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

P-ISSN 1864-9734
E-ISSN 2699-5220
ISBN 978-3-942952-94-1

www.ibai-publishing.org

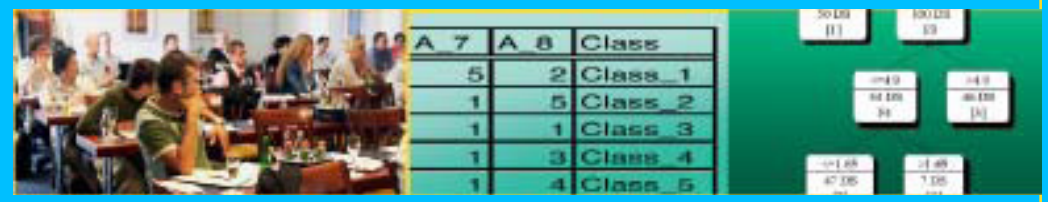
ISBN 978-3-942952-94-1

Advances in Data Mining, Proceedings, ICDM 2022

Petra Perner

Petra Perner (Ed.)

Advances in Data Mining



22nd Industrial Conference on Data Mining, ICDM 2022
New York, USA July 13-17, 2022

Industry Proceedings

ibai - publishing

Petra Perner (Ed.)

Advances in Data Mining

Industry Proceedings

22nd Industrial Conference, ICDM 2022
New York, USA, July 13 – 17 2022

ibai Publishing
www.ibai-publishing.org

Volume Editor

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

Hertha-Lindner-Str. 10-12
01067 Dresden, Germany

E-mail: pperner@ibai-institut.de

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 978-3-942952-94-1

Editorial

The twenty-second event of the Industrial Conference on Data Mining ICDM was held in New York again (www.data-mining-forum.de) running under the umbrella of the World Congress on "The Frontiers in Intelligent Data and Signal Analysis, DSA 2022" (www.worldcongressdsa.com).

At a time when we are still struggling with the corona pandemic, we scientists from different nations have gathered together for a peaceful discourse on an important research focus in the field of data mining and machine learning.

With our conference, we scientists show that we respect the opinions and work of others. That we are ready to consider them peacefully and in friendship under the critical view of the high scientific standards that this conference has.

The conference is an application-oriented conference. In recent years, we have seen a decline in interest in application-oriented research. Probably also because now innovative foundations are focused on oriented research, which requires not so time-consuming work with the experts on site. But we hope that interest will stabilize once we have overcome the consequences of Corona.

The International Program Committee has done an excellent and time-consuming job to select the best papers and provide important guidance on the work of the authors. I would like to thank all the members of the Program Committee for their efforts and that you have contributed with your top-class scientific competence.

The best papers are presented at this conference. The acceptance rate is 33%.

Thank you to all the scientists who have participated in this conference with your excellent work.

A special issue will be done after the conference in the Intern. Journal Transactions on Machine Learning and Data Mining (http://www.ibai-publishing.org/journal/mldm/about.php).

I would also like to thank those scientists who have participated in the conference with their work and have not been successful. Even if we have rejected work, we hope that the indications of the program committee will encourage you to reconsider your work and that you will perhaps face the critical scientific consideration of your work by the international program committee again next year.

The tutorial days rounded up the high quality of the conference. Researchers and practitioners got an excellent insight in the research and technology of the respective fields, the new trends and the open research problems that we like to study further.

A tutorial on Data Mining and a tutorial on Case-Based Reasoning, were held after the conference.

I also thank the members of the Institute of Computer Vision and applied Computer Sciences, Germany (www.ibai-institut.de), who handled the conference as secretariat. We appreciate the help and understanding of the editorial staff at ibai-publishing house, who supported the publication of these proceedings (http://www.ibai-publishing.org/html/proceeding.php).

Last, but not least, we wish to thank all the speakers and participants who contributed to the success of the conference. We hope to see you in 2023 in New York again

at the next World Congress on “The Frontiers in Intelligent Data and Signal Analysis, DSA 2023” (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 Medicine, Biotechnology, Chemistry, Biometry, Security, Agriculture, Drug Discovery and Food Industry, MDA (www.mda-signals.de), the workshops and tutorials.

July 2022 Petra Perner

22nd Industrial Conference on Data Mining ICDM 2022

www.data-mining-forum.de

Chair

Petra Perner
Institute of Computer Vision and Applied Computer Sciences, IBaI, Germany

Program Committee

Table of Content

Inter-and-Intra Domain Attention Relational Inference for Cabinet Temperature Prediction in Data Center
Fang Shen ……*1*

Multi-omics Survival Analysis via Adaptive Deep Sparse Canonical Correlation Analysis
Guanghui Liu ……*3*

Inter-and-Intra Domain Attention Relational Inference for Cabinet Temperature Prediction in Data Center

Fang Shen

Alibaba Group, China

Abstract. In a data center, predicting the cabinet temperature then generating alarms when an exception is detected can prevent server breakdown caused by high cabinet temperature. Each measuring point records the temperature of the cabinet over time, and each pair of measuring points may be associated with services or spaces. Therefore, the cabinet temperature prediction problem can be modeled as a graph-based prediction problem. In this case, the prediction of the cabinet temperature depends not only on its own historical temperature, but also on the temperature of cabinets related to services or in close space. Furthermore, the temperature of the cabinet is determined by various factors such as IT workloads and cold aisle temperature. Existing graph- based prediction methods do not consider the influence of these domains during the prediction, but only consider the temperature domain itself. To overcome this challenge, we propose an Inter-and-Intra domain Attention Relational Inference (I2A-RI) model: an unsupervised model that learns the relations between time series variables from different domains and utilizes the inferred interaction structure to achieve accurate dynamical predictions. Two attention modules, the guidance domain attention (GDA) module and the intra-domain attention (IDA) module, are proposed in I2A-RI, which encodes the inter-and-intra domain information to guide the learning procedure. Experiments on the real-world cabinet temperature dataset show that I2A-RI outperforms other state-of-the-art models since it takes the advantage of the ability to infer the potential interactions across domains. The benefits of the two proposed attention modules are also verified in the experiments.

