Expert Systems with Applications 40 (2013) 5916–5923

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A cost-sensitive decision tree approach for fraud detection

Yusuf Sahin [a,*], Serol Bulkan [b], Ekrem Duman [c]

[a] Department of Electrical & Electronics Engineering, Marmara University, Kadikoy, 34722 Istanbul, Turkey
[b] Department of Industrial Engineering, Marmara University, Kadikoy, 34722 Istanbul, Turkey
[c] Department of Industrial Engineering, Ozyegin University, Cekmekoy, 34794 Istanbul, Turkey

ARTICLE INFO

Keywords:
Cost-sensitive modeling
Credit card fraud detection
Decision tree induction
Classification
Variable misclassification cost

ABSTRACT

With the developments in the information technology, fraud is spreading all over the world, resulting in huge financial losses. Though fraud prevention mechanisms such as CHIP&PIN are developed for credit card systems, these mechanisms do not prevent the most common fraud types such as fraudulent credit card usages over virtual POS (Point Of Sale) terminals or mail orders so called online credit card fraud. As a result, fraud detection becomes the essential tool and probably the best way to stop such fraud types. In this study, a new cost-sensitive decision tree approach which minimizes the sum of misclassification costs while selecting the splitting attribute at each non-terminal node is developed and the performance of this approach is compared with the well-known traditional classification models on a real world credit card data set. In this approach, misclassification costs are taken as varying. The results show that this cost-sensitive decision tree algorithm outperforms the existing well-known methods on the given problem set with respect to the well-known performance metrics such as accuracy and true positive rate, but also a newly defined cost-sensitive metric specific to credit card fraud detection domain. Accordingly, financial losses due to fraudulent transactions can be decreased more by the implementation of this approach in fraud detection systems.

1. Introduction

Fraud can be defined as wrongful or criminal deception aimed to result in financial or personal gain. The two main mechanisms to avoid frauds and losses due to fraudulent activities are fraud prevention and fraud detection systems. Fraud prevention is the proactive mechanism with the goal of disabling the occurrence of fraud. Fraud detection systems come into play when the fraudsters surpass the fraud prevention systems and start a fraudulent transaction. A review of fraud domains and detection techniques can be found in Bolton and Hand (2002), Kou, Lu, Sirwongwattana, and Huang (2004), Phua, Lee, Smith, and Gayler (2005), Sahin and Duman (2010). One of the most well-known fraud domains is the credit card systems. Credit card frauds can be made in many ways such as simple theft, application fraud, counterfeit cards, never received issue (NRI) and online fraud (where the card holder is not present). In online fraud, the transaction is made remotely and only the card's details are needed. Because of the international availability of the web and ease with which users can hide their location and identity over internet transactions, there is a rapid growth of committing fraudulent actions over this medium.

There are many previous studies done on credit card fraud detection. The general background of the credit card systems and non-technical knowledge about this fraud domain can be learned from Hanagandi, Dhar, and Buescher (1996) and Hand and Blunt (2001), respectively. The most commonly used fraud detection methods in this domain are rule-induction techniques, decision trees, Artificial Neural Networks (ANN), Support Vector Machines (SVM), logistic regression, and meta-heuristics such as genetic algorithms. These techniques can be used alone or in collaboration using ensemble or meta-learning techniques to build classifiers. Most of the credit card fraud detection systems are using supervised algorithms such as neural networks (Brause, Langsdorf, & Hepp, 1999; Dorronsoro, Ginel, Sanchez, & Cruz, 1997; Juszczak, Adams, Hand, Whitrow, & Weston, 2008; Quah & Sriganesh, 2008; Schindeler, 2006; Shen, Tong, & Deng, 2007; Stolfo, Fan, Lee, Prodromidis, & Chan, 1997; Stolfo, Fan, Lee, Prodromidis, & Chan, 1999; Syeda, Zhang, & Pan, 2002; Prodromidis, Chan, & Stolfo, 2000); decision tree techniques like ID3, C4.5 and C&RT (Chen, Chiu, Huang, & Chen, 2004; Chen, Luo, Liang, & Lee, 2005; Mena, 2003; Wheeler & Aitken, 2000); and SVM (Gartner Reports, 2010; Leonard, 1993).

