to answer questions about the contents of images [21].

Another perspective with the requirement for transparency was the response of a senior lawyer at a inter departmental meeting in the Australian Federal Government

that the author attended. The lawyer stated that the results of opaque models would be acceptable if they are trusted by those who use them. This means demonstrating that the candidates, identified by opaque models, are true positives such as those showing those who will churn by, for example, changing their telephone provider. The feedback from interventions, such as audits, will indicate the extent that an opaque model is identifying these candidates. This feedback provides reassurance that the model is working and that stakeholders can trust the results.

It should also be noted that the author has seen transparent models, such as decision trees, that produced very dense and detailed business rules that were hard to unpick and interpret. This illustrates that stakeholders can also have issues understanding the results of transparent models. That is, the transparent results impose a cognitive workload that makes it difficult for stakeholders to make sense of what the model is providing.

It is also important to note that all models, regardless of what methods they use and how they are developed, have different error rates (see Serial 11 of Table 1). These need to be communicated to all interested parties so that they know how accurate the models are and what trust and hence confidence they can have in the results. If the models generate high numbers of false positives and if they impose an administrative burden on these candidates, such as the time and effort required to prove that they were not overpaid social security benefits, it can quickly undermine the confidence people have in the modelling solution.

3.5 Utility

When it comes to utility [22-24], it is surprising that few model developers provide results that indicate the economic advantages of their solutions (see Serial 12 of Table 1). The requirements with utility are twofold of (1) demonstrating that the benefits of the solution exceed its costs and (2) showing that the solution is superior in its performance compared to other alternatives that can be employed to select and treat candidates. This requires calculating both the costs and benefits of the solutions being considered.

The costs include those entailed in developing, implementing and evaluating each solution; the costs of selecting and treating candidates and the costs of the ongoing maintenance of the solutions. The benefits include both direct and indirect ones. A direct benefit is the savings achieved by the identification and treatment of fraudsters to stop them committing this offence. An example of an indirect benefit is the deterrence effects of a fraud model which, besides detecting fraudsters, deters others from carrying out this crime.

One aspect of this requirement is working out the optimal cut-off score for selecting candidates for treatments as shown in ***Figure*** 1 below. This figure indicates where the benefits of selection and treatments are optimal compared to the costs of these actions and the costs of developing, maintaining and supporting the models used for this purpose.

3.6 Other Issues

There are expectations in Table 1 that have not been addressed so far in this paper. Those that deal with security (Serial 2 in Table 1), and the KID being curated and fit for purpose (Serial 14 in Table 1) are not discussed in this paper as others have given them attention (see [25-28] for security, [29-30] for data management, [31-32] for big data, [33] for knowledge and semantic data, [34] for graph databases and [35] for knowledge bases).

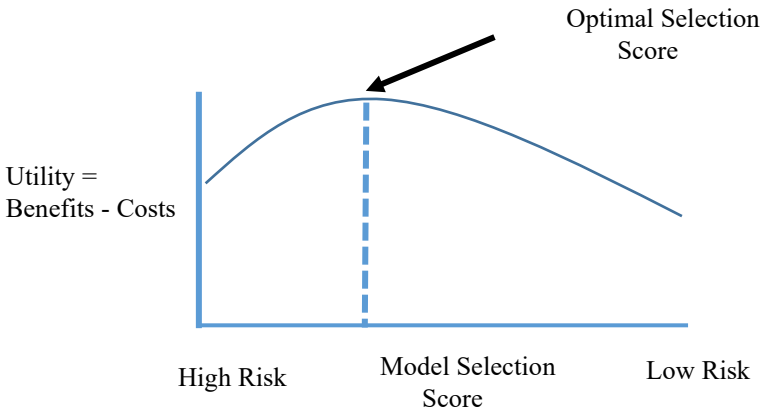


Fig. 1. Optimal Score to Select Candidates based upon Maximizing Utility

3.7 Data Wrangling

One issue with KID that is acknowledged but remains an issue with analytics is that up to 90 percent of a data scientist's time is spent doing data wrangling rather than doing mining and modelling. Data wrangling refers to the work required to extract, clean, prepare and check KID to ensure it meets the requirements of the project. This raises the question of whether data scientists should be called 'data wranglers' rather than 'data scientists'. More importantly it highlights how the productivity of data scientists is held back by the fact that they spend a disproportionate amount of their time on extraction and curation of KID when they should be spending the majority of their time mining and modelling to extract knowledge and insights from this asset.

The author has a rule of thumb that for each data scientist there should be three data analysts. Data analysts do extraction and curation of KID and data scientists check that the KID fits the requirements of the project and they do the mining and modelling. The use of three data analysts ensures there is a steady supply of KID for the data scientist to analyse. This rule increases the productivity of data scientists where their time is mostly spent on analysing KID. Robots can also be employed to assist with the extraction and curation tasks [36] thus increasing the efficiency of this process.

3.8 Measurement Issues

The need for analytical results that are valid, reliable, discriminatory and accurate are obvious requirements. Those who do model development certainly give accuracy consideration but tend not to give the same attention to ensuring that their results are valid, reliable and discriminatory (see Serials 8 to11 in Table 1). The author has not seen any data scientist report on these issues. It is usually assumed that the analytics results meet these requirements.

If reliability of classifications is taken as an example, the author has found that candidates identified by models fall into the one of the four categories shown in Table 2 below. The percentages shown are hypothetical. In practice they vary from model to model and from situation to situation.

Table 2. Positive Candidate Classification Categories

	True Positives	False Positives
Sojourners	20%	10%
Stayers	55%	15%

Sojourners, or temporary stayers, are candidates identified as positives with one run of the model and say a year later when the model is run again, they are not classified as positives. That is, they have changed from positive to negative classification. In contrast, stayers are candidates that stay positives no matter how often the model is run.

Sojourners do not warrant treatment because they are transitory cases. They will change to be negatives without any interventions by the targeting organizations. Stayers should be treated because they are classified as positives over time until they change their behaviour.

The author is of the view that reliability of an analytics model should be based on the consistency that stayers are identified over time. This means eliminating sojourners from the calculations as they are a source of error in assessing the reliability of models. This issue requires further discussion to see if a consensus can be achieved across the analytics profession about this requirement.

The approach taken by data scientists to these measurement issues differs from that of psychometricians [37-38]. Psychometricians do not neglect the requirements for accuracy, validity, reliability and discrimination when it comes to measuring behavioural attributes such as attitudes, abilities and traits. It is suggested that those who develop analytics solutions should also give these requirements due consideration. If they do this, it will give stakeholders confidence in the analytics solutions developed and that they can trust the results.

3.9 Usefulness and Acceptability of Results

The usefulness for, and so acceptability to, stakeholders of results are the overriding requirements for the successful implementation of analytics and the gaining of trust and

confidence (see Serials 6 and 7 of Table 1). In the author's experience, if these requirements are satisfied then the solution is likely to be successful. To put it simply, if the solution delivers on business outcomes and is easy and economical to use, it is likely to be acceptable to stakeholders.

3.10 Sense Making and Decision Making

One way the results of analytics can win the acceptance, trust and confidence of stakeholders is to assist them to understand issues and to reach decisions. This includes sense making which is the process of understanding the context of the problem or issue being addressed and putting it in perspective. An example is understanding candidates and their circumstances and why they were selected. It also includes decision making with an example being deciding who will be targeted and what treatments they will be given.

3.11 Challenges

There are a few challenges with these two requirements. One is how information is presented to those who make decisions. This is illustrated with three examples. The first is how information is framed. If the decision is framed as potential gains to human beings, they are more likely to be risk seeking while if the decision is framed as losses, they are more likely to be risk avoiding [39-40]. For example, a homebuyer is likely to purchase a home if the price has a high probability of doubling in the next ten years. That person is likely to be hesitant if the price is likely to halve if there is a recession in the next ten years.

The second is with the order information is presented to decision makers [41-42]. For example, decision makers can be affected more by evidence presented earlier rather than later when considering issues. This tends to happen in selection interviews where information analysed first influences these decisions more than information considered later.

The third is with 'salience effects' where information that is emphasized affects what decisions people will reach [43-44]. This refers to the fact that individuals are more likely to focus on information that is prominent and ignore that which is not. This draws attention of decision makers to issues that are striking and overriding. For example, people pay attention to sirens sounding on police cars and ambulances and make the decision to get out of the way to allow these vehicles to get through when driving on roads.

The above tendencies demonstrate that care must be exercised with the way information is presented to recipients. In some circumstances it is important to present it in a manner to encourage people to act in a responsible way such as driving safely on the roads. In other situations, it is critical to present facts objectively and impartially to ensure that decisions reached are based on the evidence.

Another issue with sense making and decision making is that decision makers differ in the way they think and act and the analytics results and KID they require to reach decisions. Some want details or minutiae, while others just want the bare essentials of

issues. Some think strategically in that they look at the big picture and the long term, while others think tactically in that they focus on the immediate situation and how this can be managed. Some think systematically, analytically and logically while others think intuitively and follow their 'gut instincts' [45-46]. A knowledge of how people think, and act can help ensure that the product or service provided is customized to everyone's idiosyncratic way of thinking and reacting to events. People are more likely to use a product or service that provides what they want and matches the way they think and operate.

3.12 Analytics Professionals

There are many analytics professionals who assist with developing solutions that meet the expectations of stakeholders, such as those listed in Table 1. Examples of these professionals are listed in Table 3.

For example, user experience architects provide products and services that please users and enable them to make informed choices if they are, for example, purchasing products. Choice architects guide candidate decision making when it comes directing citizens to act in responsible ways. Business intelligence architects ensure that users are provided with relevant, reliable and valid information to reach decisions.

Data scientists provide insights extracted by models and gleaned from mining of KID to assist recipients with comprehending issues and deciding on future action. Machine cognition scientists, machine learning scientists, cognitive scientists, knowledge scientists and robotic scientists also deal with issues to do with thinking, perceiving, judging and resolving what course of action to pursue with stakeholders. Those who perform the various engineering functions are responsible for making sure that the solutions that are deployed are well designed, are robust and reliable, and can be supported and maintained.

3.13 Multidisciplinary Teams

These examples point to the need for multidisciplinary teams made up of the right composition of analytics professionals who will develop and deliver a solution that will meet the needs and expectations of relevant stakeholders. The composition of each team will vary from project to project depending upon what skills are required and when they are needed.

Table 3. Examples of Analytics Professionals

Analytics Professionals	Explanation
User Experience Architect	Designs meaningful products and services that are intuitive, meet user needs and expectations, and are a delight to use
Choice Architect	Frames the choices presented to people especially those that will guide and reinforce desired behaviours such as paying taxes and giving up smoking cigarettes
Business Intelligence Architect	Designs the presentation of business intelligence

Data Scientist	Develops models that classify and predict and applies algorithms that discover patterns, trends and relationships in data
Data Engineer	Manages the data infrastructure and oversees designing, building, and integrating data workflows, pipelines, and the ETL process. The goal is to provide data for analysis
Machine Cognition Scientist	Carries out research and designs and develops better cognitive solutions to do with computer vision, speech processing and machine reasoning, judgment and decision making
Machine Cognition Engineer	Develops robust, reliable and supportable machine-cognition systems
Machine Learning Scientist	Carries out research and designs and develops better machine-learning solutions such as those dealing with deep learning
Machine learning Engineer	Develops robust, reliable and supportable machine-learning solutions
Cognitive Scientist	Carries out research and assists with provision of analytics solutions that assist with reasoning judgment and decision making
Cognitive Engineer	Develops robust, reliable and supportable cognitive solutions
Knowledge Scientist	Captures and represents knowledge in semantic or other symbolic form
Knowledge Engineer	Develops robust, reliable and supportable knowledge-based systems. This includes ontology engineering when it comes to semantic representation of knowledge
Robotic Scientist	Carries out research and designs and develops robotics solutions
Robotic Engineer	Develops robust, reliable and supportable robotic solutions

If the solution entails using robotics, such as a chatbot, then robotic professionals like those listed in Table **3** are needed. Similarly, if the solution requires the use of computer vision where, for example, the behaviour of customers who are contemplating buying clothes in a clothing store are observed and analysed, then the services of machine cognition specialist are relevant. If the solution requires capturing the knowledge of financial experts, then the expertise of a knowledge scientist is needed to perform this task.

An analytics project manager is essential to plan, organize and coordinate the delivery of products and services. The project manager should have well-developed people skills because this person must deal with staff and other stakeholders and their moods, needs, concerns and idiosyncrasies. The project manager should also have good management skills because he/she must orchestrate the delivery of the products and services. Project managers need to be flexible, nimble and agile to deal with project slippages and cost overruns to keep projects within budget and on time. They also must

deal with what is usually the biggest frustration with delivering outcomes and that is changes in the scope of projects.

At least one domain-knowledge expert is needed with each project. They advise on issues such as the selection of features to develop models and to mine data, guide the development of the analytics solution and whether it is providing what is needed, and interpret the results produced. A domain expert is like a navigator in that he/she helps to keep a project on track and heading in the right direction.

Each analytics team should consist of a project manager, at least one domain expert and the right mix of technical and support skills to deliver what is required. These teams can be regarded as tiger teams. These are teams of specialists brought together to work on a specific project. Members are assembled, do the project, then disband and members move to work where they are needed next on other projects.

4 Conclusion

There are five key messages in this paper. The first key message is that trust rather than transparency is the key to having the results of an analytics solution accepted. That is, transparency helps but is not essential.

The second key message is that that greater productivity would be gained from data scientists if they are relieved of the responsibilities for extraction and curation of KID and instead focused their energies on mining and modelling of these assets. This approach would provide a steady flow of insights that increase profits, lower costs, and achieve other business outcomes.

The third key message is that a variety of skill sets are required to deliver analytics capabilities and each project requires the right mix of skills.

The fourth key message is that analytics is about providing results that help people to understand issues and to reach decisions.

The fifth key message that the author learned in his long career as a data scientist is that if you provide stakeholders with what they require to perform their duties and responsibilities, and which will enhance their prestige and reputation in the eyes of their peers and their superiors, they will take ownership of the product or service. Not only will they take ownership, they will also come back wanting improvements and enhancements to the product or service. These behaviours are sure signs of the success of what was delivered.

References

1. Davis, G.B; Olson, M.H.: Management Information Systems, Conceptual Foundations, Structure and Development. McGraw Hill, New York (1985)
2. Haikowicz, S; Schleiger, E.: Artificial intelligence in Australia needs to get ethical, so we have a plan. The Conversation. 18 April (2019). https://algorithm.data61.csiro.au/artificial-intelligence-in-australia-needs-to-get-ethical-so-we-have-a-plan/
3. AI Ethics Principles. https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles
4. Kaiser, B.: Targeted. HarperCollins Publishers, London (2019)

5. O'Neil, C.H.: Weapons of Math Destruction. Crown, New York (2016)
6. Robinson, D.: K means Clustering is not a free Lunch. 16 January (2015) http://varianceexplained.org/r/kmeans-free-lunch/ and J. Keppel and S. Schmatz Anomaly Detection (Dis) advantages of k-means clustering. 4 July (2017) https://www.inovex.de/blog/disadvantages-of-k-means-clustering/
7. Meehl, P.: Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence. Echo Point Books & Media, Brattleboro Vermont (2013)
8. Grove, W.M.; Zald, D.H.; Lebow, B.S.; Snitz, B.E.; Nelson, C.: Clinical Versus Mechanical Prediction: A Meta-Analysis. Psychological Assessment. **12** (1), 19-30 (2000)
9. Kirkegaard, E.O.W.: Clinical vs. Statistical Prediction. July (2016). https://emilkirkegaard.dk/en/?p=6085
10. Yaniv, I; Hogarth, R.M.: Judgmental versus Statistical Prediction. Psychological Science. **4**, 58-62 (1993)
11. Whitecotton, S.M.; Sanders, D.E.; K.B. Norris, K.B.: Improving Predictive Accuracy with a Combination of Human Intuition and Mechanical Decision Aids. Organizational Behavior and Human Decision Processes. **76** (3), 325-348 (1998)
12. Nagar, Y: Combining human and machine intelligence for making predictions. Master of Science Thesis in Management Research, Massachusetts Institute of Technology, Sloan School of Management. (2013) https://dspace.mit.edu/handle/1721.1/82272
13. Baecke, P.; De Baets, S.; Vanderheyden, K.: Investigating the added value of integrating human judgement into statistical demand forecasting systems. International Journal of Production Economics. **191**, 85-96 (2017)
14. Farrow, D.C.; Brooks, L.C.; Hyun, S.; Tibshirani, R.J.; Burke, D.S.; Rosenfeld, R.: A human judgment approach to epidemiological forecasting. PLOS. (2017) https://doi.org/10.1371/journal.pcbi.1005248
15. Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; Mullainathan, S.: Human Decisions and Machine Predictions. Q J Econ. **133** (1), 237–293 (2018)
16. Kolyshkina, I.; Simoff, S.: Interpretability of Machine Learning Solutions in Industrial Decision Engineering. 17th Australasian Data Mining Conference. Adelaide Australia 2-5 December (2019)
17. Duez, J.: Marrying human expertise and machine intelligence to optimize decision-making. 4 February (2020) https://aibusiness.com/marrying-human-expertise-and-machine-intelligence-to-optimize-decision-making-across-the-enterprise/
18. Pontin, J.: Greedy, Brittle, Opaque, and Shallow: The Downsides of Deep Learning. Wired. 2 Feburary (2018) https://www.wired.com/story/greedy-brittle-opaque-and-shallow-the-downsides-to-deep-learning/
19. Burrell, J.: How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data and Society. 6 January (2016) https://journals.sagepub.com/doi/10.1177/2053951715622512
20. Nott, G.: How Transparent is your AI? And are 'black box' systems better? CIO. 11 January (2018) https://www.cio.com/article/3499015/how-transparent-is-your-ai-and-are-black-box-systems-better.html
21. Mascharka, D.; Tran, P.; Soklaski, R.; Majumdar, A.: Transparency by design: Closing the gap between performance and interpretability in visual reasoning. Proceedings of the IEEE conference on computer vision and pattern recognition. 4942-4950 (2018)
22. Cronbach, L.J.; Gleser, G.C.: Psychological Tests and Personnel Decisions. University of Illinois Press, Urbana IL (1965)

