Transactions on Case-Based
Reasoning for Mulitmedia Data
Vol. 1, No. 1 (2008) 37 - 46
©ISSN:1867-366X (Journal),
IBaI Publishing ISSN 1864-9734

ibai Publishing
www.ibai-publishing.org

Procurement Fraud Discovery using Similarity Measure Learning

Stefan Rüping, Natalja Punko, Björn Günter and Henrik Grosskreutz

Fraunhofer IAIS, Schloss Birlinghoven, 53754 St. Augustin, Germany
{firstName.lastName}@iais.fraunhofer.de,
WWW home page: http://www.iais.fraunhofer.de

Abstract. This paper describes an approach to detect risks of procurement fraud. It was developed within the context of a European Union project on fraud prevention. Procurement fraud is a special kind of fraud that occurs when employees cheat on their own employers by executing or triggering bogus payments. The approach presented here is based on the idea to learn a similarity measure that compares an employee (or payroll) standing-data record to a creditor record, in order to detect creditors that are suspiciously similar to employees. To this ends, it combines several simple similarity measures like address similarity or spatial similarity using a weighting scheme. The weights, that is the overall similarity function, are learned from user input specifying whether a particular pair of payroll and creditor data records are similar. This leads to an adaptive, easily transferable approach for a generic class of fraud opportunities.

Keywords: similarity measure learning, fraud detection, health care

1 Introduction

Health care fraud is responsible for losses amounting to very large amounts of money. According to some estimates, the losses amount to tens of billions of dollars each year [BH02]. A variety of approaches have been proposed to deal with fraud in general and in the context of health care. For example, [OFR06] describes an approach to detect medical claim fraud based on multilayer perceptrons. Other approaches are based on Bayesian networks [EN96,VDD04] or support vector machines [KPJ+03]. Overall, most of the approaches are based

on the idea to learn a function that classifies examples (or records) as fraudulent or non-fraudulent, using a training set of labeled examples as input.

In this paper, we do not address this general problem, but instead tackle a more specific kind of fraud, namely *procurement fraud*. Procurement fraud occurs when employees cheat their own employers by executing or triggering payments for services and products that are not delivered, work that is never done, or the like. Often, to do so fraudsters establish shadow-vendors or dummy companies, possibly incorporated by the fraudsters themselves, that they use as creditors.

The approach presented here was developed in the context of the iWebCare project[1]. Its goal is to detect hints on procurement fraud in the context of eGovernment in general and medical health care in particular. It was developed under the premise that there were no large sets of labeled training data available (here, we used the term labeled training data in the standard sense, i.e. to refer to example data classified as being fraudulent or fraud-free). The absence of such training data, which is quite frequent in fraud applications, prevents the use of one of the classification-based approaches discussed earlier. Instead, our approach is based on the idea to search for employees and creditors records that are suspiciously similar, where user feedback is used to identify what constitutes a relevant similarity. This approach is based on the assumption that many fraud cases turn out to involve payments linked with the fraudulent employee, one of his or her relatives, or to a person in his neighborhood.

Basically, our approach makes use of a set of what we call "basic similarity measures", which determine how similar an employee record and a creditor record are, for example, how close their addresses are; Thereafter, these basic similarity measures are combined by a weighting scheme. The weights are not fixed but are learned using user feedback: To this end, the system proposes a set of pairs of employees and creditors to the user and the user can provide, as feedback, whether a pair is indeed similar. Using this information, the system adapts it weights. The overall result of this approach is a function that measures the similarity of an employee and a creditor record. While this approach will not be able to detect all kinds of procurement fraud, it can be used to prevent this specific scheme of procurement fraud.

We remark that by its focus on learning a similarity measure, our approach is related to some approaches from the area of Case Based Reasoning (CBR) [AP94]. In CBR, similarity between problems plays a central role, and thus several authors have proposed approaches to build up a similarity metric (see also the discussion in the related work section).

The rest of this paper is structured as follows: In Section 2, we describe the application scenario for which the approach was developed. Thereafter, Section 3 presents the base similarity measures used. Section 4 describes how the base similarity measures are combined and how their weights are learned from user feedback. Section 5 presents some experimental results. Section 6 presents related work and finally, Section 7 summarizes and concludes.