Keywords: Data center, Cabinet temperature prediction, Relational inference, Graph neural network

Multi-omics Survival Analysis via Adaptive Deep Sparse Canonical Correlation Analysis

Guanghui Liu [0000-0002-1135-2939]

Department of Computer Science, State University of New York at Oswego, New York, USA

guanghui.liu@oswego.edu

Abstract. Survival analysis receives extensive application in cancer treatment and prediction. It is demonstrated to help understand the relationships between cases' variables and co- variates and the feasibility exerted by a range of treating choices. Existing studies have demonstrated that multi-omics fusion analysis can better stratify cancer patients with distinct prognosis than using single signature. However, those existing methods simply combine these characteristics in series, and ignore the correlation between different omics features. It is a challenging problem to investigate the underlying relationships among multi-omics features for survival analysis. In this work, we propose an adaptive multi-task learning framework via deep sparse canonical correlation analysis, which can find the best linear projections so that the highest correlation between the selected omics features can be achieved. First, we set each term in objective loss function of sparse canonical correlation analysis as a different loss task. Then, we adopt an adaptive deep network to optimize the linear transformations and weight variables of different tasks. This method can seek the maximal correlative features. In addition, we also can identify key genetic biomarkers of integrative features. Finally, we build a deep cox hazards model for survival analysis and use the selected features as input to predict patients' survival risk. We demonstrate the effectiveness of proposed method on three datasets from The Cancer Genome Atlas (TCGA). The results show the developed approach is achieving very competitive performance with comparing methods.

1 Methods

It is a challenging problem to investigate the underlying relationships among multi- omics features for diagnostic classification. In this project, we will propose an adaptive multi-task learning framework via deep sparse canonical correlation analysis, which can find the best linear projections so that the highest correlation between the selected omics features can be achieved. First, we set each term in objective loss function of sparse canonical correlation analysis as a different loss task. Then, we adopt an adaptive deep network to optimize the linear transformations and weight variables of different tasks. This method can seek the maximal correlative features.

Here, $\boldsymbol{\mathcal{L}0}$ could be treated as a main task and $\boldsymbol{\mathcal{L}i}$ (i = 1,2,3,4) as auxiliary task. So, the loss likelihood for this problem becomes as the following objective function:

$$\boldsymbol{L}(\boldsymbol{\theta}, \lambda) = \boldsymbol{\mathcal{L}0}(\boldsymbol{\theta}) + \sum \lambda \, \boldsymbol{\mathcal{L}i}(\boldsymbol{\theta}) + \sum log(1/\lambda) \qquad (1)$$

where θ is the set of all training model parameters, and λ (i = 1,2,3,4) are the regularization weights for the regularization term tasks respectively. λ are set with an initial 2value of 1 and are discouraged from decreasing too much by the negative exponential functions. The last term is an added penalty item. Small scale values λ will decrease the contribution of the task $\boldsymbol{\mathcal{L}i}$, whereas large scale λ will increase its contribution. The scale is regulated by the last term in the objective equation. The objective is penalized when setting λ too small.

We select the maximal correlative features from the original multi-omics features corresponding to the first d largest eigenvalues, and then merge these features together. Finally, using the selected integrative features as input, we build a cox model for survival analysis with deep networks. Figure 1 shows the framework of the proposed method.

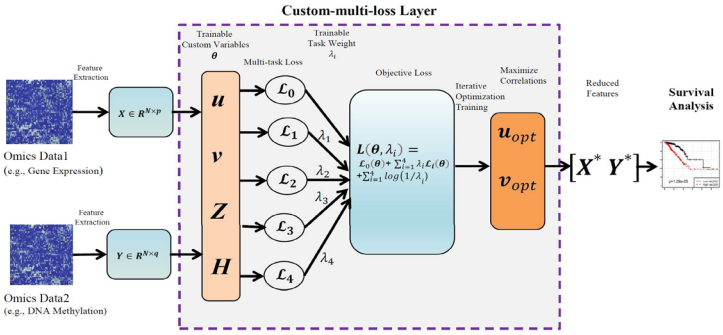


Fig. 1. Illustration of the proposed method and framework. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the input data from two omics. $\boldsymbol{X}^*$ and $\boldsymbol{Y}^*$ are the reconstruct data. $\boldsymbol{u}_{opt}$ and $\boldsymbol{v}_{opt}$ are the optimal canonical weights (eigenvalues).

2 Experiments

To further explore the effectiveness of the proposed method, we compare the developed MT-ADSCCA method with three existing machine learning survival prediction approaches: LASSO-COX, RSF, and MTLSA. For fairness, this part study runs the identical feature set in all cross-validation tests. Table 1 presents the performance comparison between the proposed method and the three methods by the measurements of the C-index on BRCA, GBMLGG and KIPAN datasets.

Methods	BRCA	GBMLGG	KIPAN
LASSO-COX	0.6807	0.7548	0.7342
RSF	0.6523	0.7426	0.7221
MTLSA	0.6894	0.7817	0.7413
Our method	0.7391	0.8449	0.7812

Table 1. Performance comparison among a range of survival prediction approaches by C-index on three datasets.

As shown in Table 1, it can be found that proposed method outperforms than other three

methods. Compared with the approaches: LASSO-COX, RSF, and MTLSA, the C-index of the MT-ADSCCA is improved on BRCA, GBMLGG, and KIPAN datasets.