23. Boudreau, J.W.: Utility analysis for decisions in human resource management. In: Dunnette M.D.; Hough, L.M. (eds.) Handbook of Industrial and Organisational Psychology. vol. 2, pp621-745. 2nd ed. Consulting Psychologists Press, Palo Alto (1991)
24. Holling, H.: Utility Analysis of Personnel Selection: An Overview and Empirical Study Based on Objective Performance Measures. Methods of Psychological Research Online. **3**(1), 5-24 (1998)
25. Schneier, B.: Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World. W.W. Norton and Company, New York (2016)
26. Bejtlich, R.: The Practice of Network Security Monitoring: Understanding Incident Detection and Response. No Starch Press, San Francisco (2013)
27. Goodman, M.: Future Crimes: Inside the Digital Underground and the Battle for our Connected World. Anchor, Norwell MA (2016)
28. Data61 Editorial Team: Data privacy is a critical business need – here's way. 20 January (2020) https://algorithm.data61.csiro.au/data-privacy-is-a-critical-business-need-heres-why/
29. Wong, P.; Bennett, R.: Everything a data Scientist Should Know about Data Management. KDNuggets. (2019) https://www.kdnuggets.com/2019/10/data-scientist-data-management.html
30. Data Management. https://www.disciplinedagiledelivery.com/agility-at-scale/data-management/;
31. Marr, B.: Data Strategy. London: Kogan Page (2017)
32. Gorelik, A.: The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science.: O'Reilly Media, Sebastopol CA (2019)
33. Blumauer, A.: Explainable AI: The Rising Role of Knowledge Scientists. Forbes Magazine. 30 December (2019) https://www.forbes.com/sites/forbestechcouncil/2019/12/30/explainable-ai-the-rising-role-of-knowledge-scientists/#44f73abf603f,
34. Bechberger D.; Perryman, J.: Graph Databases in Action. Manning Publications, Shelter Island NY (2020)
35. Cagle, K.: Why Knowledge Bases are the Next Best Thing. Forbes Magazine. 10 October (2019) https://www.forbes.com/sites/cognitiveworld/2019/10/10/why-knowledge-bases-are-the-next-big-thing/#586d59bfb3f2
36. Perez, A.: How data robots will influence the future of work. 18 July (2017) https://www.itproportal.com/features/how-data-robots-will-influence-the-future-of-work/
37. Coulacoglou, C.: Saklofske, D.: Psychometrics and Psychological Assessment. Academic Press, Boston MA (2017)
38. Hambleton, R.K.; Sireci S.G.; Zumbo, B.D.: Psychometric Methods and Practices. Routledge, Abingdon on Thames UK (2020)
39. Tversky, A.; Kahneman,D,: The Framing of Decisions and the Psychology of Choice. Science. **211** (4481): 453–58 (1981)
40. Kühberger, A.; Tanner, C.: Risky choice framing: Task versions and a comparison of prospect theory and fuzzy-trace theory. Journal of Behavioral Decision Making. 23 (3): 314–29 (2010)
41. Rey, A.; Le Goff, K.; Abadie, M.; Courrieu, P.: The Primacy Order Effect in Complex Decision Making. Psychological Research (2019). https://doi.org/10.1007/s00426-019-01178-2
42. Bansback, N.; Li, L.C.; Lynd, L.; Bryan, S.: Exploiting order effects to improve the quality of decisions. Patient Education and Counseling. **96**, 197–203 (2014)
43. Louie, K.: Integrating salience and value in decision making. PNAS. **110** (40) 15853-15854 (2013)

44. Bordalo, P.; N. Gennaioli, N.; Shleifer, A.: Salience Theory of Judicial Decisions. The Journal of Legal Studies. **44** (S1), S7-S33 (2015)
45. Graco, W.J.: Management Styles of Commanders Unpublished Paper (2020)
46. Graco, W.J.: Thinking Styles of Commanders. Presented at the International Military Testing Association. The Swiss Armed Forces College Lucerne Switzerland 27 Sep – 1 Oct (2010)

Interpreting influence of feature ranking in derivation of prediction models for screening questionnaires optimization

Leona Cilar[1], Majda Pajnkihar[1], Gregor Stiglic[1,2]

[1] University of Maribor, Faculty of Health Sciences, Zitna ulica 15, 2000 Maribor, Slovenia
[2] University of Maribor, Faculty of Electrical Engineering and Computer Science, Koroska cesta 46, 2000 Maribor, Slovenia
gregor.stiglic@um.si

Abstract. Questionnaire based screening tests have been widely used in different fields ranging from healthcare and psychology to business environment. Especially by deployment of such questionnaires in the online form it is now possible to collect large amounts of screening test data that can be used to study user characteristics and apply different data mining techniques to discover new patterns or build prediction models. We used a sample of 39775 complete depression, anxiety and stress scale questionnaires collected online. In practice such questionnaires can be used to refer users to seek help from an advanced nurse practitioner specialized in mental health. Thus, modern technology enables healthcare workers to make clinical judgments based on evidence in advanced health assessment. Different data mining approaches were used to build prediction models and study user characteristics that might influence the prediction of screening test outcomes based on a limited number of questionnaire items. This study focuses on building prediction models to achieve high prediction performance by positioning of items using feature ranking. Additionally, we provide an insight into some characteristics of online screening test users using techniques to detect careless and insufficient effort responding. Selection of smaller sets of items in screening tests can significantly reduce the time needed and workload for experts and lay population using the screening tests based on questionnaires. This paper also demonstrates the possibilities of using large survey datasets to provide guidelines that can serve experts in building screening tools of the next generation.

Keywords: Data mining, feature selection, stability of prediction models, questionnaire design, screening tests.

1 Introduction

In the last decade, awareness of the importance of mental health and mental well-being has raised, nevertheless mental health problems are still under-diagnosed [1]. World Health Organization (WHO) [2] pointed out that depression is ranked as the largest contributor to global disability, followed by anxiety disorders. Moreover, there are 300 million (4.4%) people around the world suffering from depression, and nearly the same

number suffering from anxiety. Thus, psychosocial assessment such as screenings, clinical assessments and severity measurements have big importance. Mitchell [3] states that reliable assessments and measurements of psychological health are key element of supportive care. Psychological assessment can be done using screening tests, which aim to ascertain individuals who need further assessment and care. The target of screening in the field of mental health can be mood disorders, anxiety, cognitive decline, stress, and others. Nevertheless, those screenings often present additional burden to patients and healthcare workers, unless conducted at home or online.

Early detection of different disorders and illnesses is a key goal of public health strategies. It is known that early detection of disease through screening reduced mortality rates [4]. Self-report instruments provide measure of persons behavioral status, treatment and help healthcare workers in clinical decision making. Those instruments can detect symptoms of mental health problems regardless of whether they are reported or not [5]. Nowadays, many screening tests focus on identifying people who are at high risk of developing depression, anxiety or stress as those are the most prevalent mental health problems [2, 6, 7]. The Depression Anxiety Stress Scales (DASS) is a set of three self-report scales (The Depression scale, The Anxiety scale, and The Stress Scale) developed already in 1995 by Syd Lovibond and Peter Lovibond at the University of New South Wales in order to measure depression, anxiety and stress levels. Advantage of DASS is the simultaneous interaction between depression, anxiety and stress [8]. DASS and other psychological assessments are available online from 2011 on Open Psychometrics website. The site aims to educate the public about various personality tests, their uses and meaning, various theories of personality and collect data for research and develop new measures [9].

On the other hand, self-report surveys are often challenging, because of various response styles. Some styles involve exerting little effort or even insufficient effort when responding to questionnaires. Moreover, one can have invariant response style or random response style. On the other hand, participants that are exerting effort may not provide researchers with meaningful data, because they often use socially desirable responding or disingenuous responding [10]. Careless responding may appear due to lack of motivation, concentration or insufficient language skills [11]. A big challenge nowadays is also the appearance of software that generates responses to surveys which can lead to unnecessary costs and misleading findings [12]. Careless or insufficient effort responding must be detected to ensure valid findings. Using screening questionnaires to find out the prevalence of mental health disorders can overestimate prevalence and blur distinctions between low and high prevalence population [13].

DeSimone & Harms [10] suggest analyzing the data before and after screens. On the other hand, complex procedure consisting of different analytical methods from item response theory, classical test theory along with evaluation of translatability and conceptual considerations are used to identify short version of questionnaire items [14].

In Section 2 we define a problem of using prediction models to reduce the effects of careless and insufficient effort responding. Section 3 introduces the dataset used to conduct the experiments as well as the experimental setup with description of prediction models and techniques for detection of insufficient effort responders. Sections 4 and 5 explain the results with discussion and conclusions.

2 Problem Statement

Healthcare costs are rising every year, thus causing economic burden for both patients and countries. The possible solution is investing in health promotion, disease prevention, and illness management [15]. In recent years, advanced nursing practice was suggested as a must for quality care for patients in mental health services. Evidences show that mental health nurse practitioners have the potential to make a significant contribution to quality of mental health care through flexible and innovative approach [16]. Moreover, nurses must have significantly high level of knowledge and access to modern technologies to support their clinical decision making [17]. With modern technologies, we are entering a new era where smartphones, virtual reality, robotics, telemedicine and other advancements are regular part of healthcare practices. Those technologies enable evidence-based decision making, improving patient outcomes, increasing quality of healthcare, and reducing healthcare costs [18]. Moreover, computer and information systems become an indispensable part of work in different industries, including healthcare. Relatively new concept in healthcare is cognitive informatics (CI), which presents a multidisciplinary field that combines computer science, cognitive psychology, and industrial engineering. Its recent application in healthcare is leading to new methodologies in order to provide safest and most efficient way to integrate technology in healthcare [19]. Although, recent trend in CI is focusing on design and use of electronic health records (EHRs), researchers must also take into account human factors [20]. Poorly integrated healthcare technology that does not take into account human cognitive abilities can lead to a breakdown in processes and patient harm.

The aim of this study is to build and assess prediction models on limited sets of screening test items thus reducing the time and workload needed to complete the questionnaire. We also show how to improve performance of such models by considering the positioning of items and use of feature selection algorithms to rank items in screening tests. Additionally, the focus of this study includes characterizing users and emphasizing the characteristics that might introduce bias in the results.

3 Methodology

3.1 Dataset

The DASS questionnaire data used in this study was retrieved from the Open Psychometrics website [9]. The data was last updated on 14th August 2018. After the collection of the responses from anonymous Open Psychometrics users, they had to confirm their agreement with the following statement: “Your answers on this test will be stored and used for research, and possibly shared in a way that preserves your anonymity”. DASS dataset consisted of 39,775 completed questionnaires. Only questionnaire data with no missing values were included in the database. The questions were randomly shuffled for each user; therefore, the database also includes the information on position of the question in the online questionnaire. There is a long (42-items) and short (21-items) version of the questionnaire available. Each scale consists of 14 items, but they are

mixed inside the 42 items of DASS [21, 22]. In this study three 14 item scales for depression, anxiety and stress are used. Each of 42 items is rated on a four-point scale representing a frequency of that symptom in the last week. To obtain a score for each of the three scales, the answers are summed to obtain a score between 0 and 42. Fig. 1 represents the distribution of depression, anxiety and stress scores for all participants. Ceiling effect of respondents answering with all maximal values can be seen in the depression scale but is not as prevalent in the other two scales.

3.2 Experimental Setup

All experiments were performed using R programming language for statistical computing [23]. The initial dataset was split into two subsets where randomly selected 10% of data was used for initial feature ranking step. The remaining 90% of data was used to build prediction models based on the ranking of features on the 10% of data reserved for feature ranking. Permutation based feature importance as implemented by Baniecki & Biecek [24] and originally proposed by Fisher, Rudin and Dominici [25] was used in the feature ranking step. Permutation based feature importance is calculated by permuting the values of the feature of interest and observing the loss in performance caused by this permutation. Performance was measured using root mean square error (RMSE).

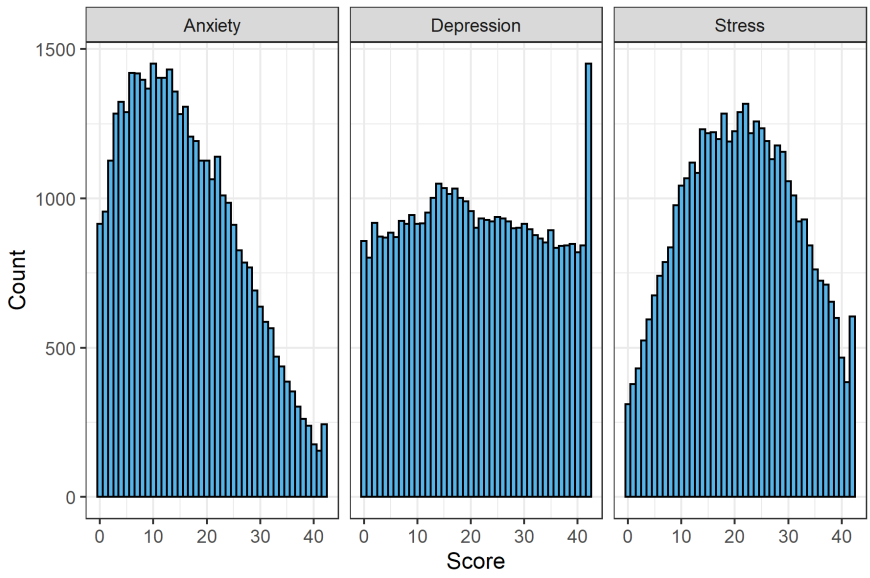


Fig. 1. Distribution of cumulative score for depression, anxiety and stress scales

Regression Models. After the feature selection step, we obtained sets of top three and bottom three ranked features as well as sets of top seven and bottom seven ranked features. The four sets of features were obtained for each of the depression, anxiety and stress scales separately. All four sets of features were used to measure the performance

of a prediction model built using top3, bottom3, top7 and bottom7 features sets. This way we simulated a scenario where extreme cases of a smaller subset of questionnaire items were used to build a prediction model predicting the final anxiety, depression or stress score. To build regression models we used *bigglm* function from the R package *glmnet* [26] version 3.0.2. to build regression models without penalization. Additionally, we used Random Forest [27] and XGBoost [28] based regression models to compare the predictive performance of conceptually different machine learning based methods.

To measure the performance of the regression models we used RMSE. As there were no missing values, the regression models were built and evaluated using 100 bootstrapping iterations on all available data (n = 35,797) to compare the performance of the top3 vs. bottom3 and top7 vs. bottom7 models. This step of experimental setup was used to find out whether feature importance ranking of the questions set in a questionnaire plays a significant role in cases where only a few questions are asked and only an approximate score estimation is required.

In the second part of the experimental setup we evaluated the performance of the regression models for different position of the questions in the questionnaire. We would like to note that questions were randomly shuffled for each user as described in 3.1 which allowed us to form subsets of data that differed in mean position of the same three/seven questions. To group users based on mean position of the questions the mean position was rounded to integer values. Subsets of data for each mean position in questionnaire were then used to build and evaluate the regression models. Finally, we calculated and visualized RMSE and 95% confidence intervals for depression, anxiety and stress subscales based on 100 bootstrap iterations.

Detection of Insufficient Effort Responders. To provide more insight into the results obtained from the prediction models described in 3.2.1, we employed two techniques used for detection of insufficient effort responders also known as careless responding detection techniques.

One of the simplest techniques to detect users which do not provide valid responses is called long string (LS) and was proposed by Johnson [29]. In LS detection we assume that a user is consistently entering the same response to consecutive questions. Originally proposed LS technique calculates the length of the longest string of consistent responses anywhere in the questionnaire. Since the data in our study allows to observe the LS calculation in relation to different position of the LS in the questionnaire due to randomly shuffled questions for each participant, we adapted the LS method to our needs. Therefore, we calculated the mean length of the same response in seven consecutive questions from each position for each participant. The results were then averaged over all participants for each starting position in the questionnaire to calculate mean LS and corresponding confidence intervals.

Another frequently used technique to detect insufficient effort responders is called individual response variability (IRV). It is usually defined [30, 31] as the standard deviation (SD) of participant's responses to all questions. However, since our aim was to detect patterns of careless responding in relation to a position of the question in the questionnaire, we adapted IRV accordingly. In our study IRV was calculated as SD of the seven consecutive questions. This allowed us to calculate the IRVs starting at different positions in the questionnaire where the position ranged from 1 to 35 in a set of 42 questions answered by each participant.

Both techniques were used to test our assumption that careless responding increases towards the end of the questionnaire when the motivation of the responders drops.

4 Results

As mentioned above, we conducted two experiments to confirm our assumptions of the question ranking importance in online questionnaire-based screening tests. In the initial experiment we used three subsets of data from a DASS dataset (n = 39,775) where the data was split in the depression, anxiety and stress subsets where each subset consisted of 14 questions that were used as features for the initial feature importance calculation. For each of the three subsets we selected top3, top7, bottom3 and bottom7 sets of features that were selected based on the 10% random sample of questionnaires.

4.1 Prediction Models

Initially the four feature sets (top3, top7, bottom3, bottom7) were used to build four models on each of the three datasets using 100 bootstrap iterations. The results provided as mean RMSE with corresponding 95% confidence intervals are presented in Table 1. To some extent the results might be surprising as both ensemble classifiers performed worse than multiple linear regression, but it needs to be noted that top and bottom sets were selected based on multiple linear regression permutation-based feature importance. Additionally, the performance of both, Random Forests and XGBoost could be optimized by cross-validation based tuning of the parameters. However, our aim was not to compare different machine learning approaches, but to show that it is important how we rank questions in the screening tests. The only combination where the prediction model based on the bottom ranked questions was bottom7 for stress prediction. The differences between top and bottom sets in RMSE were small in both, three and seven question predictive models. This indicates that in stress scale there might be much more questions with high influence on predictive performance than in anxiety or depression scales.

Based on the results in Table 1 it is evident that ranking questions based on permutation-based feature importance improves predictive performance. Some more sophisticated feature ranking methods might also be applied, but this was not the focus of this study. We were more interested in observing what is the effect of presenting the top ranked features at the beginning of the online screening test vs. displaying them to the user towards the end of the screening process.