[1] `http://iwebcare.iisa-innov.com/`

2 The Scenario

The approach presented in this paper was developed in the context of the project iWebCare, a project funded by the European Union whose aim is to develop a flexible fraud detection web service platform which shall help to combat fraud. In particular, iWebCare is intended to be used in the context of health care. One main objective of this project is that the resulting system should be operated by people who are no experts in the area of data mining. Thus, the resulting system should be easy to use and should hide most of the complexity of the data mining techniques from the user.

The situation was also characterized by the lack of large sets of labeled data, that is of records that are classified as been fraudulent or fraud-free. This prevented a classification-learning approach. As alternative, an approach was chosen where the system tries to find suspicious patterns in the data. Different approach where used to find suspicious patterns, for example subgroup discovery [GRW08,Wro97]. Another approach, which is described in detail in this paper, is to search for suspiciously similar payroll and creditor data sets.

2.1 Information in the Given Datasets

The data sets at hand consist of thousands of records with creditor standing data, and of records with employee standing data (also called payroll data). The payroll information includes the title, surname, forename, address, postcode, and telephone number of every employee. The creditor dataset contains informations like the title, name, address, postcode, telephone number and alternative payees.

2.2 Use Cases

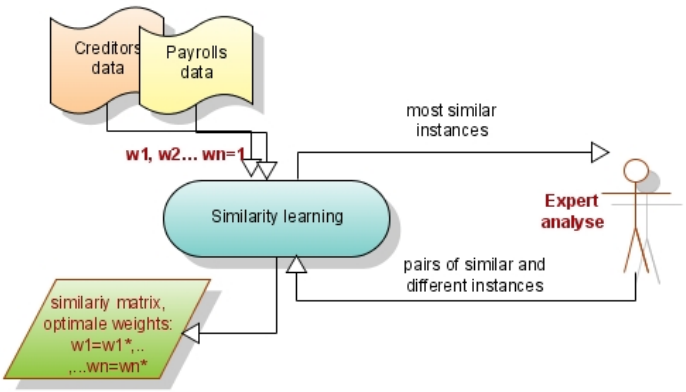


Fig. 1. Use-Case Optimize

The iWebCare system should be accessible via a Web front-end from actors involved in the health care business. The interaction of the users with the similarity approach described in this paper can be modeled by the following two use-cases:

- *Find pairs of payroll and creditor records that are similar*
  The user provides a data set with payroll data and another data set with creditor data to the system. The system search for the most similar pairs in the payroll and creditor datasets, using the similarity measure described in Section 4. These pairs are presented to the user. The user or fraud investigator can now decide whether these similarities appear suspicious to him and can start an investigation if it seems appropriate to him.
- *Learn/Optimize*
  In this use-case, the user trains the system in order to optimize the similarity function that assigns a similarity to every combination of a payroll and a creditor entry. To this end, the user selects a data set with standing data about employees (in particular addresses), and a second data set with standing data about creditors.
  The user submits this input to the system which computes the most similar pairs of payroll and creditor entries. The resulting list of pairs is shown to the user, which can now select pairs that are indeed similar, as well as pairs that are not similar. This annotation of the list of similar pairs can be sent to the system in order to optimize the parameters of the similarity function. Once the optimized parameters have been computed, an updated list of similar pairs is computed and shown to the user. The whole process can be repeated several times, as illustrated in Figure 1.

3 Base Similarity Measures

Given a payroll and a creditor record that are both structured, that is composed of several parts like address, postcode etc, there is no obvious way to express the "similarity" of the two. However, it is possible to compute different similarities for single components of the records; as a simple example, it is possible to compute the textual similarity of the name using the edit distance. Our approach is then to compute several of these similarities, which we call "base similarities" and thereafter to combine them to a more complex similarity function. We will now describe the different base similarities used in the iWebCare project:

3.1 Exact Matching Similarity

This is the simplest base similarity measure returning 1 for identical attributes and 0 otherwise. It operates on all attributes in the given dataset.

3.2 Text Similarities

There are several base similarity measures computed using text similarities accounting for the name, the address and the combination of the name and address.

The text is first transformed into word vectors using the *term frequency – inverse document frequency* (TF-IDF) which accounts for the overall appearance of the single words, hence emphasizing infrequent words. $tfidf$ is the product of two values, that is $tfidf = tf_{i,j} \cdot idf_i$. The first one is the term frequency of a term t_i in a document d_j given by

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{k,j}$ denotes the occurrences of term t_k in the document d_j. The second is the inverse document frequency denoted as

$$idf_i = log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

where $|D|$ is the total number of documents and the denominator is the number of documents d containing the term t_i.

