Integrating knowledge engineering and data mining in e-commerce fraud prediction

Timo Polman, Marco Spruit[1]

[1] Institute of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
tppolman@gmail.com, m.r.spruit@uu.nl

Abstract. The number of merchants and consumers that participate in b2c e-commerce is still growing. Overall fraud rates have stabilized in recent years but for post-payment transactions in the Netherlands the fraud percentage remains unacceptably high. Companies often have a great deal of knowledge about fraudulent orders, and how to recognize them. Fraud prevention is often aided by automated recognition systems that are created through data mining. There have been few studies examining the combination of explicit domain knowledge and data mining. This study analyses the incorporation of domain knowledge in data mining for fraud prediction based on a historical dataset of 5,661 post-payment orders.

Keywords: Data mining, knowledge discovery in databases, knowledge engineering, automated fraud detection.

1 Introduction

Data mining and the broader *knowledge discovery in databases* or KDD have made a transition from an academic discipline to that of applied science, with usage in almost every field. Particularly interesting are the *predictive* data mining methods that are able to classify new data after being trained with historical data. Often, researchers try to gain knowledge exclusively from data, while in many domains there also is a great deal of relevant domain knowledge available. It has been pointed out quite early that incorporating domain knowledge in data mining might yield better results [1].

Many KDD and data mining publications [2-9] stress the importance of knowledge (engineering) in the process of knowledge discovery but not as an *explicit part of the classification itself*. This has been done in some more recent case studies including genes-disease associations [5], medical diagnosis [2], and indirect lending [9]. However, no such study has yet been performed in the e-commerce fraud prevention domain. While data mining is often used in e-commerce fraud detection [10-12], the incorporation of domain knowledge has yet to be examined for this domain. This research aims to fill this gap by examining the following research question:

"What improvements in e-commerce fraud prediction rates are possible when integrating expert domain knowledge and data mining techniques?".

2 Case Study

E-commerce merchants often suffer from fraud. Payment fraud – where the fraudster evades payment – is the best known and most practiced type of fraud. A payment method with a very high fraud rate is payment-on-credit. Customers then receive their package with an invoice inside it. The invoice can be posted or a regular (online) bank transfer can be used. The payment method is also popular because payment for the customer is postponed until approximately two weeks after delivery. The main problem associated with this payment method is that fraudsters can fake (a part of) their identity when purchasing. The goods are received but never paid for. The faked identity makes tracing afterwards very difficult and cost-ineffective. An undesired consequence is that customers who have honest intentions at the moment of purchase, default because they have run out of cash.

A medium sized online merchant in the Netherlands has provided us with a dataset of 5,661 post payment orders and the expert knowledge they possessed. Too many pay-on-credit orders in the past year have become uncollectible. For the payment method to remain feasible, an increase in fraud detection rates is necessary.

3 Methodology

3.1 Modeling the Knowledge System

The knowledge system will be classifying an order in terms of *suspicious* or *not suspicious* based on its set of attributes. A knowledge system is formed from domain knowledge, in our case possessed by two experts. As described by Schreiber et al. [13], this knowledge has to be *elicited* in order to be usable in a knowledge system. Furthermore, Shreiber et al. describe various methods for knowledge elicitation of which we will use structured and unstructured interviews.

Through analysis of the unstructured interviews performed we have identified the task template *classification*. A classification task takes object features as input, and gives an object class as output, *suspicious* or *not suspicious* in our case.

We have assembled structured interviews in order to gain insight in the experts' decision making process. These interviews have enabled us to form 13 rules, together forming a knowledge base. The rules apply to order characteristics, both already saved by the e-commerce software, for example *total order amount*, and new ones, to be deducted from the available data, for example *free email address*. We would present the experts with the rules already formed and ask them what rules were missing. This method was adapted from Schweickert et al. [14].

The classification will be performed by evaluating these rules.

Example of a classification rule as described above.

```
if (free_email_address and risk_products)
then score = score + 5
```

This rule will add a score of 5 to an order if the email address used originates from a list of known free email providers e.g. (hotmail, live, gmail), and one or more of the products ordered belongs to a list of products that are relatively more often bought by fraudsters. The sum of the scores generated by all the rules will decide which label will be assigned to the order.

3.2 Modeling the Data Mining Classifiers

The data mining will also be performing a classification task. Data mining classifiers are generated by *learning algorithms* that use training data. Different classification techniques employ different *learning algorithms* [15]. The input of an algorithm is the prepared input dataset, and its output is a model that can "predict the class labels of records it has never seen before" [15]. We will be evaluating eight different algorithms because the knowledge enhancement might perform differently over classifiers – as shown in a previous study [9].
We want to prevent *model over-fitting*; that is, if a generated model fits *noise*; "pays attention to parts of the data that are irrelevant" [16] or ". . . [fits] to data by chance" [17]. There are multiple techniques to avoid over-fitting, for example splitting the dataset into a test set and a training set. The classifier algorithm will train on the training set and can be validated on the test set. We will use cross-validation, a technique to split up the dataset multiple times in different training and test sets, to minimize the loss of using a smaller training set. More specifically we will use the *K-fold cross-validation* variant [15, 18].