Table 1. Mean RMSE of linear regression, Random Forests and XGBoost based prediction models on depression, anxiety and stress datasets

		Depression	
	Linear regression	Random Forest	XGBoost
Top3	4.926 (4.920-4.931)	4.973 (4.968-4.979)	4.913 (4.907-4.918)
Bottom3	5.833 (5.827-5.840)	5.935 (5.930-5.941)	5.823 (5.817-5.829)
Top7	2.918 (2.915-2.920)	2.989 (2.986-2.992)	3.024 (3.021-3.027)
Bottom7	3.444 (3.441-3.448)	3.527 (3.523-3.531)	3.579 (3.574-3.583)
		Anxiety	
	Linear regression	Random Forest	XGBoost
Top3	5.132 (5.127-5.137)	5.454 (5.449-5.460)	5.124 (5.119-5.130)
Bottom3	5.454 (5.449-5.460)	5.632 (5.626-5.639)	5.416 (5.411-5.421)
Top7	3.454 (3.450-3.458)	3.518 (3.514-3.521)	3.546 (3.542-3.551)
Bottom7	3.837 (3.833-3.840)	3.906 (3.902-3.909)	3.914 (3.911-3.918)
		Stress	
	Linear regression	Random Forest	XGBoost
Top3	5.132 (5.127-5.137)	5.216 (5.210-5.221)	5.124 (5.119-5.130)
Bottom3	5.454 (5.449-5.460)	5.632 (5.626-5.639)	5.416 (5.411-5.421)
Top7	2.889 (2.886-2.892)	3.000 (2.997-3.003)	3.021 (3.018-3.025)
Bottom7	2.828 (2.824-2.831)	2.928 (2.924-2.930)	2.945 (2.941-2.948)

Therefore, all four sets of features were also used to build prediction models on different subsets of questionnaires with different mean position of the questions. Figure 2 presents the RMSE results with corresponding 95% confidence intervals for mean position of the questions ranging from 10 to 32.

From Figure 2 we can observe higher variance of the predictive performance in models built on samples with extremely low or extremely high mean position of the questions due to smaller sample size in those subsets. Expectedly, it can be observed that top3- and top7-based prediction models performed better than prediction models based on bottom sets of features. However, especially in top3 vs. bottom3 prediction models it can be observed that position of the question in a screening test plays an important role for top ranked features. In other words, if the screening test consists of many questions, the gain of predicting the score from a small set of top-ranked questions is much higher if we present those questions to the user immediately at the beginning. The difference between the top3 and bottom3 based classifier was the highest in case of depression prediction, especially when the questions were asked at the beginning of the screening test. Again, we show that ranking questions can play an important role and can provide very accurate results already after the top 3 questions are answered.

4.2 Insufficient effort responding

In the second set of experiments, we focused on demonstration of the importance of providing highly influential questions at the beginning of the screening test where we assume the motivation and concentration of the users is still high. We measured mean LS and mean SD of the responses for any seven consecutive questions starting at different position in the screening test. The position ranged from 1, where the first seven

questions were considered, up to 35, where the last seven questions were used to calculate both measures.

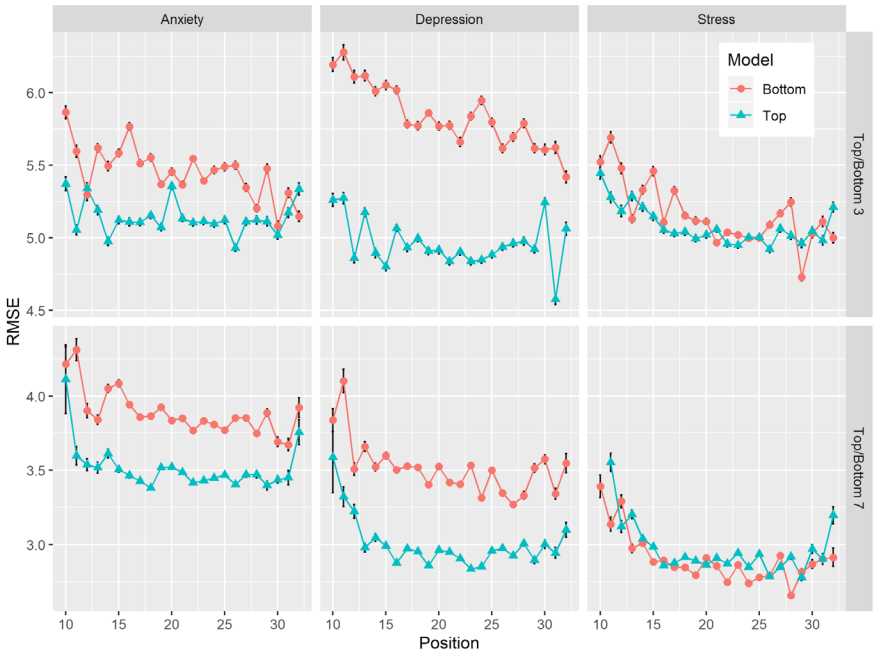


Fig. 2. Mean RMSE of linear regression for different mean position of the top3, top7, bottom3, bottom7 sets of features using 100 bootstrap iterations.

Figure 3 shows clear evidence that users provided significantly longer sequences of the same answer towards the end of the questionnaire. Moreover, the most significant increase in the mean length of the LS can be observed in the first 10 questions where mean LS increases from 2.973 (2.958-2.987) to 3.279 (3.262-3.295) representing an increase of 10.3%.

Similar to the increasing trend of LS, we can observe the opposite trend in SD of seven consecutive responses as shown in Figure 3. Again, the drop in SD is more significant in the initial 20 questions where it drops from 0.799 (0.796-0.801) to 0.753 (0.749-0.756) representing a decrease of 5.8%. Both measures show that as the users progress through the questions they tend to provide more equal responses and consequently there is less variance in responses. In our case this makes a task of predicting the score of depression, anxiety or stress score simpler, but leaves a question of reliability of data containing so many insufficient effort responses open.

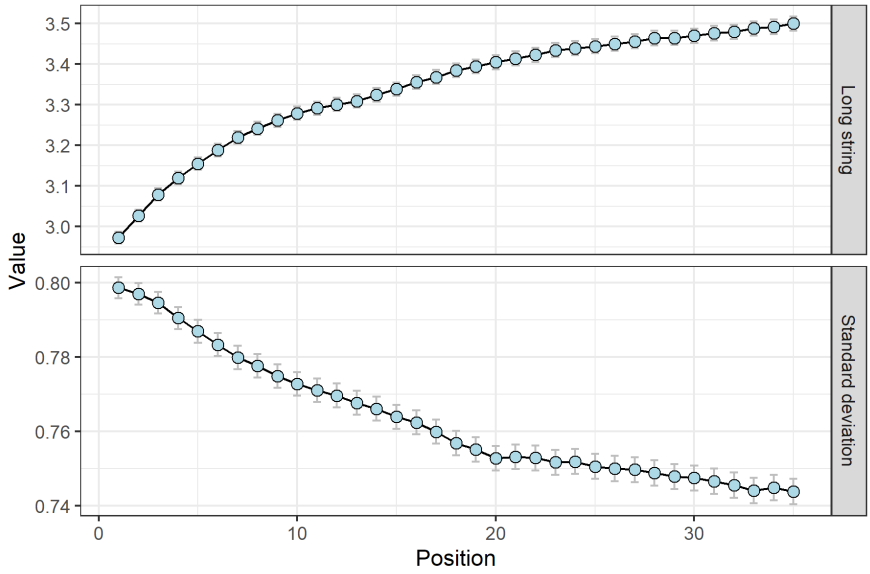


Fig. 3. Mean Long string and Standard deviation measures in relation to a position of the question in a screening test.

5 Conclusions and Future Work

In this paper we studied a large database from an online screening test for depression, anxiety and stress. We showed that it is possible to build prediction models that allow accurate prediction of all three mental health condition screening test scores by using only a limited set of questions instead of the whole questionnaire. Using smaller sets of items in screening tests can significantly reduce the time needed to get the result of the screening. As we show in this study, the results can be improved by using feature ranking techniques. Furthermore, we also demonstrated that screening tests consisting of many items reduce the quality of data. By analyzing the results from more than 39,000 questionnaires, we demonstrated that motivation and concentration of the users drop when they progress through the questionnaires with large number of items.

When observing the prediction performance in relation to the position of the items in the screening test, we noticed an interesting paradox – i.e. when more careless responding is present (towards the end of the screening test) the predictive performance of the models is actually better than using data where less careless responding is present. However, this can be explained by the fact that careless responding introduces more correlation between items in the test and consequently less features are needed to achieve good predictive performance. It needs to be noted that predictive performance results in such cases include bias introduced by careless responding.

The open problem for the future therefore remains how to reduce the bias introduced by careless or insufficient effort responding. Even with the large datasets collected it

remains a challenge to assess the quality of data, especially as we are usually not provided with expert assessment of the users' health condition when the screening tests are performed online. Techniques that allow interpretable solutions and integrate feature pre-selection [32] might represent one of the directions to solve similar problems. This study provides some initial approaches in the assessment of using prediction models to predict the outcomes of the screening tests much earlier and with less workload for the user. At the same time, we emphasize the importance of detecting insufficient effort responders and their influence on the predictive performance when the above-mentioned approach is used.

Acknowledgment

This work was supported by the Slovenian Research Agency grants ARRS-N2-0101 and ARRS-P2-0057.

References

1. Poulsen, K.M., Pachana, N.A., McDermott, B.M.: Health professionals' detection of depression and anxiety in their patients with diabetes: the influence of patient, illness and psychological factors. J Health Psychol 21(8), 1566-1575 (2016).
2. World Health Organization.: Depression and Other Common Mental Disorders: Global Health Estimates. World Health Organization, Geneva (2017).
3. Mitchell, A.J.: Screening for Psychosocial Distress and Psychiatric Disorders in Medicine: From Concepts to Evidence. In: Grassi, L., Riba, M., Wise, T. (eds.) Person Centered Approach to Recovery in Medicine. Integrating Psychiatry and Primary Care. Springer, Cham (2019).
4. Perrier, M.-J., Martin Ginis, K.A.: Narrative interventions for health screening behaviours: A systematic review. Journal of Health Psychology 22(3), 375–393 (2017).
5. Valente, S.M., Saunders, J.: Screening for Depression & Suicide: Self-Report Instruments that Work. Journal of Psychosocial Nursing and Mental Health Services 43(11), 22-31 (2005).
6. Ritchie, H., Roser, M., https://ourworldindata.org/mental-health, last accessed 2020/2/11.
7. Charlson, F., van Ommeren, M., Flaxman, A., Cornett, J., Whiteford, H., Saxena, S.: New WHO prevalence estimates of mental disorders in conflict settings: a systematic review and meta-analysis. The Lancet 394(10194), 240-248 (2019).
8. Lee, D.: The convergent, discriminant, and nomological validity of the Depression Anxiety Stress Scales-21 (DASS-21). Journal of Affective Disorders 259, 136-142 (2019).
9. Open-Source Psychometrics Project, https://openpsychometrics.org, last accessed 2020/2/11.
10. DeSimone, J.A., Harms, P.D.: Dirty Data: The Effects of Screening Respondents Who Provide Low-Quality Data in Survey Research. J Bus Psychol 33, 559-577 (2018).
11. Frany, G.: Identifying careless responders in routine outcome monitoring data. Universiteit Leiden, Leiden (2016).
12. Dupuis, M., Meier, E., Cuneo, F.: Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. Behavior Research Methods 51, 2228-2237 (2019).

13. Thombs, B.D., Kwakkenbos, L., Levis, A.W., Benedetti, A.: Addressing overestimation of the prevalence of depression based on self-report screening questionnaires. CMAJ 190(2), 44-49 (2018).
14. Abma, F., Bjorner, J.B., Amick III. B.C., Bültmann, U.: Two valid and reliable work role functioning questionnaire short versions were developed: WRFQ 5 and WRFQ 10. Journal of clinical epidemiology 105, 101-111 (2019).
15. Potter, P.A., Griffin Perry, A., Stockert, P.A., Hall, A.M.: Fundamentals of Nursing. 9th ed. Elsevier (2017).
16. Tranter, S., Robertson, M.: Improving the physical health of people with a mental illness: holistic nursing assessments. Mental Health Practice, Jan 9;23(1) (2020).
17. Ricard, N., Page, C., Laflamme, F.: Advanced nursing practice: a must for the quality of care and mental health services. Sante Ment Que 39(1), 137-157 (2014).
18. Wulfovich, S., Meyers, A.: Introduction to Digital Health Entrepreneurship. In: Wulfovich, S., Meyers, A. (eds.) Digital Health Entrepreneurship. Health Informatics. Springer, Cham (2020).
19. Hettinger, A.Z., Hoffman, D.J., Weldon, D.L.M., Blumenthal, H.J.: Cognitive informatics in healthcare. Clinical Engineering Handbook, 887–890 (2020).
20. Chen, E.T.: Examining the Influence of Information Technology on Modern Health Care. In: Management Association, Information Resources. Data Analytics in Medicine: Concepts, Methodologies, Tools, and Applications, 1943-1962 (2020).
21. Lovibond, P.F., Lovibond, S.H.: The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behaviour Research and Therapy 33, 335-343 (1995).
22. Psychology Foundation of Australia, http://www2.psy.unsw.edu.au/dass/, last accessed 2020/2/11.
23. R Development Core Team.: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2005).
24. Baniecki, H., Biecek. P.: modelStudio: Interactive Studio with Explanations for ML Predictive Models. Journal of Open Source Software 4(43),1798 (2019).
25. Fisher, A., Rudin, C., Dominici, F.: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research 20(177), 1-81 (2019).
26. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33(1), 1 (2010).
27. Breiman, L.: Random forests. Machine learning 45(1), 5-32 (2001).
28. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 785-794 (2016).
29. Johnson, J. A.: Ascertaining the validity of individual protocols from web-based personality inventories. Journal of Research in Personality 39, 103-129 (2005).
30. Dunn, A.M., Heggestad, E.D., Shanock, L.R., Theilgard, N.: Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. Journal of Business and Psychology 33(1), 105-121 (2018).
31. Curran, P.G.: Methods for the detection of carelessly invalid responses in survey data. Journal of Experimental Social Psychology 66, 4-19 (2016).
32. Perner, P. Improving the accuracy of decision tree induction by feature preselection. Applied Artificial Intelligence, 15(8), 747-760 (2001).

Predicting Inactive Users in Telecom

Cong Dan Pham[0000−0001−5184−640X], Phi Hung Nguyen, Xuan Vinh Chu, Van Hung Trinh, and Duc Hai Nguyen

Viettel High Technology, Viettel Group, Hanoi, Vietnam
danpc@viettel.com.vn, hungnp22@viettel.com.vn, vinhcx2@viettel.com.vn, hungtv7@viettel.com.vn, haind13@viettel.com.vn

Abstract. In recent years, artificial intelligence has invaded a myriad of industrial fields to improve labor productivity and service quality. In telecom, there have been many researches on churn prediction. In our paper, we propose a new approach called "inactivity user prediction". We define what "inactivity" is and treat the problem as a classic machine learning problem that deals with binary classification. To solve this problem, we use Gradient-Boosted Forests to find customers whose consumption is likely to reduce. These findings may aid in decreasing churn rate and tailoring campaigns in response to changes in consumption.

Keywords: Inactivity user· churn · similarity model · artificial intelligence · machine learning.

1 Introduction

Churn prediction has garnered interests from businesses in a wide range of industries, especially ones that follow the subscription business model. Service providers seek to minimize the number of customers unsubscribing by pushing appropriate customer care campaigns, whether it be discounts or personalized pricing plans. For such campaigns to be efficient at all, service providers have to successfully identify flight risk (i.e. correctly spot customers who will terminate subscriptions). In the context of telecom, such customers can be identified by their recent contacts with customer service, if any, and their usage history. Since manually singling out abnormalities in usage with business rules is both taxing and inflexible, telecom companies have looked into Machine Learning (ML) as a means of enhancing customer retention; some have made their results available to the public. We extrapolate the problem, beyond potential dropouts, to users who will consume significantly less than usual but not necessarily stop using the service altogether. This paper defines this generalized problem, which we term "Inactivity Prediction", as a traditional ML classification problem.

Let us explain the organization of this paper. In Section 2, we present related works on data science in telecom. Previous researches are concentrated on churn prediction and customer behavior analysis. In section 3, we define the problem of inactive users prediction, describing input data and used classification methods. The data engineering is present in Section 4, including feature importance and

PCA, bagging and resampling techniques. Next, simulation and results are mentioned in Section 5. We show the results on prediction and sampling methods, running time and further discussion. In Section6, we give a conclusion and open further research in the future.

2 Related works

The ML workflow for churn prediction was detailed in Shin-Yuan Hung et al. [5] and Kiran Dahiya et al. [2]. The authors outlined the main steps to create an ML model and evaluate it before deploying it in production. Both Shin-Yuan Hung and B.Q. Huang et al. [4] utilized the latest six months' transaction data including billing data, call detail records (CDR), customer care, etc. Shin-Yuan Hung compared the performance of Feed-forward Neural Network (FNN) with that of C5.0 Decision Tree (DT) while B.Q. Huang used C4.5 DT, FNN (or Multilayer Perceptron in the paper), and Support Vector Machine (SVM). Churn customers only account for a small percentage of the total number of users, which means a striking imbalance in churn vs. non-churn distribution. An imbalanced dataset may jeopardize the training process of an ML model. Furthermore, accuracy - the metric typically used in classification problems - may be misleading and of little use in this case. Therefore, our problem calls for more suitable metrics, namely Recall, Precision and F1 scores, as previously done in Saad Ahmed Qureshi et al. [6], T. Vafeiadis et al. [9], V. Umayaparvathi et al. [8]. The problem of imbalanced data was addressed in Saad Ahmed Qureshi and Hui Li et al. [11]. The authors utilized different resampling techniques to process the data before feeding it to an ML model. The imbalanced-learn Python library aided us greatly in testing different over-sampling and under-sampling methods. As we progressed in the research, the dimension of our data, i.e. the number of features, proved to be a problem and feature selection became a pressing need. Utku Yabas et al. [10] discussed the effect of reducing dimensionality. Besides "the curse of dimensionality", which essentially says too many features in data can hurt predictive power, high dimensionality also results in long sessions of waiting for an ML model to converge (if it does) and/or impractical levels of memory consumption. Therefore, in this paper, we also took several attempts at alleviating the dimensionality problem. Based on the above sources as well as a literary survey done by Bandara et al. [1], the consensus is that DT algorithms, and Forest (ensemble of DTs) algorithms by extension, are popular choices when it comes to churn prediction with ML.

Our paper mentioned one of the topics of data science in Telecom, using call detail record data (CDR). The analysis of telecom data has been studied in a wide range within Industrial Conference on Data Mining ICDM, mostly fraud detection, see [3]. The analysis of the telecom data is similar to the analysis of log file data, the work was studied in for example [7].

3 Defining the problem

3.1 How to mark a user as "inactive" or not?

Definition for "churn" is straightforward as the moment when a customer terminates a service is an objective recorded timestamp. On the other hand, there isn't one universal way to gauge usage nor one to dictate whether a drop in usage is significant. Usage can be quantified by the number of outbound SMSs, or the total cost incurred by a user, etc. In this paper, usage was measured by the total duration of both inbound and outbound calls. A customer's status would be "inactive" for a certain week if his/her total usage during that week was less than 40% of the total usage in the previous week. If a customer doesn't register any call during a week, his/her status in that week would also be "inactive", regardless of the usage in the previous week.