Afterwards the resulting word vectors are compared using the cosine distance which is the normalized scalar product, hence representing the angle between the two vectors x and y. Using this distance, our text similarity measure for the vector representation x and y of a payroll and a creditor record is defined as:

$$textSim(x, y) = 1 - cosDist(x, y) = 1 - \frac{\sum_i (x_i \cdot y_i)}{\sqrt{\sum_i {x_i}^2} \cdot \sqrt{\sum_i y_i^2}}$$

3.3 Post Code Similarity

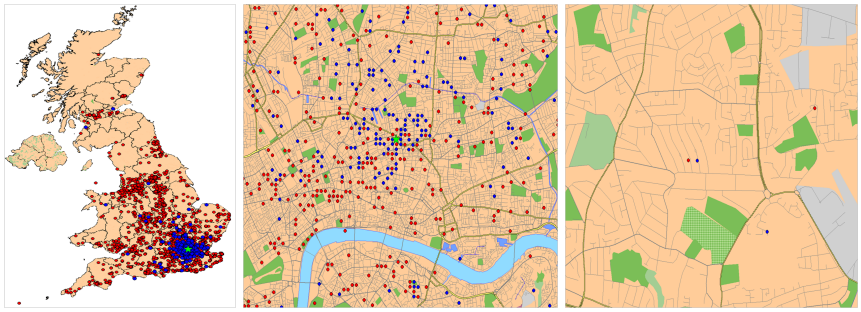
Another similarity measure is based on the post codes. The English postcode, which we examined in our experiments, is made up of two strings separated by a blank; the first block represents a coarse localization, which is refined by the following characters. The first block contains the area part (letters corresponds to the city) and the district part (numbers), the second block is split into the sector part (number) and the unit part (letters). As with most postcodes, the more letters of two postcodes match from left to right, the closer these two postcode areas are. The similarity is then calculated by traversing from left to right through the four parts of the post code, increases the value of the similarity by 0.25 for each consecutive match.

3.4 Spatial Similarity

The spatial similarity of a creditor and a payroll is computed based on the Euclidean distance $dist(x, y)$ of their geographic locations.

$$geoSim(x, y) = \frac{1}{1 + dist(x, y)}$$

The coordinates of these locations (Figure 2) are obtained by decoding the corresponding postcodes by means of available mapping software.



(a) (b) (c)

Fig. 2. Spatial Similarity. The blue and red dots represent the locations of the creditors and payrolls in Great Britain (2(a)) and London (2(b)). In 2(c) there is shown a highly suspicious and a more usual pair of a creditor and a payroll.

4 Similarity Measure Learning

Given the base similarities of several pairs of creditors and payrolls (c, p) and the corresponding fraud annotations which are given by the expert as described in the use-case (2.2), the distance of these pairs can be represented as vectors $\boldsymbol{dist(c,p)} = \mathbf{1} - \boldsymbol{sim(c,p)} = \mathbf{1} - (sim_1(c,p), ..., sim_n(c,p))^T$ in the space spanned by the base similarities (here, n refers to the number of base similarities). These vectors are then passed to the distance metric learning algorithm introduced by [XNJR02]. This algorithm optimizes the overall distance measure based on a diagonal matrix $\mathbf{A}$:

$$dist_A(c,p) = \sqrt{\boldsymbol{dist(c,p)}^T \cdot \mathbf{A} \cdot \boldsymbol{dist(c,p)}}$$

The goal is to minimize the distance between the similar pairs $\mathcal{S}$:

$$min_A \sum_{(c,p)\in\mathcal{S}} dist_A(c,p)^2 \tag{1}$$

To exclude the trivial solution $A = 0$, we define a constraint for the distinct pairs $\mathcal{D}$:

$$\sum_{(c,p)\in\mathcal{D}} dist_A(c,p) \geq 1 \tag{2}$$

Using Equations 1 and 2 we can solve the given convex problem by minimizing the following equation:

$$g(\mathbf{A}) = g(A_1, ..., A_n) = \sum_{(c,p)\in\mathcal{S}} dist_A(c,p)^2 - log \sum_{(c,p)\in\mathcal{D}} dist_A(c,p)$$

Unlike in the original algorithm in [XNJR02] we consider only a diagonal matrix $\mathbf{A}$ to achieve a linear combination of the base similarity measures.