3.3 Integrating KE in KDD and Data Mining

We will incorporate the *suspicious* classification from the knowledge system as a field in the training dataset, together with all the other order characteristics, and evaluate the performance of all eight classifiers – i.e. algorithms – with and without this variable.

4 Evaluation Methods

4.1 Measuring Classifier Performance

In order to answer our research question, we must know how to measure classifier performance. A classifiers performance can be ranked in *true positives* (correctly classified as positive), *true negatives* (correctly classified as negative), *false positives* (classified as positive while negative) and *false negatives* (classified as negative while positive). We will evaluate the performance of the different classifiers in the following measures deducted from these numbers; *area under the ROC curve* and *total cost.*

Area Under the Curve (AUC). A sophisticated measurement based on receiver operating characteristic (ROC) graphs. Recently, data mining studies have been using this method, originating from signal detection theory [19] in measuring classifier performance, for example [9, 20]. "The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance." [21]

Misclassification Costs. The costs of false positives and false negatives are good indicators. The use of misclassification costs has since long been applied within the medical domain [22], and increasingly more authors stress the importance of applying misclassification costs [23, 24] in data mining. The most relevant measure for our case company is *total costs* involved with a certain model choice.

4.2 Statistical Evaluation

Statistical evaluation of classifier performance over multiple datasets poses some problems in estimating the variance [25] and significance. According to a study [26], research in machine learning often assesses a significant difference in the wrong way. We will use the non-parametric *Wilcoxon Signed-Ranks Test* as proposed in [26] instead of the paired t-test. This test ranks the differences between pairs and then compares sum of the ranks for the positive and negative differences. With this test, we will try to reject our null hypothesis:

H_0: The incorporation of domain knowledge through an attribute added to the dataset yields no difference in AUC outcomes.

We will reject H_0 if the observed difference exceeds the 95% confidence interval.

5 Results

5.1 Overview

Table 1 shows the AUC and the standard deviation (SD) for the three best performing classifiers. Seven out of eight classifiers tested showed an increase when the knowledge-induced attribute was added.

Table 1. Area Under the Curve (AUC)

	No Domain Knowledge		Domain Knowledge	
Classifier	AUC	SD	AUC	SD
Naïve Bayes	0.717	0.033	0.727	0.033
Logistic Regression	0.728	0.033	0.738	0.032
AdaBoost	0.683	0.031	0.701	0.028
Average[1]	0.649		0.658	

[1] The averages are calculated based upon all eight classifiers as listed above.

5.2 Statistics

We have statistically evaluated the performance of all eight classifiers combined using the Wilcoxon Signed-Rank Test (1).

$$Z = -1.542, P = 0.123, \alpha = 0.05 \tag{1}$$

P is not equal to, or smaller than our chosen level of significance, which means we cannot reject our null hypothesis. We do not observe a statistically significant difference in data mining classifiers performance when incorporating knowledge engineering.

5.3 Case Study Results

For our case study company the most important performance indicator is cost. In a previous case study [27], false positive and false negative costs were calculated.
For each classifier we have calculated the total fraud related costs with and without incorporating the domain knowledge attribute. The estimated fraud cost decrease as a percentage of the turnover was 6.68% on average, for all eight classifiers. When applying the Wilcoxon Signed-Rank Test (2) we observe a significant increase.

$$Z = -2.100, P = .036, \alpha = .05 \tag{2}$$

5.4 Explanation of Results

Why did we fail to find statistical valid improvements when comparing the AUC of the different methods? First, our evaluation methods, both the classifier performance indicator we have chosen and the statistical test are very robust. This also implicates that a significant improvement is less likely to observe.

Second, as explained in our case description, an unknown part of the instances classified as fraud were *unintentional.* Domain knowledge about fraud is mostly unable to distinguish these from paying customers.

6 Conclusion and Discussion

Research Question. *What improvements in e-commerce fraud prediction rates are possible when integrating expert domain knowledge and data mining techniques?*

We did not observe a significant increase when comparing the AUC of all our classifiers. For the case-relevant measure, total costs, we have calculated that the cost

reduction by integrating domain knowledge for the eight classifiers we have chosen could be 6.68 % on average, a significant decrease.