3.2 Format of Input Data

Data was tabulated into a matrix in which each row presents a feature vector containing relevant information about a user's usage over a window of 8 weeks. Each field in that row, or each column in the matrix, described a weekly attribute (e.g. the total number of outbound SMSs over a certain week). Our data had 71 weekly features, so a vector covering usage information over 8 weeks would be 71 x 8 = 568 fields long. Each weekly feature would show up in the vector 8 times, each time for a week in the 8-week window, as illustrated in Figure 1. Usage data in each window of 8 weeks, from week 1 to week 8, from week 2 to week 9, and so on, was represented by a matrix.

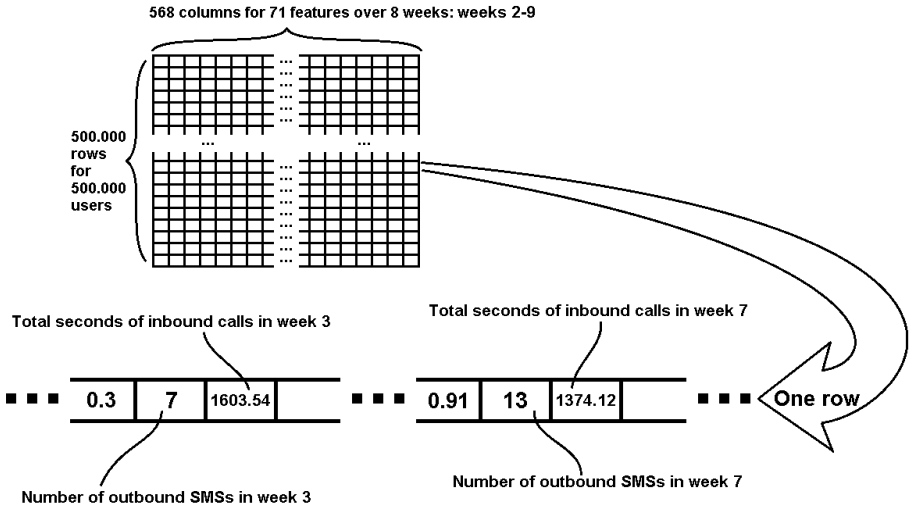


Fig. 1. Usage Data Matrix – A block of 8-week-long vectors.

3.3 Classification methods

In a ML classification problem, each feature vector has an associated label. In this case, the label of a user's feature vector over 8 weeks (e.g. week 1 to week 8) was the "inactive" status of that user in the week immediately after those 8 weeks (e.g. week 9). This status is determined by the difference between that user's usage in week 8 and his/her usage in week 9. This subtlety extends a classification problem to a forecasting problem: an ML model has to classify users when their true labels are yet to be known and only confirmed in the following week. We employed Gradient-Boosted Forest (implementation in LightGBM library provided by Microsoft) as our prediction model. Each predictor was trained with a matrix made of 4 vertically-concatenated matrices representing 4 consecutive windows of 8 weeks (e.g. weeks 1-8, weeks 2-9, weeks 3-10 and weeks 4-11). Because each of the 4 matrices has 500.000 rows for 500.000 users, the matrices used in training had 2.000.000 rows and 568 columns. The rationale for using 4 windows in training was that seasonality could plausibly manifest in the usage of some customers. Some users may exhibit biweekly tendencies, some monthly, others no cyclic behavior at all. Our training method should take this possibility into account. Each trained predictor was evaluated with each of the 4 matrices representing the next consecutive 4 windows, which allowed us to inspect how the model's reliability changed over time. Due to the imbalanced nature of the data, accuracy would tell us little about the predictive power of each model. For instance, for weeks 2-9, 28.7% of the users were labeled inactive. If a very bad model predicts all users to be active regardless of their feature vectors, it still achieves an accuracy score of 71.3%. Moreover, for the purpose of customer retention, we value correctly predicting inactive users much more than predicting active users. Mistaking an active user for an inactive one may lead to less efficient campaigns due to overestimating the target customers; however, overlooking an inactive user may lead to losing him/her as a customer, directly impacting revenue.

$$Precision = \frac{\text{Number of correctly identified inactive users}}{\text{Number of users predicted to be inactive}}$$

Precision answers the question "Given the fact our model has predicted a user to be inactive, what is the probability that user will actually be inactive?" In other words, it gauges the credibility of an "inactive" verdict.

$$Recall = \frac{\text{Number of correctly identified inactive users}}{\text{Number of all users who are actually inactive}}$$

Recall answers the question "Of all the users who are actually going to be inactive, how many have been detected by the model?" In other words, it measures how sensitive the model is to a user who's going to be inactive, or how willing it is to hand out an "inactive" verdict. The two following extreme scenarios illustrate why we need to balance precision and recall. Suppose, for both scenarios, we have 200 users, 20 of whom are truly going to be inactive. In the first scenario, we have a model that is very reluctant to label a user "inactive". If it

only gives us one "inactive" verdict and that rare verdict happens to land on a customer who's going to be inactive, the model's precision would be a perfect $1/1 = 1.0$ but its recall score would be very low as it has missed out on all the other inactive users. In fact, the recall would only be $1/20 = 0.05$. At the opposite extreme, consider a model that predicts every user as "inactive". Since all the users who are truly going to be inactive get detected by the model, its recall would be a perfect 1.0; yet such predictions would be of little value and the precision would be low, $20/200 = 0.1$ to be exact. Therefore, we relied on F1, which is the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}$$

4 Data Engineering

4.1 Reducing dimensionality with feature importance and PCA

To mitigate problems caused by the training matrices' prohibitive size, we first randomly sampled rows from those matrices to produce smaller matrices, with which predictors are trained. These predictors in turn reported on the importance of each feature, which was defined as the number a feature served as a splitting criterion in building a DT. Based on those reports, we eliminated some features, one by one, starting from the least important feature, until the predictive performance started to suffer. In the end, we removed 136 from 568 columns and were left with 432 columns, with imperceptible loss in performance. Most of the removed columns were associated with roaming usage and unusual services (e.g. loaning). The top 8 features that were consistently rated as the most important are listed in Table VI. To further reduce the dimensionality, we tried bringing our data to a lower-dimensional space with PCA (implemented by the scikit-learn Python library). We tested retaining 400, 350, and 300 components (and hence matrix columns) and recorded the changes in performance, as listed in Table II. For the sake of convenience and brevity, PCA was applied to only the training matrix composed of weeks 5-12, weeks 6-13, weeks 7-14, and weeks 8-15.

4.2 Bagging data to decrease memory consumption

The initial results in the table I demonstrated that the ML models were stable for 4 weeks after training, implying the relationship between the input vectors and the labels didn't change by any considerable amount in this time frame. The discovery that there was negligible concept drift in the customer behaviors led us to the realization we could reuse a model that was trained at least 4 weeks ago. This conclusion, in turn, inspired us to take another approach to decrease memory consumption which is bootstrap aggregating (or bagging). Instead of having all the training data flow into a single predictor, we let a group of predictors (an ensemble of Forests) work on the data, each training with only a subset

of the data. Thus, the process of sequentially training each base predictor in the ensemble did not overwhelm the memory, allowing us to extend our training data beyond 4 windows of 8 weeks. We compared the result of using 4 windows (weeks 1-8, weeks 2-9, weeks 3-10, weeks 4-11) with that of using 7 windows (those 4 windows with the addition of weeks 5-12, weeks 6-13, and weeks 7-14). To train with 7 windows of data, each base predictor received roughly 4/7 rows randomly sampled (without replacement) from each window so that the active portion of training data in memory would be the same as when training with 4 windows of data. We tested with 5, 10, and 25 Forests in the ensemble to ensure convergence (i.e. no further improvement yielded by adding more Forests to the ensemble).

4.3 Resampling imbalanced data

We tested different under-sampling and over-sampling algorithms offered by the imbalanced-learn Python library. Due to memory constraint, we resampled each 500.000-row matrix before joining them into the training matrix instead of performing the process on the whole 2.000.000-row training matrix. Again, we only resampled windows of data from weeks 5-12, weeks 6-13, weeks 7-14, and weeks 8-15. Individual models, each trained with data resampled under one of the resampling algorithms, were evaluated before the best-performing ones were put into an ensemble. Their respective impacts are presented in Table 6. The runtime each algorithm took to resample a matrix is listed in Table 8.

5 Simulation

5.1 Results

Prediction The results on prediction are presented in the Tables 1, 2, 3, 4, 5, 6, as follows:

Table 1: Evaluation Metrics for One-Forest Predictor

Train Data	Test Data	Recall	Precision	F1
Weeks 1-8 + Weeks 2-9 + Weeks 3-10 + Weeks 4-11	Weeks 5-12	0.5646	0.8528	0.6794
	Weeks 6-13	0.5545	0.8502	0.6712
	Weeks 7-14	0.5550	0.8678	0.6770
	Weeks 8-15	0.5556	0.8763	0.6800
Weeks 2-9 + Weeks 3-10 + Weeks 4-11 + Weeks 5-12	Weeks 6-13	0.5505	0.8558	0.6701
	Weeks 7-14	0.5505	0.8724	0.6750
	Weeks 8-15	0.5560	0.8783	0.6809
	Weeks 9-16	0.5833	0.8645	0.6965
Weeks 3-10 + Weeks 4-11 + Weeks 5-12 + Weeks 6-13	Weeks 7-14	0.5546	0.8707	0.6776
	Weeks 8-15	0.5582	0.8764	0.6820
	Weeks 9-16	0.5839	0.8610	0.6959
	Weeks 10-17	0.5720	0.8699	0.6902

Weeks 4-11 +	Weeks 8-15	0.5576	0.8768	0.6817
Weeks 5-12 +	Weeks 9-16	0.5858	0.8600	0.6969
Weeks 6-13 +	Weeks 10-17	0.5751	0.8684	0.6920
Weeks 7-14	Weeks 11-18	0.5794	0.8754	0.6973
Weeks 5-12 +	Weeks 9-16	0.5848	0.8600	0.6962
Weeks 6-13 +	Weeks 10-17	0.5741	0.8711	0.6921
Weeks 7-14 +	Weeks 11-18	0.5793	0.8751	0.6971
Weeks 8-15	Weeks 12-19	0.5785	0.8795	0.6980

Table 2: Evaluation Metrics Affected by Increasingly Reduced Dimensionality

Retained Principle Components	Test Data	Recall	Precision	F1
None (PCA not applied)	Weeks 9-16	0.5848	0.8600	0.6962
	Weeks 10-17	0.5741	0.8711	0.6921
	Weeks 11-18	0.5793	0.8751	0.6971
	Weeks 12-19	0.5785	0.8795	0.6980
400	Weeks 9-16	0.5177	0.8049	0.6301
	Weeks 10-17	0.5146	0.8016	0.6268
	Weeks 11-18	0.5244	0.8151	0.6382
	Weeks 12-19	0.5181	0.8148	0.6334
350	Weeks 9-16	0.5157	0.8037	0.6283
	Weeks 10-17	0.5159	0.8022	0.6280
	Weeks 11-18	0.5258	0.8127	0.6385
	Weeks 12-19	0.5123	0.8133	0.628
300	Weeks 9-16	0.5145	0.7972	0.6253
	Weeks 10-17	0.5161	0.8010	0.6277
	Weeks 11-18	0.5282	0.8170	0.6416
	Weeks 12-19	0.5129	0.8125	0.6288

Table 3: Recall Scores for N-Forest Ensembles

Test Data	1 Forest	5 Forests	10 Forests	25 Forests
Weeks 9-16	0.5854	0.5823	0.5823	0.5820
Weeks 10-17	0.5738	0.5710	0.5707	0.5703
Weeks 11-18	0.5772	0.5763	0.5756	0.5750
Weeks 12-19	0.5762	0.5755	0.5752	0.5734

Table 4: Precision Scores for N-Forest Ensembles

Test Data	1 Forest	5 Forests	10 Forests	25 Forests

Weeks 9-16	0.8609	0.8620	0.8633	0.8638
Weeks 10-17	0.8711	0.8783	0.8797	0.8801
Weeks 11-18	0.8751	0.8823	0.8872	0.8879
Weeks 12-19	0.8895	0.8900	0.8906	0.8907

Table 5: F1 Scores for N-Forest Ensembles

Test Data	1 Forest	5 Forests	10 Forests	25 Forests
Weeks 9-16	0.6969	0.6951	0.6955	0.6954
Weeks 10-17	0.6919	0.6921	0.6923	0.6921
Weeks 11-18	0.6956	0.6972	0.6982	0.6980
Weeks 12-19	0.6994	0.6990	0.6990	0.6977

Table 6: Top 8 Important Features Features

Main Balance (not counting Free Credits)
Call Duration during Weekdays
Total Call Duration
Call Duration during Weekend
Call Duration between 6am and Noon
Total Number of Calls
Cost Incurred by Calling
Call Duration between Noon and 6pm

Sampling methods Result on sampling methods, with explanations in Subsection 4.2, see the following table:

Table 7: Effects of Resampling Data on Evaluation Metrics

Resampling Algorithms		Test Data	Recall	Precision	F1
None		Weeks 9-16	0.5848	0.8600	0.6962
		Weeks 10-17	0.5741	0.8711	0.6921
		Weeks 11-18	0.5793	0.8751	0.6971
		Weeks 12-19	0.5785	0.8795	0.6980
Oversampling	SMOTE	Weeks 9-16	0.5903	0.8506	0.6970
		Weeks 10-17	0.5779	0.8540	0.6893
		Weeks 11-18	0.5822	0.8688	0.6972
		Weeks 12-19	0.5796	0.8695	0.6956
	Borderline SMOTE 1	Weeks 9-16	0.5911	0.8477	0.6965
		Weeks 10-17	0.5808	0.8548	0.6916
		Weeks 11-18	0.5814	0.8654	0.6955
		Weeks 12-19	0.5819	0.8660	0.6961

	Borderline SMOTE 2	Weeks 9-16	0.5823	0.8580	0.6938
		Weeks 10-17	0.5751	0.8598	0.6892
		Weeks 11-18	0.5773	0.8684	0.6935
		Weeks 12-19	0.5727	0.8767	0.6928
	ADASYN	Weeks 9-16	0.5898	0.8498	0.6963
		Weeks 10-17	0.5790	0.8524	0.6896
		Weeks 11-18	0.5782	0.8723	0.6955
		Weeks 12-19	0.5759	0.8722	0.6938
Undersampling	Near Miss	Weeks 9-16	0.7638	0.4269	0.5477
		Weeks 10-17	0.7619	0.4403	0.5581
		Weeks 11-18	0.7613	0.4438	0.5607
		Weeks 12-19	0.7602	0.4587	0.5722
	Edited Nearest Neighbors	Weeks 9-16	0.7540	0.6349	0.6893
		Weeks 10-17	0.7449	0.6450	**0.6914**
		Weeks 11-18	0.7551	0.6540	**0.7010**
		Weeks 12-19	0.7474	0.6624	**0.7023**
	Repeated Edited Nearest Neighbors	Weeks 9-16	0.8282	0.5377	0.6520
		Weeks 10-17	0.8205	0.5446	0.6547
		Weeks 11-18	0.8307	0.5565	0.6665
		Weeks 12-19	0.8216	0.5660	0.6703
	Neighborhood Cleaning Rule	Weeks 9-16	0.7308	0.6635	**0.6955**
		Weeks 10-17	0.7201	0.6707	**0.6945**
		Weeks 11-18	0.7312	0.6816	**0.7055**
		Weeks 12-19	0.7222	0.6855	**0.7034**
	One-Sided Selection	Weeks 9-16	0.5913	0.8533	**0.6985**
		Weeks 10-17	0.5794	0.8643	**0.6937**
		Weeks 11-18	0.5859	0.8706	**0.7004**
		Weeks 12-19	0.5838	0.8716	**0.6992**
	Tomek Links	Weeks 9-16	0.5909	0.8526	**0.6980**
		Weeks 10-17	0.5784	0.8613	0.6920
		Weeks 11-18	0.5859	0.8701	**0.7003**
		Weeks 12-19	0.5846	0.8715	**0.6997**
Ensemble of Four		Weeks 9-16	0.6585	0.7693	0.7096
		Weeks 10-17	0.6512	0.7782	0.7090
		Weeks 11-18	0.6564	0.7877	0.7161
		Weeks 12-19	0.6518	0.7926	0.7153

5.2 Running time

Testing environment: Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60Ghz with 8Gb RAM. Each LightGBM model took 5-7 minutes to converge while taking less than 5 seconds to generate predictions. Most of the default hyperparameters were untouched, save for num_leaves which was set at 512 and max_depth which was set at 8. For example, a Forest trained with a 2.000.000-row matrix of data from weeks 1-8, weeks 2-9, weeks 3-10, and weeks 4-11 took 6 minutes and 27 seconds to complete learning and 4 seconds to make predictions for 500.000 users.

Table 8: Runtime taken by each resampling algorithm (in seconds)

Resampling Algorithms		Weeks 5-12	Weeks 6-13	Weeks 7-14	Weeks 8-15
Oversampling	SMOTE	434	425	469	490
	Borderline SMOTE 1	1791	1741	1838	1849
	Borderline SMOTE 2	1756	1749	1836	1820
	ADASYN	1895	1868	1970	1937
Undersampling	Near Miss	1141	1130	1155	1112
	Edited Nearest Neighbors	4851	4939	4715	4606
	Repeated Edited Nearest Neighbors	6532	6350	6619	6531
	Neighborhood Cleaning Rule	7174	7354	4895	4835
	One-Sided Selection	7556	7559	7171	6956
	Tomek Links	6262	6257	5941	5839

5.3 Discussion

Prediction The fact call duration ranked quite high in Table VI while roaming-related features lagged behind is in line with the findings in Yihui et al. [11]. However, in their research, roaming was still a valuable factor to take into consideration while it was a dispensable one in our data. Another curious finding was that most features related to free credits stood quite low in the ranks of importance. While the effects of being more and more aggressive with PCA (i.e. keeping fewer components) are ambiguous, it is clear that the performance was severely impacted. This indicates the inherent chronological structure of the data, which is lost during the dimension reduction with PCA, was meaningful to the prediction process. As listed in Table II, as we put more predictors in the ensemble, the ensemble as a whole is stricter on giving an "inactive" verdict, resulting in lower recall. For the same reason, the sacrifice made in recall was compensated by the rise in precision. These competing factors result in the lack of a clear trend in F1 scores as more Forests were added to the ensemble. The ambivalent plateau in F1 went against our expectation that more data coverage would lead to better results. Thus, usage data covering 4 windows (and perhaps even fewer) of 8 weeks was sufficient for training ML models. Moreover, the results yielded by a Forest trained with data from only the first 4 windows and those by a Forest (or rather an ensemble of one), trained with data from

both the first 4 windows and the next 3 windows, were comparable. This observation aligned with the postulation that there was little concept drift beyond the 4 weeks after the window of time covered by the training data. Based on the evaluation metrics, we concluded that we needn't proceed with a bagging ensemble.