The optimization of $g(\mathbf{A})$ is then accomplished using the Newton-Raphson Method. The final similarity measure is the linear combination of the base similarity measures described above using the weights represented by the computed diagonal elements $A_1, ..., A_n$ and can be used for later fraud detection.

4.1 Ontology-based Meta-data

The basic similarities described in Section 3 and the weights learned as described in this section are not restricted to data with the exactly same structure as the training dataset. Instead, based on the domain specific ontology of health care data developed in the context of iWebCare, it is possible to apply the similarities and weights to other data sets. Basically, all that is needed is some ontology-based meta-data that describes what kind of information is represented in the different columns of the data set.

This allows to easily transfer a model learned by our approach to similar, but not identical data sets: relevant attributes and their corresponding similarity measure can be identified using meta data. After this, the weight learned on one data set can be directly applied on the new data, such that the iterated user interaction already starts with a sensible first model, which is optimized further in the following iterations.

5 Preliminary Results

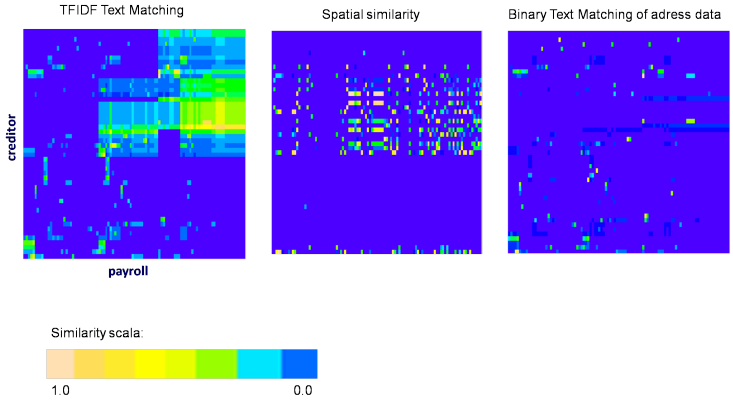


Fig. 3. Heat maps for three basic similarities: TFIDF similarity (left), spatial similarity (middle) and exact text matching (right).

In this section, we will briefly illustrate some results of our approach. The similarities between the different payroll and creditor records are visualized using

heat maps. In particular, Figure 3 shows the similarity of payroll and creditor data for the three basic similarity measures "TDFIF similarity" (left), "spatial similarity" (middle) and "exact text matching" (right) described in Section 3. The different payroll records are plotted on the x-axis, while the creditor records are plotted on the y-axis. The colors in the diagram shows how similar the employee in the x-axis is to the creditor in the y-axis: the brighter the cell, the more similar the two records.

Figure 4 illustrates the combined similarity. On the right, it shows the learned overall similarity assigned to every pair of payroll and creditor data record. For comparison, we have simply averaged the different base similarities and plotted the resulting averaged similarity on the left. It is easy to see that the averaged similarity is more blurry than the learned similarity, which shows a number of quite distinct peaks.

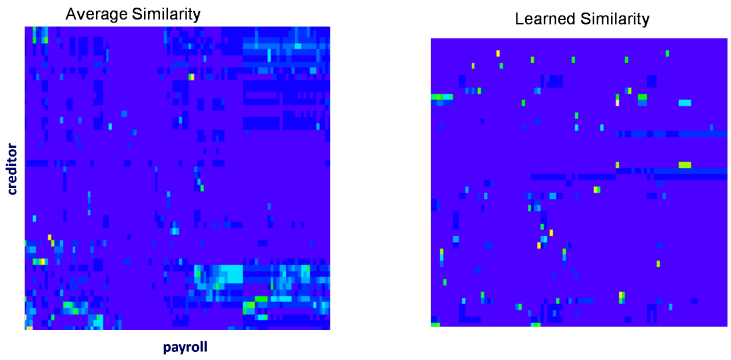


Fig. 4. Heat maps for the combined similarity. The image on the left results from simply averaging the base similarites, while the image on the right shows the result using learned weights.

6 Related Work

Similarity Learning has also been investigated in the area of Case Based Reasoning (CBR) [AP94], which attempts to solve problems via analogy with others problems for which a solution is known. CBR builds up on the assumption that similar problems have similar solutions and therefore similarity plays a central role here.