Discussion. The addition of knowledge engineering in data mining as a research topic poses some difficulties. The integration of domain knowledge however is often only applicable in a very specific area, and its associated costs are relatively high, since the process of knowledge elicitation is highly time-consuming. Also, it can be difficult to determine whether a (lack of) performance increase originates from the nature of the data, or the quality of the knowledge system created.

Further Research directions. This issue deserves further empirical study; we are especially interested in the performance of our method when applied to other datasets and domains. Also, another promising line of research would be examining different types of knowledge systems and machine learning integration.

Acknowledgements. We would like to thank Total Internet Group for providing sample data, their domain knowledge and a great deal of support. Especially Joost Schildwacht and Joachim de Boer at TIG have been very helpful, both on the scientific and practical domains.

References

1. Pazzani, M., Kibler, D.: The Utility of Knowledge in Inductive Learning. Machine Learning. 9, 57-94 (1992).
2. Alonso, F., Caraça-Valente, J.P., González, A.L., Montes, C.: Combining expert knowledge and data mining in a medical diagnosis domain. Expert Systems with Applications. 23, 367-375 (2002).
3. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0 Step-by-step data mining guide, http://www.crisp-dm.org/CRISPWP-0800.pdf, (1999).
4. Daniëls, H.A.M., Feelders, A.J.: Integrating Economic Knowledge in Data Mining Algorithms. Tilburg University, Center for Economic Research (2001).
5. Dinu, V., Zhao, H., Miller, P.L.: Integrating domain knowledge with statistical and data mining methods for high-density genomic SNP disease association analysis. Journal of Biomedical Informatics. 40, 750-760 (2007).
6. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD process for extracting useful knowledge from volumes of data. Commun. ACM. 39, 27-34 (1996).
7. Kopanas, I., Avouris, N., Daskalaki, S.: The Role of Domain Knowledge in a Large Scale Data Mining Project. Methods and Applications of Artificial Intelligence. p. 746 (2002).
8. Langseth, H., Nielsen, T.D.: Fusion of Domain Knowledge with Data for Structural Learning in Object Oriented Domains. Journal of Machine Learning Research. 4, 339-368 (2003).
9. Sinha, A.P., Zhao, H.: Incorporating domain knowledge into data mining

classifiers: An application in indirect lending. Decision Support Systems. 46, 287-299 (2008).
10. Chan, P.K., Wei Fan, A.L., Stolfo, J.: Distributed Data Mining in Credit Card Fraud Detection. IEEE intelligent systems and their applications. 1094, 67-74 (1999).
11. Quah, J.T.S., Sriganesh, M.: Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications. 35, 1721-1732 (2008).
12. Sánchez, D., Vila, M.A., Cerda, L., Serrano, J.M.: Association rules applied to credit card fraud detection. Expert Systems with Applications. 36, 3630-3640 (2009).
13. Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N., Van de Velde, W., Wielinga, B.: Knowledge engineering and management. MIT Press, London (2000).
14. Schweickert, R., Burton, A.M., Taylor, N.K., Corlett, E.N., Shadbolt, N.R., Hedgecock, A.P.: Comparing knowledge elicitation techniques: a case study. Artif Intell Rev. 1, 245-253 (1987).
15. Pang-Ning, T., Steinbach, M., Kumar, V.: Chapter 5. Classification: Alternative Techniques. Data Mining. pp. 207-326. Addison Wesley (2005).
16. Moore, A.: Decision Trees Tutorial Slides, http://www.autonlab.org/tutorials/dtree.html, (2005).
17. Fayyad, U., Stolorz, P.: Data mining and KDD: Promise and challenges. Future Generation Computer Systems. 13, 99-115 (1997).
18. Kirkos, E., Spathis, C., Manolopoulos, Y.: Data Mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications. 32, 995-1003 (2007).
19. Provost, F., Fawcett, T.: Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. Presented at the Proceedings of the third international conference on knowledge discovery and data mining (1997).
20. Provost, F., Domingos, P.: Tree Induction for Probability-Based Ranking. Machine Learning. 52, 199-215 (2003).
21. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters. 27, 861-874 (2006).
22. Ambrosino, R., Buchanan, B.G.: The use of physician domain knowledge to improve the learning of rule-based models for decision-support. Proc AMIA Symp. 192-196 (1999).
23. Phua, C., Lee, V., Smith, K., Gayler, R.: A comprehensive survey of data mining-based fraud detection research. Artificial Intelligence Review. (2005).
24. Weiss, G., Provost, F.: The effect of class distribution on classifier learning: an empirical study. Rutgers Univ. (2001).
25. Nadeau, C., Bengio, Y.: Inference for the Generalization Error. Machine Learning. 52, 239-281 (2003).
26. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research. 7, 30 (2006).
27. Stolte, V.: Onderzoek naar een e-commerce fraudedetectie strategie, (2009).