Sampling methods Due to the low separability between the two classes of inactive and non-inactive, artificially synthesized samples might just be as "noisy" as the original data and thus are not conducive to better predictive power. This hypothesis is confirmed by the fact results associated with oversampling algorithms were consistently the same (or slightly worse) than the result obtained without any resampling method. On the other hand, some of the undersampling methods did yield better F1 scores, mainly by striking a balance between recall and precision. Algorithms that produced at least 3 F1 scores higher than the baseline (without resampling) are underlined in table VII, with those F1 scores in bold. The ensemble of models trained with data resampled by those four algorithms attained the best F1 scores over the 4 test windows, proving that diversity of models was indeed useful in this case.

6 Conclusion

Our experiment suggested that there was a lack of concept drift in the usage data over at least a month. Thus, further research aiming to track changes in user behaviors over time should look into data that cover longer time periods. The delicate chronological structure of the data necessitates a more appropriate way to reduce dimensionality. Our preliminary results on resampling hint at a potential method to improve training data's quality and hence require deeper analysis.

Acknowledgment

We would like to extend our gratitude to Viettel Telecom for providing us with necessary resources, including data and computing units.

References

1. Bandara, W., Perera, A., Alahakoon, D.: Churn prediction methodologies in the telecommunications sector: A survey. In: 2013 International Conference on Advances in ICT for Emerging Regions (ICTer). pp. 172–176. IEEE (2013)
2. Dahiya, K., Bhatia, S.: Customer churn analysis in telecom industry. In: 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions). pp. 1–6. IEEE (2015)
3. Ferreira, P., Alves, R., Belo, O., Cortesão, L.: Establishing fraud detection patterns based on signatures. In: Industrial Conference on Data Mining. pp. 526–538. Springer (2006)

4. Huang, B.Q., Kechadi, T.M., Buckley, B., Kiernan, G., Keogh, E., Rashid, T.: A new feature set with new window techniques for customer churn prediction in land-line telecommunications. Expert Systems with Applications **37**(5), 3657–3665 (2010)
5. Hung, S.Y., Yen, D.C., Wang, H.Y.: Applying data mining to telecom churn management. Expert Systems with Applications **31**(3), 515–524 (2006)
6. Qureshi, S.A., Rehman, A.S., Qamar, A.M., Kamal, A., Rehman, A.: Telecommunication subscribers' churn prediction model using machine learning. In: Eighth International Conference on Digital Information Management (ICDIM 2013). pp. 131–136. IEEE (2013)
7. Reichle, M., Perner, P., Althoff, K.D.: Data preparation of web log files for marketing aspects analyses. In: Industrial Conference on Data Mining. pp. 131–145. Springer (2006)
8. Umayaparvathi, V., Iyakutti, K.: Attribute selection and customer churn prediction in telecom industry. In: 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE). pp. 84–90. IEEE (2016)
9. Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., Chatzisavvas, K.C.: A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory **55**, 1–9 (2015)
10. Yabas, U., Cankaya, H.C., Ince, T.: Customer churn prediction for telecom services. In: 2012 IEEE 36th Annual Computer Software and Applications Conference. pp. 358–359. IEEE (2012)
11. Yihui, Q., Chiyu, Z.: Research of indicator system in customer churn prediction for telecom industry. In: 2016 11th International Conference on Computer Science & Education (ICCSE). pp. 123–130. IEEE (2016)

Causal Inference via Conditional Kolmogorov Complexity using MDL Binning

Daniel Goldfarb[1&2] and Scott Evans[2]

[1] Northeastern University, Boston MA 02115, USA
[2] GE Research, Niskayuna NY 12309, USA
goldfarb.d@northeastern.edu

Abstract. Recent developments have linked causal inference with Algorithmic Information Theory, and methods have been developed that utilize Conditional Kolmogorov Complexity to determine causation between two random variables. We present a method for inferring causal direction between continuous variables by using an MDL Binning technique for data discretization and complexity calculation. Our method captures the shape of the data and uses it to determine which variable has more information about the other. Its high predictive performance and robustness is shown on several real-world use cases.

Keywords: Causal Inference, Minimum Description Length, Kolmogorov Complexity

1 Introduction

Kolmogorov Complexity is the length of the shortest binary program to create a given string X and measures the descriptive complexity of individual sets of data or probability distributions. The Conditional Kolmogorov Complexity, K(X|Y), is the size of the smallest program required to create string X given input Y. As described in [2] [3] [4], when $K(X) + K(Y|X) < K(Y) + K(X|Y)$ we can conclude $X \rightarrow Y$ (X causes Y). The challenge is that Kolmogorov complexity is incomputable due to the Halting problem. Methods such as compression techniques and Stochastic Complexity [1], [4] have been developed to estimate K(X) and K(X|Y). We will map feature data and their corresponding probability distributions to binary strings to determine causal features in data by estimating Kolmogorov Complexity and Conditional Kolmogorov Complexity of these strings. [2] [3] [4] have several approaches that will be leveraged. We will go beyond these approaches by using an MDL Binning technique to discretize continuous data and treat the binning techniques as the model in terms of the MDL principle. The estimated complexity is then the cost of describing random variable X using a given binning technique, and K(X) is approximated as the minimum complexity of X over all binning techniques.

The main contributions of our work are as follows:

- We present a new MDL binning technique to provide estimates of Kolmogorov Complexity of continuous data as a discrete probability distribution.

- We develop a method for using MDL binned distributions to determine conditional K(X|Y).
- We show how this approach gives improved robustness in determining causal direction over state-of-the-art techniques.

2 Prior Work

2.1 Causal Inference via Algorithmic Information Theory

Budhathoki et al. used a couple methods to infer causality using algorithmic information theory [3] [5]. He inferred the most likely causal direction between random variables by identifying lowest K-complexity. If $K(X) + K(Y|X) < K(Y) + K(X|Y)$ then $X \rightarrow Y$. These authors used a tree packing algorithm to compress binary data to compute the complexities using the MDL principle. Since the packing algorithm does not support the compression of non-binary data off-the-shelf, binarization of the data is required as a pre- processing step. Marx and Vreeken [10] developed a similar causal inference algorithm that uses the same inference for predicting causal direction. Their method uses regression to compress the data by encoding functional relations which allows for the ability to make causal inference on continuous data.

2.2 The Minimum Description Length Principle in Coding and Modeling

Barron et al. outlined the principles of MDL in a handful of applications for data compression and statistical modeling [1]. One of these applications is Density Estimation which utilizes a histogram density function to assign points to bins. Our calculation of the tradeoff between model cost and error cost falls closely with the principles in this application. We extended this method to iteratively calculate and log complexities for a variety of bin numbers to determine the minimal Kolmogorov complexity estimation.

3 Encoding A Distribution to Estimate Kolmogorov Complexity

The essence of computing the complexity of a continuous random variable is in how we discretize it via the binning technique. Hence, we will briefly outline our two proposed techniques before describing our complexity estimation algorithm. After that we will continue with the analysis of the binning techniques by comparing their performance in concise visual plots.

3.1 Definitions of Binning Techniques

In order to compute complexities, we must first find a binning assignment that is simple enough that it doesn't cost too much but still captures the essence of the distribution:

- Uniform: assign equal sized bins to span the range of points
- Greedy: iteratively add variable sized bins to minimize complexity.

These binning techniques are performed iteratively over number of bins in order to find the best binning strategy for the given sampling of points. Once the optimal technique is found, the complexity of the distribution is defined as the Kolmogorov complexity estimation given those optimal bins.

3.2 Computing Kolmogorov Complexity for Sampled Distribution X: K(X)

Our method for estimating the complexity of a random variable is as follows:

Algorithm 1 Calculate $Complexity(X, bins)$

Initialize $complexity = 0$
for $b \in bins$ **do**
  Calculate and store Shannon code for b
  complexity += 1 + $log_2(len($Shannon code for $b))$
end for
for $p \in X$ **do**
  cl = code length of bin_p
  $mean$ = avg of points in bin_p
  complexity += $log_2(cl) + log_2(|p - mean|))$
end for
return $complexity$

Algorithm 2 Calculate $K(X)$

Initialize $best\_complexity = \infty$
for $B \in$ set of possible binning strategies) **do**
  if $Complexity(X, B) < best\_complexity$ **then**
    $best\_complexity = Complexity(X, B)$
  end if
end for
return $best\_complexity$

$$K(X) := \#bins + \Sigma_{b \in bins} log(CL(b)) + \Sigma_{p \in X}[log(CL(bin_p)) + log(|p - mean(bin_p)|)]$$

We see that the calculation of complexities is split into three parts, defined as follows:

- Model Cost: Total number of bins and length of Shannon code for each bin.
- Code Length Cost: Length of bin Shannon code for each point in X.
- Error Cost: Difference between each point's value and the mean of all points in its bin.

The tradeoff, via the MDL principle, between these three components is explored in the next section.

3.3 Comparing Uniform vs. Greedy Binning Techniques

As seen in Figure 2, the greedy method finds a local optimum earlier than the uniform method, but over time the best global optimum is found by the uniform method. The greedy decisions made early on are not beneficial for binning in the long run. The dataset used in this example is a 1,000-point bimodal normal distribution with a 40/60 spread, but the same sentiment follows with toy and use case datasets used.

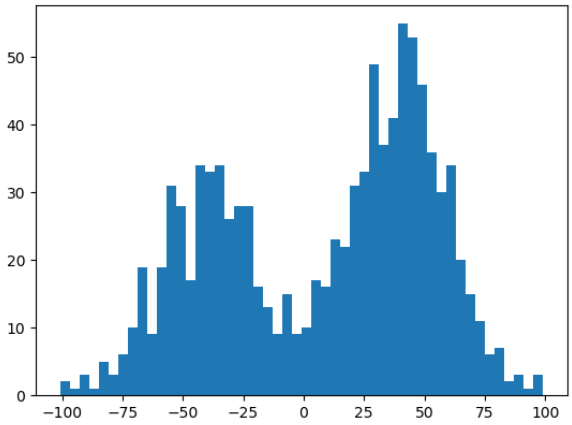


Fig. 1. Toy Bimodal Distribution

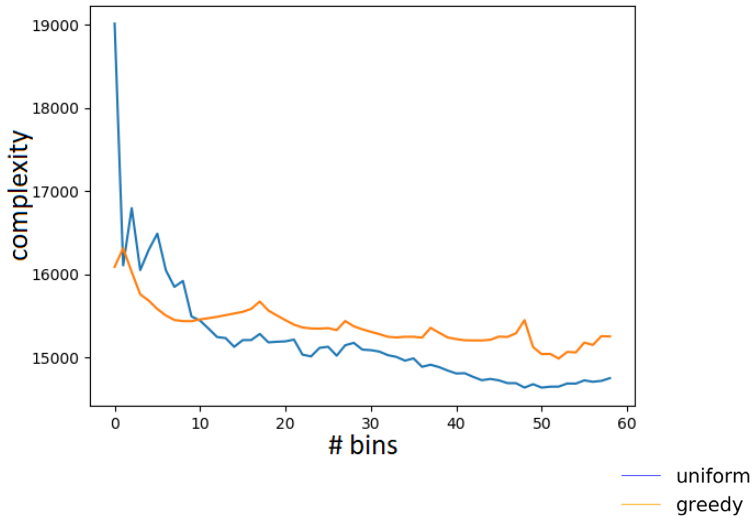


Fig. 2. Comparing the optimization of Uniform and Greedy binning techniques

Since for the uniform method, we are trying on binning for each of the n bin options and computing K-complexity takes linear time, the runtime is $O(n^2)$. However, the greedy method additionally has to test m possible partitions for each step, which results in a runtime of $O(n^2 m)$. For both of these reasons, we will be focusing on the uniform method from now on.

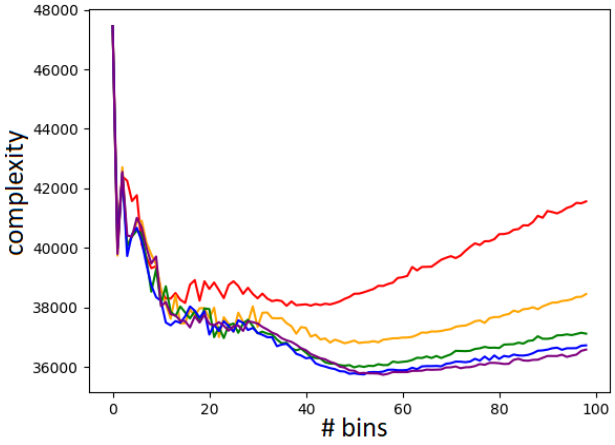


Fig. 3. Comparing the optimization of the Uniform binning technique on five different normal distribution sampling sizes

Figure 3 shows the uniform binning complexity plots for 500, 1,000, 1,500, 2,000, and 2,500 point normal samplings from top to bottom. We find that for the 1,000-2,500 point samplings, the uniform binning technique produces relatively similar optimal bin numbers (~55). However, for the 500-point distribution (in red), not enough points were sampled which resulted in a simpler representation of 20 uniform bins. We conclude that at least 1,000 points is enough to properly represent this distribution and larger sample sizes have little effect on the optimal binning strategy. We use this knowledge moving forward by narrowing our use cases to only those with > 500 rows.

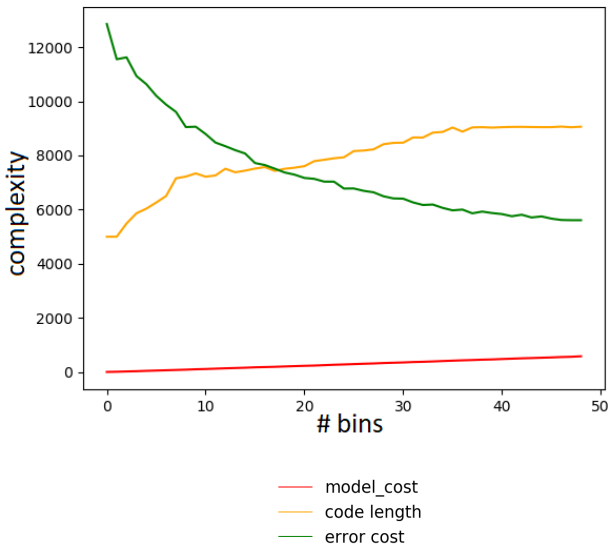


Fig. 4. Tradeoff between cost components for the normal distribution

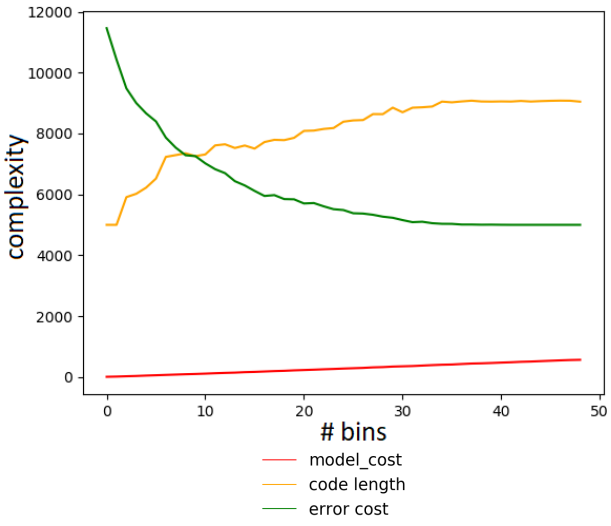


Fig. 5. Tradeoff between cost components for the skew distribution

Computing optimal complexities via the MDL principle poses a tradeoff between model cost, code length cost, and error cost. We see from various types of probability distributions in Figures 4, 5, 6 that as we increase the number uniform bins, the model cost increases to store information about the type of binning technique used, the code

length cost increases in order to represent the code lengths of a larger number of bins, and the error cost decreases since there is a lower average difference between values of points and the mean of their respective bins. As defined in subsection B, the sum of these three values are summed up and kept track of at every binning iteration. The minimum of sums is defined as the estimated Kolmogorov complexity, with optimal bins being calculated by its corresponding uniform bin number.

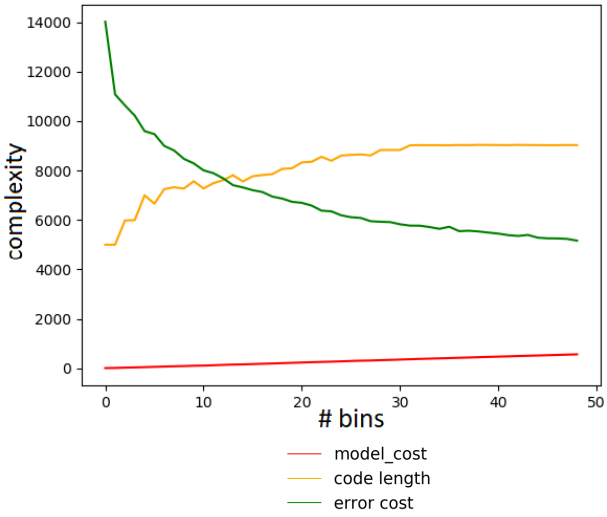


Fig. 6. Tradeoff between cost components for the bimodal distribution

4 Inferring Causality

Recall that if K(X) + K(Y|X) < K(Y) + K(X|Y) then it is most likely that X → Y. We already defined how to estimate K(X) and K(Y) so what's left is to compute K(X|Y) and K(Y|X). We find these conditional complexities in an analogous way to the original ones, however we take binning information from the conditional variables for free without incorporating the corresponding model cost.