In particular, [XF06] presents an approach to build a similarity metric in two steps. First the compatibility on individual attributes is calculated and thereafter the individual compatibilities are fused into an overall similarity criteria, based on estimates of the utility between known cases.

In [PPM02], a system for the classification of images is presented which is based on a similar idea than the one presented here. Basic image features are computed and a weighting of these features is calculated with the goal to approximate the similarity between images that have been specified as similar by the user. Unlike here, the user is always presented 3 images and asked which two are the most similar to get an unbiased rating.

As for the other approaches to fraud discovery that are not based on similarity learning, [OFR06], [EN96,VDD04] and [KPJ+03] describe specific approaches to detect fraud using methods like support vector machines resp. neural networks. A review of methodologies applied to identify fraud in various domains like e-commerce or telecommunication is giben in [BH02].

7 Summary and Discussion

This paper presented an approach to find hints on procurement fraud by identifying identical persons in payroll and creditor data. Because there are different ways to mask this frequently occurring fraud case, we combine different base similarity measures which detect among others geographical proximity and names or addresses with similar spelling. The different base similarity measures are combined to yield a composed similarity measure. The weightings of these base similarities are determined based on users feedback. The similarity measure learning approach incorporates ontological information about the data attributes and can thus easily be transferred to other data sets.

The overall framework could be easily extended to incorporate other, possibly more elaborated basic similarity measures. For example, it would be possible to define similarities based on social network mining techniques, in order to detect less obvious relationships and similarities between creditors and employees. The work presented in this paper represents a starting point in that it describes a few simple base similarities and at the same time a framework for learning the weightings of different base similarities, which could easily be extended to incorporate further, possibly more elaborate base similarities.

Acknowledgments This work was partially supported by Project IST-2005-028055 with the title *iWebCare : Integrated Web Services Platform for the facilitation of Fraud Detection in health care e-government services.*

References

[AP94] Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Commun.*, 7(1):39–59, 1994.

[BH02] R. Bolton and D. Hand. Statistical fraud detection: A review. *Statistical Science*, 17(3):235–255, 2002.

[EN96] Kazuo J. Ezawa and Steven W. Norton. Constructing bayesian networks to predict uncollectible telecommunications accounts. *IEEE Expert: Intelligent Systems and Their Applications*, 11(5):45–51, 1996.

[GRW08] Henrik Grosskreutz, Stefan Rüping, and Stefan Wrobel. Tight optimistic estimates for fast subgroup discovery. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *ECML/PKDD (1)*, volume 5211 of *Lecture Notes in Computer Science*, pages 440–456. Springer, 2008.

[KPJ+03] Hyun-Chul Kim, Shaoning Pang, Hong-Mo Je, Daijin Kim, and Sung Yang Bang. Constructing support vector machine ensemble. *Pattern Recognition*, 36(12):2757–2767, 2003.

[OFR06] Pedro A. Ortega, Cristián J. Figueroa, and Gonzalo A. Ruz. A medical claim fraud/abuse detection system based on data mining: A case study in chile. In Sven F. Crone, Stefan Lessmann, and Robert Stahlbock, editors, *DMIN*, pages 224–231. CSREA Press, 2006.

[PPM02] Petra Perner, Horst Perner, and Bernd Müller. Similarity guided learning of the case description and improvement of the system performance in an image classification system. In Susan Craw and Alun D. Preece, editors, *ECCBR*, volume 2416 of *Lecture Notes in Computer Science*, pages 604–612. Springer, 2002.

[VDD04] Stijn Viaene, Richard A. Derrig, and Guido Dedene. A case study of applying boosting naive bayes to claim fraud diagnosis. *IEEE Trans. on Knowl. and Data Eng.*, 16(5):612–620, 2004.

[Wro97] Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In Jan Komorowski and Jan Zytkow, editors, *Proc. First European Symposion on Principles of Data Mining and Knowledge Discovery (PKDD-97)*, pages 78–87. Springer, 1997.

[XF06] Ning Xiong and Peter Funk. Building similarity metrics reflecting utility in case-based reasoning. *Journal of Intelligent and Fuzzy Systems*, 17(4):407–416, 2006.

[XNJR02] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart J. Russell. Distance metric learning with application to clustering with side-information. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 505–512. MIT Press, 2002.