4.1 Computing Kolmogorov Complexity for X given Y : K(X|Y)

Our method for estimating the conditional complexity of one random variable *X* given another random variable *Y* is as follows:

Algorithm 3 Calculate $K(X|Y)$

Initialize $complexity = 0$
$bins_Y$ = optimal uniform bins for Y
$bins_X$ = $len(bins_Y)$ bins scaled for X
$counts_X$ = bin counts for $bins_X$
$counts_Y$ = bin counts for $bins_Y$
while $\exists i, j$ such that $counts_Y[i] > counts_X[i]$ and $counts_Y[j] < counts_X[j]$ **do**
  $diff\_i = |counts_Y[i] - counts_X[i]|$
  $diff\_j = |counts_Y[j] - counts_X[j]|$
  $counts\_diff = min(diff\_i, diff\_j)$
  $counts_X[i]$ += $counts\_diff$
  $counts_X[j]$ -= $counts\_diff$
  $complexity$ += $1 + log_2(len(bins_X))$
end while
$shannon_X$ = Shannon codes for $counts_X$
for $p \in X$ **do**
  $mean$ = avg of points in $bins_X[p]$
  complexity += $log_2(shannon_X[p]) + log_2(|p - mean|))$
end for
return $complexity$

$$K(Y|X) := \#binsbalanced + \Sigma_{b \in bins} I(balanced)[log(CL(b)) + log(ind(b))] + \Sigma_{p \in X}[log(CL(bin_p)) + log(|p - mean(bin_p)|)]$$

The difference here is that we use the optimal bin number for Y to compute K(X) and then iteratively balance out code lengths of bins to better fit the distribution of X. The penalty for balancing out each code length is an additional model cost encoding along with the bin indices of both edited counts.

In some cases, we find that K(X|Y) > K(X) due to additional cost of indicating the indices of the new balanced bins. To follow probabilistic axioms, we define our Kolmogorov complexity estimate to be min(K(X|Y), K(X)) to follow that if Y gives no information about X, then K(X|Y) := K(X).

For our inferences, we will no longer incorporate error cost since we find that it is commonly uniformly distributed from bin to bin since they are always being compared to bin means. However, the error cost is still used when performing the tradeoff to find the optimal binning strategy. This also allows the model cost to play a bigger role in determining overall complexity which is important for our bin balancing algorithm. After we lay out some prior work to provide a better description of the area, we will be using this method on several real-world use cases and one toy example.

5 Data Selected and Use Case

Our real-world use cases are all datasets that are accessible from a handful of sources [UCI Machine Learning Repository, Econometrics Toolbox by James P. Lesage, & Kaggle] and took inspiration from two papers [5] [6]. The first is the ORIGO paper from Budhathoki et al. This paper lays out a similar algorithmic information theory method for discrete data, so we made sure to use their use cases for comparison purposes. The other paper by Mooij et al. uses a less related method for inferring cause and effect but provides ~100 different variable pair examples with intuitive ground truth from which we extracted the use cases that made sense for our method. Namely, those with > 500 rows and preferably no binary features. We also include a toy solar power use case which was the industrial application that motivated us to perform causal inference on continuous data.

It should be noted that each of the datasets are normalized to have minimum value 0, maximum value 100, and the rest of the points scaled accordingly. This detail makes sure that when computing the optimal bin strategy, error costs are bounded below by 1 and above by ceil(log2(100)) = 7. This bounding of error costs allows the scaling of our model to datasets with extremely large values so that the error costs do not overrule the model costs and code length costs when performing the iterative binning complexity tradeoff.

Table 1 provides a summary of the results when running our causal inference technique on the use cases. Both sides of the inequality are given along with the computed percent change between K(Y) + K(X|Y) and K(X) + K(Y|X) to provide the likelihood of causality. Namely, since the inequality holding implies that it is most likely that X causes Y, then a larger magnitude in difference implies a greater likelihood. In order to normalize over datasets with many points, we provide this likelihood metric in percent change.

Table 1. Causal Inference Summary Results

dataset	X	Y	$K(Y)+K(X\|Y)$	$K(X)+K(Y\|X)$	result	% change
car evaluation	safety	evaluation	24221	22527	$X \to Y$	7.5
abalone	sex	length	34939	28366	$X \to Y$	23.2
abalone	sex	diameter	35368	28366	$X \to Y$	24.7
abalone	sex	height	29226	28366	$X \to Y$	3.0
adult	education	income	348980	384909	$Y \to X$	-9.3
concrete	cement	strength	15129	15129	inconclusive	0
concrete	water	strength	15013	15258	$Y \to X$	-1.6
concrete	superplast	strength	14374	14331	$X \to Y$	0.3
county	population	employment	46261	46086	$X \to Y$	0.3
housing	rooms	value	8061	7903	$X \to Y$	2.0
toy solar	solar	power	15029	14680	$X \to Y$	2.4

Out of 11 pairwise causal examples, our method predicted 8 of them to be causal, 2 of them to be non-causal, and 1 of them as inconclusive. We came to the inconclusive result because K(Y|X) > K(Y) and K(X|Y) > K(X), so our evaluation found K(Y)+K(X|Y) = K(X)+K(Y|X). So our precision score for this probing of examples is 80% which is on-par with other continuous causal inference methods [7] [8]. This result is using a conclusivity threshold of 0 meaning that any positive percent change will predict causal, any positive percent change will predict not causal, and a 0 percent change will predict inconclusive. In addition to the high precision, we find that our inferences are robust with respect to the threshold used. Figure 7 shows that for conclusivity thresholds between 0% and 5%, precision values stay at 80±6%.

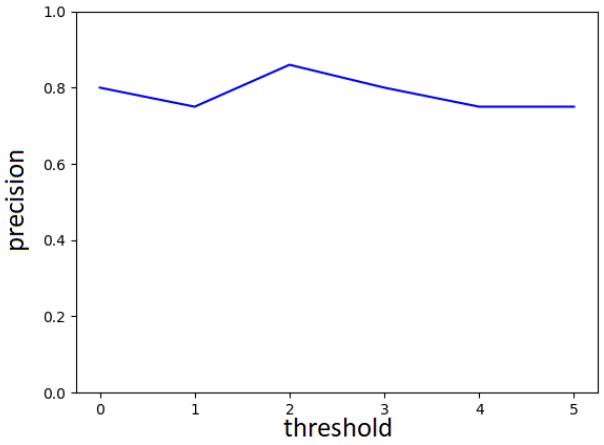


Fig. 7. Comparing the precisions of six causal inference classifiers to show the robustness of our method

We will now dig into some of the important use case examples and elaborate on the intuition behind their ground truths.

5.1 Car Evaluation Dataset

The car evaluation dataset contains 1728 rows of 6 features about used cars for sale. The labels to be predicted are the evaluation prices. The two features that we are isolating are the safety rating of the car and its evaluation label. We hypothesize that since safer cars have special technologies that they should cause higher prices. However, a higher price does not imply a safer car. Plenty of sports cars are expensive due to their fancy engines but are not necessarily safer, so this is a strictly causal and non-symmetric relationship. We were able to correctly infer this direction of causality.

5.2 Abalone Dataset

The abalone dataset contains 4177 rows of 8 features about different abalone shellfish. The labels to be predicted are the age of the abalone using number of rings on the shell as a proxy. The two features that we are isolating are the sex of the abalone and its size (length, diameter, height). It is common in other species that on average males are larger than females. We hypothesize that the same relationship holds for abalones as well. On the other hand, changing the size of an abalone to be larger does not make it more likely to be male, so this is a causal relationship. We were able to correctly infer this direction of causality.

5.3 Housing Dataset

The housing dataset contains 506 rows of 14 features about the details and neighborhoods of Boston apartments. The labels to be predicted are the values of the homes in 1, 000's of dollars. The two features that we are isolating are the number of rooms in the apartment and the value. More rooms in an apartment adds value by providing more space and accommodation. However, an increase in price does not imply more rooms in the apartment. For example, an apartment may be more expensive due to its location with respect to the city or the neighborhood it belongs to. So we infer that this is a causal relationship where number of rooms causes apartment value. We were able to correctly infer this direction of causality.

5.4 Toy Solar Power Dataset

This toy example is inspired by the industrial application of the input and output to solar panels. Each of the 1,000 points is an instance in time where the intensity and instantaneous value of power generation are logged. Given a normal distribution with a single mean and infrequent extreme intensities for X, the outputted power generation distribution, Y, is an extremely skewed normal distribution.

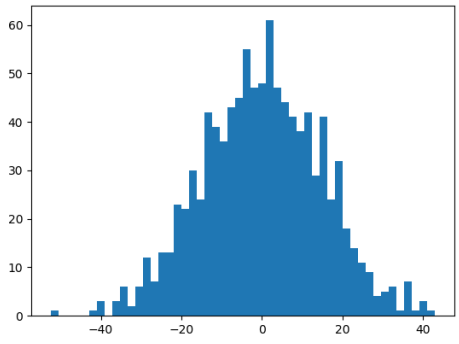


Fig. 8. Toy Solar Distribution

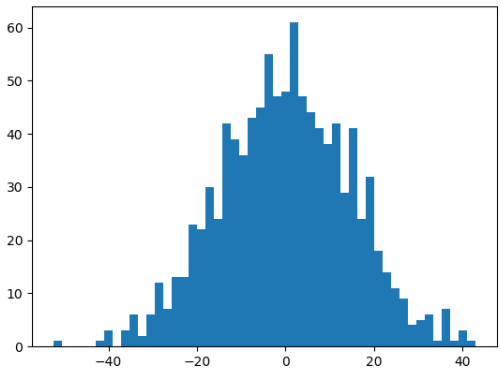


Fig. 9. Toy Power Distribution

This shape follows from the clipping behavior of power generation. Solar panels are not perfect, they have a clipping point where any more solar intensity does not increase the marginal power. As a result, we see a plateauing behavior in the power and thus a hard clip in its probability distribution. So given a level of solar intensity, we should be able to tell what the level of power generation was at that point. However, this is not possible going from Y to X. If we are given a point of power generation that is above the clipping line, it is impossible to recover what the level of solar intensity was at that instance. Hence, this is a causal relationship. We were able to correctly infer this direction of causality.

6 Conclusion

We have introduced a causal inference technique that uses the MDL binning principle to compress and compute Kolmogorov complexities for continuous data. We applied our method to numerous real-world examples with intuitive ground truths and showed competitive prediction precision against state-of-the-art methods and robustness over various conclusivity thresholds.

For future work, we are interested in applying our pairwise causal inference method to feature selection. Given a dataset X and set of labels Y, we want to extract a causal feature set such that the only features used in the new X′ have a causal relationship with Y. This in turn would produce a causal machine learning model for which we know that each feature is a causal predictor of Y.

References

1. Barron, Andrew, Jorma, Rissanen, and BinYu. ”The minimum description length principle in coding and modeling.” IEEE Transactions on Information Theory 44.6 (1998): 2743-2760.
2. Janzing, Dominik, and Bernhard Scholkopf. ”Causal inference using the algorithmic Markov condition.” IEEE Transactions on Information Theory 56.10 (2010): 5168-5194.
3. Budhathoki, Kailash, and Jilles Vreeken. ”Causal inference by compression.” 2016 IEEE 16th international conference on data mining (ICDM). IEEE, 2016.
4. Budhathoki, Kailash, and Jilles Vreeken. ”MDL for causal inference on discrete data.” 2017 IEEE International Conference on Data Mining (ICDM). IEEE, 2017.
5. Budhathoki, Kailash, and Jilles Vreeken. ”Causal inference by stochastic complexity.” arXiv preprint arXiv:1702.06776 (2017).
6. Mooij, Joris M., et al. ”Distinguishing cause from effect using observational data: methods and benchmarks.” The Journal of Machine Learning Research 17.1 (2016): 1103-1204.
7. Janzing, Dominik, et al. ”Information-geometric approach to inferring causal directions.” Artificial Intelligence 182 (2012): 1-31.
8. Peters, Jonas, et al. ”Causal discovery with continuous additive noise models.” The Journal of Machine Learning Research 15.1 (2014): 2009- 2053.
9. Marx, Alexander, and Jilles Vreeken. ”Telling cause from effect using MDL-based local and global regression.” 2017 IEEE international conference on data mining (ICDM). IEEE, 2017.

Survival Analysis of Breast Cancer Utilizing Integrated Features with Ordinal Cox Model and Auxiliary Loss

Isabelle Bichindaritz[1[0000-0003-1712-490X]], Guanghui Liu[1[0000-0002-1135-2939]], and Christopher Bartlett[1[0000-0003-3450-0786]]

[1] Department of Computer Science, State University of New York at Oswego, New York, USA
{ibichind, guanghui.liu, cbartle3}@oswego.edu

Abstract. Survival analysis has currently become an essential statistical research hotspot that models the time-to-event information with data censorship handling. Such technique has been widely used in cancer treatment and prognosis, and has been also proven to be useful for understanding the relationships between patients' variables and covariates (e.g. clinical and genetic features) and the effectiveness of various treatment options. Despite the advances in this direction, limitations still exist. In this study, we propose a novel method for survival prediction of breast cancer using bidirectional long short-term memory, ordinal Cox model network and auxiliary loss. First, we use weighted gene co-expression network analysis algorithm to reduce the gene expression data and DNA methylation data feature dimension and extract cluster eigengenes respectively. These eigengenes will be merged as input to the model. Then, we build an ordinal cox proportional hazards model for survival analysis and use long short-term memory method to predict patient survival risk. We add an adaptive auxiliary loss to the original objective to improve the ability of optimizing the learning process in training and regularization. The auxiliary loss will add extra gradient flow during back propogation, thereby helping to reduce the vanishing gradient problem for earlier layers and helping to decrease the loss of the main task. We use the cross validation method and the concordance index to evaluate the prediction performance. Stringent cross-validation tests on the benchmark dataset demonstrates the efficacy of the proposed method, which achieves very competitive performance with existing state-of-the-art methods.

Keywords: Survival Analysis, Breast Cancer, Genetics, Feature Selection, Ordinal Cox Model, Auxiliary Loss, Bidirectional Long Short-Term Memory.

1 Introduction

Breast cancer is one of the most common forms of disease worldwide. It is reported that more than forty thousand women and four hundred men in the United States died from breast cancer annually before 2016. These data emphasize the importance of a more profound understanding of factors that trigger breast cancer and contribute to its development. Gene expressions driven by many elements are frequently used as markers of breast cancer progression. Two easily measurable elements are mRNA and DNA

methylation. Both mRNA and DNA methylation levels are differentially expressed in several tissue types [20]. To understand the interaction between different types of genomic features requires more sophisticated modeling and analysis. In particular, the causal relationships between gene expression data and DNA methylation data have been extensively studied [29]. The influence of mRNAs and methylations in cancer have been introduced many times. Epigenetic regulation of mRNA via DNA methylation at CpG sites is heritable, with methylation patterns referred to as epigenetic markers. The transmission failure of somatic epigenetic markers is related to aberrant mRNA expression, leading to disease phenotypes [17]. Extensive perturbations of DNA methylation have been noted in cancer, causing changes in gene regulation that promote oncogenesis. Understanding both epigenetic changes and somatic DNA mutations show promise for improving the characterization of malignancy and predicting treatment response and prognosis [7].

One goal of long-term cancer research is to be able to identify prognostic factors that affect patients' survival time, which in turn allows clinicians to make early decision on treatment [16]. In this study, we focus on breast cancer, which is the most prevalent subtype. Consequently, to explore the utility of gene expression data and DNA methylation data for cancer diagnosis, gene expression and methylation of tumors from patients with breast cancer will be analyzed to identify potential cancer-specific survival risk. Gene expression and DNA methylation features will be used to predict survival. Existing studies have demonstrated that combining gene expression data and methylation data can better stratify cancer patients with distinct prognosis than using single signature [11]. However, these existing methods simply combining these features in series made of the set of genes and have ignored the strong ordinal relationship between the survival times of different patients. Deep learning techniques are used directly in deep survival models to learn the hazard function. These models overcome many of the restrictions of cox models like the proportionality assumption. It is noticed that recurrent neural network (RNN) could be adopted to model the time-to-event distributions. RNN models often segment the time by the same interval and then could use the discrete time method to deal with the survival analysis problem. Long short-term memory (LSTM), which is a variant of the traditional RNN and has been widely used in sequence data modeling, also could be adopted to model the long-term dependency of time-varying covariates. Motivated by all these considerations, we present a novel method for survival prediction of breast cancer using bidirectional long short-term memory (biLSTM) [8], Cox model [14] network and auxiliary loss from mRNA and DNA methylation data. The efficacy of the proposed method was demonstrated through cross-validation tests on the benchmark dataset. The proposed method is not limited to breast cancer and can be applied for survival analysis of other cancer types having many samples.

2 Related Work

Early studies on cancer prognosis often focus on the use of single-model biomarkers. However, in these studies, some useful supplementary information between different

data modalities was ignored. With the advances of modern genomic technologies, integrative analysis on heterogeneous data to find important information for diagnosis, staging, and prognosis for cancers has received considerable attention [11]. Multi-features fusion analysis is being used extensively by pathologists in clinical practice. Some studies have explored a combination of different genomic biomarkers for survival analysis. [11] proposed an integrative pathway-based directed random walk (DRW) method on survival prediction of breast cancer utilizing the interaction between gene expression and DNA methylation. [31] integrated image data and genomic data to improve the survival prognosis of breast cancer patients. [3] constructed a novel framework that can predict the survival outcome of renal cell carcinoma patients by combining image features and gene expression features. These existing studies have suggested that different modalities of data complement each other and provide better patient stratification when used together. Although the combination of genomic features can better predict the clinical prognosis of cancer patients, simple combination of these features may bring redundant features, thus reducing the prediction performance, hence feature selection is the key step of multimodal feature fusion. In the existing research, the authors usually simply concatenate the multimodal data, and then apply the traditional feature selection methods to select the components related to cancer prognosis.

In clinical practice, pathologists make a diagnosis and predict prognosis by clinical exam. The clinical behavior of breast cancer is quite diverse, ranging from slow-growing localized tumors to aggressive metastatic disease [6]. Therefore, prognostic markers play a crucial role in stratification of patients for personalized cancer management, which could avoid either overtreatment or undertreatment [2]. For instance, patients classified into a high-risk group may benefit from closer follow-up, more aggressive therapies, and advanced care planning [30]. Cox proportional hazard model [14] is among the most popular survival prediction models. Recently, based on the Cox model, several regularization methods have been proposed in the literature. The least absolute shrinkage and selection operator Cox model (LASSO-COX) [19; 23] applies lasso feature selection method to select components that are related to cancer prognosis. Random survival forests (RSF) [9] computes a random forest using the log-rank test as the splitting criterion. It computes the cumulative hazards of the leaf nodes and averages them over the ensemble. Cox regression with neural networks by a one hidden layer multilayer perceptron (MLP) [28] was proposed to replace the linear predictor of the Cox model. It was showed that some novel networks were able to outperform classical Cox models [1]. DeepSurv [10] is a Cox proportional hazards deep neural network and a survival method for modeling interactions between a patient's covariates and treatment effectiveness in order to provide personalized treatment recommendations. DeepSurv is developed upon Cox proportional assumption with acutting-edge deep neural network. MTLSA [13] is the recently proposed model which regards survival analysis as a multi-task learning problem. It transforms the problem into a series of binary classification, and uses a multi-task learning method to model the event probability at different times. Although much progress has been made using above approaches, nevertheless, the prediction performance of the existing methods is still far from satisfactory, and there still exits much room for further improvement. In addition, these methods assumed that the survival information of one patient is independent from another, and

thus miss the strong ordinal relationship between the survival times of different patients.

3 Materials

In this study, the used survival analysis benchmark datasets including gene expression data, DNA methylation data, and clinical data. The clinical data are included in the main clinical file downloaded from The Cancer Genome Atlas (TCGA) [24], which provides an extensive collection of genomics and clinical outcome data for large cohorts of patients of more than 30 types of cancers. The main files contain 1097 breast cancer patients' clinical annotations and information. In our case, two clinical variables are used: Overall Survival Status (1 if the patient deceased, 0 if he/she is living at the time of the last follow-up) and Overall Survival (Months), which represent the number of months between diagnosis and date of death or last follow-up. In clinical data, patients with missing follow-up were excluded.

The gene expression data and DNA methylation data of breast cancer patients were obtained from the TCGA dataset of the Broad Institute GDAC Firehose [4]. Gene expression data from mRNA sequencing consisted of 20,533 genes. mRNA expression profiles were transformed from Illumina HiSeq 2000 RNA-seq readcounts to normalized reads per kilobase per million (RPKM). DNA methylation data were obtained as a gene-level feature of 20,106 genes by selecting the probe having a minimum correlation with expression data for each gene. We removed genes having gene expression values of 0. Gene expression data, DNA methylation data and clinical data were merged and filtered to keep matching records. We removed patients whose survival months were not recorded or wrongly so as negative values. So, among 1097 patients, we extracted 476 instances that had both mRNA sequencing and DNA methylation data. The benchmark dataset including gene data and survival data was obtained. The gene and clinical characteristics for the selected patients are summarized in Table 1.

Table 1. Gene and clinical characteristics.

Characteristics	**Summary**
Instance no.	476
Gene no.	
Methylation	20106
mRNA	20533
Survival status	
Living	413
Deceased	63
Follow up (months)	0.03-282.69
Age (years)	
Range	26-90
Median	57.23

4 Methods

4.1 Gene Feature Extraction

The large number of genes of mRNA and methylation posed a challenge to obtaining sufficient statistical power. Although, in deep learning, end-to-end method [25] has demonstrated excellent performance on various difficult problems, we still cannot use this method to our train survival model because the extracted high-level representation may be too coarse to accurately describe local features and the low-level features of the network are too precise and lack semantic information [15]. In addition, the number of integrated original genes is too large with more than forty thousand and the computational complexity is too high, which may cause errors in losses calculations. Denoising Autoencoder (DA) [26] has proven to be effective in selecting robust features against input noise and extracting more specific cancer-related pathways or genes [22]. But in our case, comparative experimental results demonstrate that DA method is not competitive. A recently developed weighted network mining algorithm called local maximum quasi-clique merging (lmQCM) [3], which could detect weak quasi-clique modules in weighted graphs with application in functional gene cluster discovery, has been successfully applied to gene co-expression analysis. lmQCM uses hierarchical clustering and does not allow overlap between modules. Another well-known gene clustering algorithm is weighted gene co-expression network analysis (WGCNA) [12], WGCNA is a powerful technique used to extract co-expressed gene networks from gene expressions, and widely used in genomic data analysis.

In our study, we tested the effectiveness of methods lmQCM and WGCNA respectively. By comparing the effects, we chose WGCNA as gene feature extraction method. Instead of focusing on individual genes, we firstly use the WGCNA algorithm to cluster genes into coexpressed modules, then summarized each module as an eigengene. Modules are clusters of highly interconnected or correlated genes. The eigengene of a module is defined as the first principal component which is considered to be the representative of gene expression profiles in a module. This method not only greatly improves statistical power, but also enables us to focus on important biological processes or gene variations related to coexpressed gene modules, making the results easier to explain than a single gene, because coexpressed modules are usually closely related to specific genomes that participate in the same biological process or are located on the same chromosome band. We use WGCNA algorithm to extract the features of mRNA and methylation respectively. Thus, WGCNA algorithm yields 12 coexpressed gene modules (features) for methylation data and 26 coexpressed gene modules for mRNA data. It is worth noting that to avoid overfitting, we applied gene feature selection methods to the training set and test set in cross-validation respectively.

4.2 Ordinal Cox Model

In survival analysis, prediction of the time duration until a certain event occurs is the goal and the death of a cancer patient is the event of interest in our study. Cancer patients can be divided into two categories i.e., censored patients and non-censored patients. For censored patients, the death events were not observed for them during the

follow-up period, and thus their genuine survival times are longer than the recorded data; while for non-censored patients their recorded survival times are the exact time from initial diagnosis to death. We use a triplet (x_i, t_i, δ_i) to represent each observation in survival analysis, where x_i is the feature vector, t_i is the observed time, and δ_i is the censoring indicator. Here, $\delta_i = 1$ or $\delta_i = 0$ indicates a non-censored or censored instance, respectively.

The primary goals in survival analysis are estimating the survival function and hazard function [27], both of which can be used to model the distribution of the event time over the timeline. Survival function $s(t \mid x)$ represents the probability that the event has not happened earlier than a specified time *t*. We define *O* as the variable of the true occurrence time for the event of interest. And $P_r(O)$ is the probabilistic density function (P.D.F.) of the true event time. So we have,

$$s(t \mid x) = P_r(O \geq t \mid x) \tag{1}$$

By defining the survival function $s(t \mid x)$ as the probability that a patient will survive after time *t*, the hazard function $h(t \mid x)$ that can assess the instantaneous rate of death is defined as follows:

$$h(t \mid x) = \lim_{\Delta t \to 0} \frac{P_r(t \leq O \leq t + \Delta t \mid O \geq t; x)}{\Delta t} \tag{2}$$

Where $x = (x_1, x_2, \cdots, x_n)$ corresponds to the covariate variable of dimensionality *n*. Among the hazards modeling methods, cox proportional hazard model [14], which is built based on the hypothesis that the hazard ratio between two instances is time-independent, is defined as:

$$h(t \mid x) = h_0(t) \exp(\theta^T x) \tag{3}$$

Here, $h_0(t)$ is the baseline hazard, and $\theta^T x_i$ is called survival function, in which $\theta = (\theta_1, \theta_2, \cdots \theta_n)$ can be estimated by minimizing its corresponding partial likelihood function. The partial likelihood is defined as follows:

$$L(\theta) = \prod_{i:\delta_i = 1} \frac{\exp(h(t_i \mid x_i))}{\sum_{j \in R(t_i)} \exp(h(t_j \mid x_j))} \tag{4}$$

Where T_i denotes the event time, δ_i is a binary value indicating whether the event happened or not, and $R(t_i)$ denotes the set of all individuals at risk at time t_i, which represents the set of patients that are still at risk before time t_i. Therefore, the coefficient vector can be learned via minimizing the negative partial log-likelihood function of the Cox model, which is defined as follows [21]:

$$L_Z(\theta) = -\sum_{i=1}^{n} \delta_i \left(\theta^T x_i - \log \sum_{j \in R(t_i)} \exp(\theta^T x_j) \right) \tag{5}$$

Although we could use the above Cox model to directly make survival prediction, it does not take the ordinal survival information between different patients (e.g., the survival time for patient A is longer than that for patient B) into consideration. In the hazard ratio based model, the ordinal relationship of the hazard risk between patient i and patient j can be easily derived by calculating the ratio (i.e., rec_{ij}):

$$rec_{ij} = \frac{h(t \mid x_i)}{h(t \mid x_j)} = \frac{h_0(t)\exp(\theta^T x_i)}{h_0(t)\exp(\theta^T x_j)} = \exp(\theta^T (x_i - x_j)) \quad (7)$$

In practice, if $rec_{ij} \geq 1$, the survival time for patient i should be shorter than that for patient j, and vice versa. By utilizing the above ordinal relationship indicated by Cox model, we design a ranking loss function to capture the ordinal survival information among different patients as follows:

$$ordLoss = -\sum_{i=1}^{n}\sum_{j \neq i} I * \max(0, 1 - rec_{ij}) = -\sum_{i=1}^{n}\sum_{j \neq i} I * \max(0, 1 - \exp(\theta^T (x_i - x_j)) \quad (8)$$

Where $I = 1$ if the survival time for patient i is shorter than that for patient j. Otherwise, $I = 0$.

By combining the Cox negative partial log-likelihood function L_Z with the above ordinal loss *ordLoss*, the objective loss function can be formulated as a multi-task model:

$$L(\theta) = L_Z(\theta) + \lambda * ordLoss \quad (9)$$

Where λ is a multi-task weight coefficient. Many previous methods which learn multiple tasks simultaneously use a naive weighted sum of losses, where the loss weights are uniform, or crudely and manually tuned. However, the model performance is extremely sensitive to weight selection. These weight hyper-parameters are expensive to tune. Therefore, it is desirable to find a more convenient approach which is able to learn the optimal weights. We developed a way of combining multiple loss functions to adaptively learn multiple objectives.

4.3 Adaptive Auxiliary Loss

In this study, we used methylation or gene expression features to make survival analysis for breast cancer patients. Our main task is obtaining the training model. The main task has a corresponding loss L_{main}, which can be the expected return loss used for calculating the policy gradient. We use the Cox negative partial log-likelihood function as the main loss L_{main}, i.e. $L_{main}=L_Z$. To improve data efficiency, besides the main task, one has access to one or more auxiliary tasks that share some unknown structure with the main task [18]. Here, the ordinal loss can be used as auxiliary loss of an auxiliary task, i.e. $L_{aux}=ordLoss$. Our goal is to optimize the main loss L_{main}. However, using gradient-based optimization with only the main task gradient $\nabla_\theta L_{main}$ is often slow and unstable, due to the high variance of the deep network. Thus, auxiliary tasks are commonly used to help to learn a good feature representation. We can combine the main loss with the loss from the auxiliary tasks as:

$$L(\theta) = L_{main}(\theta) + \lambda L_{aux}(\theta) \quad (10)$$

Where λ is the weight of the auxiliary task. Under the intuition that modifying θ to minimize L_{aux} will improve L_{main} if the two tasks are sufficiently related, we propose to modulate the weight λ at each learning iteration t by how useful the auxiliary task is for the main task given θ_t. θ_t is the set of all model parameters at training step t. We assume that we update the parameters θ_t using gradient descent on this combined objective:

$$\theta_{t+1} = \theta_t + \alpha \nabla_{\theta_t} L(\theta_t) \tag{11}$$

Where α is the gradient step size. At each optimization iteration, we want to efficiently approximate the solution to:

$$\underset{\lambda_t}{argmin}\, L\left(\theta_t + \alpha \nabla_{\theta_t} (L_{main} + \lambda_t L_{aux})\right) \tag{12}$$

The weights λ_t of the auxiliary task need to be tuned. However, tuning the parameters λ_t becomes more computationally intensive as the number of iteration increases. We try to look for a cheap heuristic to approximate λ_t which is better than keeping λ_t constant and does not require hyper-tuning.

Our goal is to find an auxiliary task with a weight to make L_{main} decrease the fastest. Specifically, $V_t(\lambda)$ is defined as the speed at which the loss of the main task decreases at time step *t*.

$$\begin{aligned} V_t(\lambda) &= \frac{dL_{main}(\theta_t)}{dt} \approx L_{main}(\theta_{t+1}) - L_{main}(\theta_t) \\ &= L_{main}(\theta_t + \alpha \nabla_{\theta_t} L(\theta_t)) - L_{main}(\theta_t) \\ &\approx L_{main}(\theta_t) + \alpha \nabla_{\theta_t} L_{main}(\theta_t)^T \nabla_{\theta_t} L(\theta_t) - L_{main}(\theta_t) \\ &= \alpha \nabla_{\theta_t} L_{main}(\theta_t)^T \nabla_{\theta_t} L(\theta_t) \end{aligned} \tag{13}$$

We can simply calculate the gradient to update λ :

$$\begin{cases} \nabla_{\lambda_t} V_t(\lambda_t) = \dfrac{\partial V_t(\lambda_t)}{\partial \lambda_t} = \alpha \nabla_{\theta_t} L_{main}(\theta_t)^T \nabla_{\theta_t} L_{aux}(\theta_t) \\ \lambda = \lambda + \beta \nabla_{\lambda} V_t(\lambda) \end{cases} \tag{14}$$

Where β is the gradient step size. This update rule is based on the dot product between the gradient of the main task and the gradient of the auxiliary task. The auxiliary loss will add extra gradient flow during backpropogation, thereby helping to reduce the vanishing gradient problem for earlier layers. Intuitively this approach leverages the online experiences to determine if an auxiliary task has been useful in decreasing the main task loss. The update rule is a product of a derivation of maximizing the speed at which the main task loss decreases.

So, we can obtain the gradient of empirical loss of the main loss L_{main} and the auxiliary loss L_{aux} respectively at search point *t* as:

$$\begin{cases} \nabla_{\theta_t} L_{main}(\theta_t) = -\sum_{i=1}^{n} \delta_i \left(x_i - \dfrac{\sum_{j \in R(t_i)} x_j \exp(\theta_t^T x_j)}{\sum_{j \in R(t_i)} \exp(\theta_t^T x_j)} \right) \\ \nabla_{\theta_t} L_{aux}(\theta_t) = -\sum_{i=1}^{n} \sum_{j \neq i} I * \max\left(0, 1 - (x_i - x_j) \exp(\theta_t^T (x_i - x_j))\right) \end{cases} \tag{15}$$

4.4 System Algorithm Flow Chart

LSTM is a kind of recurrent neural network architecture which is widely used in all kinds of sequential data. LSTM was designed to prevent vanishing gradients. Bidirectional LSTMs (biLSTM) are the bidirectional RNNs counterpart based on LSTM. Deep

bidirectional RNNs can be implemented by replacing each hidden sequence with the forward and backward sequences and ensuring that every hidden layer receives input from both the forward and backward layers at the level below. If LSTM is used for the hidden layers, we get deep bidirectional LSTM [5]. Because of biLSTM's ability of keeping previously observed data, it is very suitable for exploring the relationship within long sequence data, and that's why we chose biLSTM to model the survival behaviors. The auxiliary loss added to the original objective helps to improve the ability of optimizing the learning process in biLSTM training and regularization.

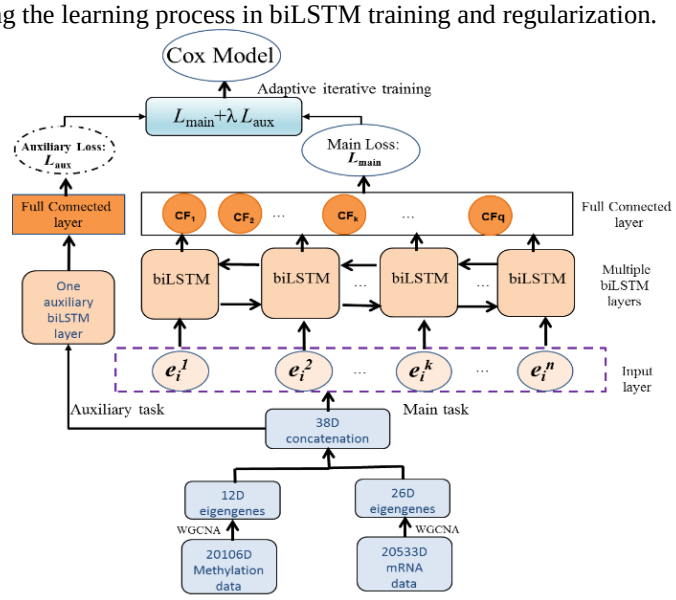


Fig. 1. Illustration of the proposed model and framework.

Fig.1 shows the algorithm process of our proposed method. There are several stages including the gene co-expression cluster stage, main/auxiliary biLSTM network stage and the COX model stage etc. In the gene co-expression cluster stage, mRNA data and methylation data could be reduced in feature dimension. WGCNA algorithms are used to cluster genes. So, mRNA and methylation eigengenes are obtained respectively. The directly concatenated eigengenes of mRNA and methylation will be the main task input features for the machine learning network to train the model. Meanwhile, we also use the concatenated eigengenes as auxiliary task input. In the main task, Multiple biLSTM layers, TimeDistributed layers, dropout layers, and full Connected layer are used to predict patient survival risk with proposed ordinal loss function, and then the main loss (i.e., L_{main}) is obtained. In the auxiliary task, we use one auxiliary biLSTM layer and a fully connected layer to obtain the auxiliary loss (i.e., L_{aux}). We can combine the main loss with the loss from the auxiliary tasks as: $L_{main} + \lambda L_{aux}$. We use a proposed adaptive optimization iteration method to tune the weight (λ) of auxiliary loss. Finally, through iterative training, the deep cox proportional hazard model is built for survival analysis to ensure that the ordinal relationship among the survival time of different patients can be preserved.

5 Experimental Results and Discussions

In this section, we assess the performance of the proposed method and carry out experiments on the training set through 10-fold cross validation. Specifically, we firstly use WGCNA algorithms to cluster genes and obtain methylation eigengenes and mRNA eigengenes. Meanwhile, we also use DA algorithm and lmQCM method to reduce the gene dimensions. We test these methods and compare the effectiveness of every one. By comparing the effects, we choose WGCNA method. Then the Cox proportional hazards model is built on the clustered eigengene features in the training set. After that, we compare the proposed method with only the main task loss method. In order to demonstrate the effectiveness of our proposed method, we also compare with five other machine learning methods. For survival stratification prediction, the median risk score predicted by the cox proportional hazards model is used as a threshold to split patients into low-risk and high-risk groups. Finally, we test if these two groups have distinct survival outcomes using Kaplan-Meier estimator and rank test. The survival curves are drawn by applying different methods.

We evaluate the performance of the proposed method and other comparing methods using the Concordance index (C-index). C-index quantifies the fraction of all pairs of patients whose predicted survival times are correctly ordered and is calculated as following:

$$C-index=\frac{1}{|k|}\sum_{i=1}^{m}\sum_{j:t_i<t_j} I(F(x_i)<F(x_j)) \tag{16}$$

Where k is the set of validly orderable pairs when $t_i<t_j$; $|k|$ represents the number of comparable pairs among them; $F(x)$ is the prediction of survival time; I is the indicator function of whether the condition in parentheses is satisfied or not. The C-index estimates the probability that, for a random pair of individuals, the predicted survival times of the two individuals have the same ordering as their true survival times. As the C-index only depends on the ordering of the predictions, it is very useful for evaluating proportional hazards models. This is because the ordering of proportional hazards models does not change over time, which enables us to use the relative risk function instead of a metric for predicted survival time.

5.1 Performance Comparison of Different Gene Feature Selection

In this section, we tested three gene feature selection methods: DA, lmQCM, and WGCNA. We set the number of DA encoder layer nodes as 100, and activation function as 'sigmoid'. In lmQCM, we set parameters with $\gamma=0.30$, $t=1$, $\alpha=1$, and $\beta=0.4$. To WGCNA, we set minModuleSize=*30*. Through these three different methods, methylation and mRNA features after dimensionality reduction can be obtained. We use DA algorithm to obtain 100 methylation features and 100 mRNA features respectively. We use lmQCM algorithm to obtain 33 methylation features and 24 mRNA features respectively. Similarly, we use WGCNA algorithm to obtain 12 methylation features and 26 mRNA features respectively. We combine methylation and mRNA features in series and obtain 200, 57, and 38-dimensional features in three dif-

ferent methods. We test each integrated feature / model and compare performance between different methods with C-index value. Table 2 summarizes the performance comparison of three methods. For the sake of fairness and convenience, we only carry out the same loss function, i.e. main loss (L_{main}), and the same biLSTM structure.

Table 2. Comparison of performance of three gene feature selection methods with C-index.

Gene Selection Methods	C-index
DA	0.5235
lmQCM	0.6102
WGCNA	0.6147

As shown in Table 2, it can be found that WGCNA and lmQCM methods have better performance than DA. In the cross validation on the standard data set, WGCNA is superior to lmQCM, but performance improvement is not obvious. Compared with the DA and lmQCM methods, the C-index of WGCNA is improved by 9.12% and 0.45%. Considering the similar computational complexity of WGCNA and lmQCM, we decided to adopt WGCNA method to extract gene features.

5.2 Performance Comparison with only Main Task Loss

In this section, we evaluate the performance concerning the proposed method with auxiliary loss. We use WGCNA algorithm to obtain 38-dimensional integrated features which are main task inputs, meanwhile, this integrated features will also be used as auxiliary task input features. We will compare the proposed method, which combines the main loss (L_{main}) with the auxiliary loss (L_{aux}) from the auxiliary task, with only the main loss (L_{main}). The experiments compare the performance of ten-fold cross validation between two methods on the standard data set and run 500 iterations with a constant step size of 0.01. Table 3 summarizes the performance comparison of the proposed method and only the main loss method with the measurements of the Concordance Index and convergence iteration. Convergence iteration is defined as number of training steps for reaching the goal. In this study, we set the goal (i.e. threshold of loss) as 1e-4.

Table 3. Performance comparison between two different loss methods with C-index and convergence iteration

Methods	C-index	Convergence iteration
The proposed method	0.6385	<300
Only main loss method	0.6147	>450

As demonstrated in Table 3, in the cross validation on standard data sets, the proposed method is superior to only the main loss method. Compared with the only the main loss method, the C-index of the proposed method is improved by 2.38%. From this comparison experiment, we found that the convergence speed of the proposed method is obviously faster than that of only the main loss method. The proposed method, by leveraging the combination of main loss and auxiliary loss, can dynamically adapt the weights for

the auxiliary task to perform better than or as well as the best main task. It means that using an adaptive auxiliary loss could give marginal improvement over the baseline. Additionally, we can see that the auxiliary task depends on the main task.

5.3 Comparison with Different Survival Prediction Methods over Cross-validation Test

We compare the prediction effects of our proposed biLSTM method with five machine learning methods: RSF [9], LASSO [23], MLP [1], DeepSurv [10], and MTLSA [13]. The C-index is used to evaluate the prediction performance. For the sake of fairness, we carry out the same feature set in all cross validation tests.

Table 4. Performance comparison among different survival prediction methods by the measurements of C-index (along with their standard deviations).

Methods	C-index
The proposed method	0.6385 (0.0115)
MTLSA	0.6048 (0.0332)
DeepSurv	0.6123 (0.0467)
MLP	0.6089 (0.0663)
LASSO	0.5644 (0.0097)
RSF	0.5529 (0.0178)

Table 4 summarizes the performance comparisons between the proposed method, MTLSA, DeepSurv, MLP, Lasso, and RSF by the measurements of C-index. From Table 4, we find that the cross validation of the proposed method on the standard training set is better than the other five methods. Compared with the methods: RSF, LASSO, MLP, DeepSurv, and MTLSA, the C-index of the proposed method is improved by 8.56%, 7.41%, 2.96%, 2.62% and 3.37%. As can be seen from Table 4, firstly, the prognosis power of the regularized Cox models (i.e., RSF and LASSO) is inferior to the other deep model based methods (i.e., MLP and DeepSurv). This is because the deep model can better represent gene features than the hand-crafted low-level features. Secondly, the proposed biLSTM method can achieve higher C-index values than the comparing methods, which demonstrates the advantage of LSTM that can represent the integrated patterns of sequential mRNA data and methylation data. The experiments also demonstrate the efficacy of the proposed method.

5.4 Survival Stratification Prediction

Another important task in survival analysis is to stratify cancer patients into subgroups with different predicted outcomes, by which we can develop personalized treatment plans during cancer disease progression. The median risk score method is used in the training set as a threshold to stratify patients in the test set into low-risk and high-risk groups, and then test if these two groups have significantly different survival time using the log-rank test. Better prognosis prediction performance comes with smaller p-value

from the log-rank test. We show the stratification performance of different prediction methods in Fig. 2.

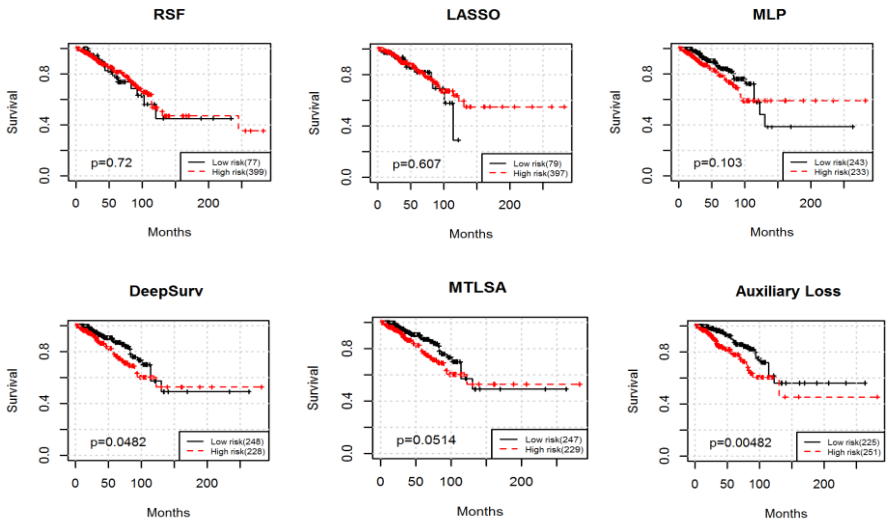


Fig. 2. The survival curves by applying different methods.

As shown in Fig. 2, the proposed prediction method (Auxiliary Loss) achieves significantly superior stratification performance (log-rank test P =0.00482) when compared with the other methods (log-rank test P =0.72, 0.607, 0.103, 0.0482 for RSF, LASSO, MLP, DeepSurv and MTLSA, respectively) on mRNA and methylation datasets, which shows the advantage of using auxiliary loss. In addition, it is worth noting that the proposed method could provide better prognostic prediction than the comparing methods, this is because our proposed model considers both the ordinal characteristics and the integrative patterns in survival analysis.

6 Conclusion

In this study, we have developed a survival analysis framework for breast cancer patients, in which we take patients' ordinal survival information into consideration. Ten-fold cross-validation experiments on the mRNA gene expression data, DNA methylation data and clinical data were carried out. Experimental results demonstrate the superiority of the proposed method over the existing RSF, Lasso, MLP, DeepSurv, and MTLSA methods. The good performances of the proposed method come from the use of the combined bidirectional LSTM predictor and ordinal information. Experimental results also show the importance of gene expression and DNA methylation signatures for breast cancer survival analysis. In this work, we have shown that dynamically combining an auxiliary task and adaptively adjusting the weights for the auxiliary task in an online manner can give a significant performance improvement for biLSTM Cox model network. The proposed method uses the idea that auxiliary tasks should provide

a gradient update direction that helps to decrease the loss of the main task. Our method is not limited to breast cancer and can be applied to other cancer types having many samples in TCGA.

Our future work will focus on integrative pathway-based survival prediction for breast cancer. We will also study how to exploit the best relationship between the auxiliary tasks and the main task.

References

1. Amiri, Z., Mohammad, K., Mahmoudi, M., Zeraati, H., and Fotouhi, A.: Assessment of gastric cancer survival: using an artificial hierarchical neural network. Pak J Biol Sci. 11(8),1076-84 (2008).
2. Chen, J.-M., Qu, A.-P., Wang, L.-W., Yuan, J.-P., Yang, F. et al.: New breast cancer prognostic factors identified by computer-aided image analysis of HE stained histopathology images. Scientific reports. 5,10690 (2015).
3. Cheng, J., Zhang, J., Han, Y., Wang, X., Ye, X. et al.: Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. Cancer research. 77(21),e91-e100 (2017).
4. Deng, M., Br ägelmann, J., Kryukov, I., and Saraiva-Agostinho, N.: FirebrowseR: an R client to the Broad Institute's Firehose Pipeline. Database. 2017 (2017).
5. Graves, A., Jaitly, N., and Mohamed, A.-r.: Hybrid speech recognition with deep bidirectional LSTM. In: 2013 IEEE workshop on automatic speech recognition and understanding, 2013, pp. 273-278. IEEE (2013).
6. Gulati, S., Martinez, P., Joshi, T., Birkbak, N.J., Santos, C.R. et al.: Systematic evaluation of the prognostic impact and intratumour heterogeneity of clear cell renal cell carcinoma biomarkers. European urology. 66(5),936-948 (2014).
7. Han, L., Yuan, Y., Zheng, S., Yang, Y., Li, J. et al.: The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. Nature communications. 5,3963 (2014).
8. Hochreiter, S., and Schmidhuber, J.: Long short-term memory. Neural computation. 9(8),1735-1780 (1997).
9. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., and Lauer, M.S.: Random survival forests. The annals of applied statistics. 2(3),841-860 (2008).
10. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T. et al.: Deep survival: A deep cox proportional hazards network. stat. 1050,2 (2016).
11. Kim, S.Y., Kim, T.R., Jeong, H.-H., and Sohn, K.-A.: Integrative pathway-based survival prediction utilizing the interaction between gene expression and DNA methylation in breast cancer. BMC medical genomics. 11(3),68 (2018).
12. Langfelder, P., and Horvath, S.: WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 9(1),559 (2008).
13. Li, Y., Wang, J., Ye, J., and Reddy, C.K.: A multi-task learning formulation for survival analysis. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1715-1724 (2016).
14. Lin, D.Y., Wei, L.-J., and Ying, Z.: Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika. 80(3),557-572 (1993).
15. Liu, G., Zhang, W., Qian, G., Wang, B., Mao, B. et al.: Bioimage-based Prediction of Protein Subcellular Location in Human Tissue with Ensemble Features and Deep Networks. IEEE/ACM transactions on computational biology and bioinformatics. (2019).

16. Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J. et al.: An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell. 173(2),400-416. e11 (2018).
17. Long, C.R., Westhusin, M.E., and Golding, M.C.: Reshaping the transcriptional frontier: epigenetics and somatic cell nuclear transfer. Molecular reproduction and development. 81(2),183-193 (2014).
18. Papoudakis, G., Chatzidimitriou, K.C., and Mitkas, P.A.: Deep reinforcement learning for Doom using unsupervised auxiliary tasks. arXiv preprint arXiv:1807.01960. (2018).
19. Shao, W., Cheng, J., Sun, L., Han, Z., Feng, Q. et al.: Ordinal multi-modal feature selection for survival analysis of early-stage renal cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018, pp. 648-656. Springer (2018).
20. Suzuki, H., Maruyama, R., Yamamoto, E., and Kai, M.: DNA methylation and microRNA dysregulation in cancer. Molecular oncology. 6(6),567-578 (2012).
21. Sy, J.P., and Taylor, J.M.: Estimation in a Cox proportional hazards cure model. Biometrics. 56(1),227-236 (2000).
22. Tan, J., Hammond, J.H., Hogan, D.A., and Greene, C.S.: Adage-based integration of publicly available pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions. MSystems. 1(1) (2016).
23. Tibshirani, R.: The lasso method for variable selection in the Cox model. Statistics in medicine. 16(4),385-395 (1997).
24. Tomczak, K., Czerwińska, P., and Wiznerowicz, M.: The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary oncology. 19(1A),A68 (2015).
25. Tsubaki, M., Tomii, K., and Sese, J.: Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics. 35(2),309-318 (2019).
26. Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning, 2008, pp. 1096-1103 (2008).
27. Wang, P., Li, Y., and Reddy, C.K.: Machine learning for survival analysis: A survey. ACM Computing Surveys (CSUR). 51(6),1-36 (2019).
28. Xiang, A., Lapuerta, P., Ryutov, A., Buckley, J., and Azen, S.: Comparison of the performance of neural network methods and Cox regression for censored survival data. Computational statistics & data analysis. 34(2),243-257 (2000).
29. Yang, X., Han, H., De Carvalho, D.D., Lay, F.D., Jones, P.A. et al.: Gene body methylation can alter gene expression and is a therapeutic target in cancer. Cancer cell. 26(4),577-590 (2014).
30. Yu, K.-H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C. et al.: Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nature communications. 7,12474 (2016).
31. Yuan, Y., Failmezger, H., Rueda, O.M., Ali, H.R., Gräf, S. et al.: Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. Science translational medicine. 4(157),157ra143-157ra143 (2012).

Author´s Index

Announcement

World Congress DSA 2021

The Frontiers in Intelligent Data and Signal Analysis July 12 - 23, 2021, New York, USA

www.worldcongressdsa.com

We are inviting you to our fourth World congress on the Frontiers of Signal and Image Analysis DSA 2021 to New York, Germany.

This congress will feature three events:

- the 17th International Conference on Machine Learning and Data Mining MLDM (www.mldm.de),
- the 21th Industrial Conference on Data Mining ICDM (www.data-mining-forum.de),
- and the 16th International Conference on Mass Data Analysis of Signals and Images in Artificial Intelligence&Pattern Recognition MDA-AI&PR (www.mda-signals.de).

Workshops and Tutorial will also be given.

Come to join us to the most exciting event on Intelligent Data and Signal Analysis.

Sincerely your,
Prof. Dr. Petra Perner

MLDM www.mldm.de

icdm www.data-mining-forum.de

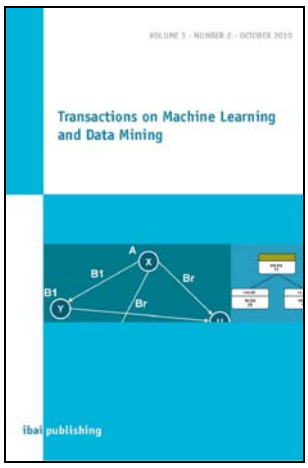
mda www.mda-signals.de

Journals by ibai-publishing

The journals are free on-line journals but having in parallel hardcopies of the journals. The free on-line access to the content of the paper should ensure fast and easy access to new research developments for researchers all over the world. The hardcopy of the journal can be purchased by individuals, companies, and libraries.

Transactions on Machine Learning and Data Mining

P-ISSN: 1865-6781 E-ISSN: 2509-9337



The International Journal "Transactions on Machine Learning and Data Mining" is a periodical appearing twice a year. The journal focuses on novel theoretical work for particular topics in Data Mining and applications on Data Mining.

Net Price (per issue): EURO 100
Germany (per issue): EURO 107 (incl. 7% VAT)

Submission for the journal should be send to:
info@ibai-publishing.org

For more information visited: www.ibai-publishing.org/journal/mldm/about.html

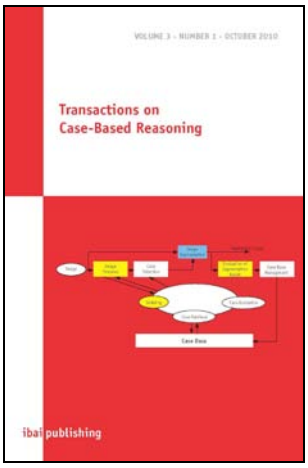
Transactions on Case-Based Reasoning

P-ISSN: 1867-366X E-ISSN: 2509-9345

The International Journal “Transactions on Case-Based Reasoning" is a periodical appearing once a year.

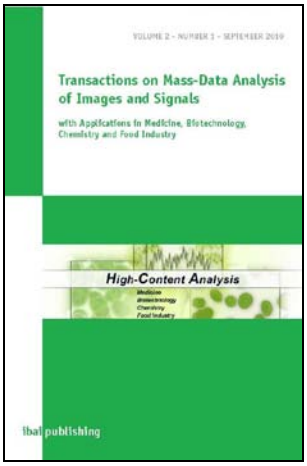
Net Price (per issue): EURO 100
Germany (per issue): EURO 107 (incl. 7% VAT)

Submission for the journal should be send to:
info@ibai-publishing.org



For more information visited: www.ibai-publishing.org/journal/cbr/about.html

Transactions on Mass-Data Analysis of Images and Signals ISSN: 1868-6451 E-ISSN: 2509-9353



The International Journal "Transactions on Mass-Data Analysis of Images and Signals" is a periodical appearing once a year.

The automatic analysis of images and signals in medicine, biotechnology, and chemistry is a challenging and demanding field. Signal-producing procedures by microscopes, spectrometers and other sensors have found their way into wide fields of medicine, biotechnology, economy and environmental analysis. With this arises the problem of the automatic mass analysis of signal information. Signal-interpreting systems which generate automatically the desired target statements from the signals are therefore of compelling necessity. The continuation of mass analyses on the basis of the classical procedures leads to investments of proportions that are not feasible. New procedures and system architectures are therefore required.

Net Price (per issue): EURO 100
Germany (per issue): EURO 107 (incl. 7% VAT)

Submission for the journal should be send to:
info@ibai-publishing.org

For more information visited: www.ibai-publishing.org/journal/massdata/about.php