Credit card fraud detection is an extremely difficult, but also popular problem to solve. There comes only a limited amount of data with the transaction being committed. Also, there can be past transactions made by fraudsters which also fit a pattern of normal (legitimate) behavior (Aleskerov, Freisleben, & Rao, 1997). Furthermore,

* Corresponding author. Tel.: +90 533 5566217; fax: +90 216 3480292.
E-mail addresses: ysahin@marmara.edu.tr (Y. Sahin), sbulkan@marmara.edu.tr (S. Bulkan), ekrem.duman@ozyegin.edu.tr (E. Duman).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.05.021

the problem has many constraints. First of all, the profiles of normal and fraudulent behaviors change constantly. Secondly, the development of new fraud detection methods is made more difficult by the fact that the exchange of ideas in fraud detection, especially in credit card fraud detection is severely limited due to security and privacy concerns. Thirdly, data sets are not made available and the results are often censored, making them difficult to assess. Even, some of the studies are done using synthetically generated data (Brause et al., 1999; Dorronsoro et al., 1997). Fourthly, credit card fraud data sets are highly skewed sets. Lastly, the data sets are also constantly evolving making the profiles of normal and fraudulent behaviors always changing (Bolton & Hand, 2002; Kou et al., 2004; Phua et al., 2005; Sahin & Duman, 2010). So, credit card fraud detection is still a popular challenging and hard research topic. Visa reports about credit card frauds in European countries state that about 50% of the whole credit card fraud losses in 2008 are due to online frauds (Ghosh & Reilly, 1994). Many papers reported huge amounts of losses in different countries (Bolton & Hand, 2002; Dahl, 2006; Schindeler, 2006). Thus new approaches improving the classifier performance in this domain have both financial implications and research contributions. Defining a new cost-sensitive approach is one of the best ways for such an improvement due to the characteristics of the domain.

Although traditional machine learning techniques are generally successful in many classification problems, having a high accuracy or minimizing the misclassification errors is not always the goal for the classifier developed. In the applications of real-world machine-learning problem domains, there are various types of costs involved and Turney defined nine main types of costs (Turney, 2000). However, most of the machine-learning literature does not take any of these costs into account, only a few of the remainings take the misclassification cost into consideration. Turney also stated that the cost of misclassification errors occupies a unique position in their taxonomy (Turney, 2000). Nevertheless, according to the Technological Roadmap of the ML-netII project (European Network of Excellence in Machine Learning), cost-sensitive learning is stated to be one of the most popular topics in the future of machine learning research (Saitta, 2000; Zhou & Liu, 2006). Thus, improving classifier performance of a fraud detection system by building cost-sensitive classifiers is the best way to enable recovery of large amounts of financial losses. Besides, the customer loyalty and trust will also be increased. Also, cost-sensitive classifiers have been shown to be effective in addressing the class imbalance problem (Thai-Nghe, Gantner, & Schmidt-Thieme, 2010; Zhou & Liu, 2006).

Most of the past studies work on constant misclassification cost matrices or cost matrices composed of a number of constant heterogeneous misclassification costs; however, each false negative (FN) has a unique misclassification cost inherent to it. Accordingly, each FN should be prioritized in some way to show the misclassification cost difference. For example, a fraudulent transaction with a larger transaction amount or a larger usable card limit should be more important to detect than one with a smaller amount or usable card limit. A constant cost matrix or a combination of constant cost matrices cannot depict this picture. So, this study is one of the pioneers to take such cases into account while working with classification problem under variable misclassification costs. This is one of the gaps in the literature of credit card fraud detection aimed to be filled by this study.

In this study, a new cost-sensitive decision tree induction algorithm that minimizes the sum of misclassification costs while selecting the splitting attribute at each non-terminal node of the tree is developed and the classification performance is compared with those of the traditional classification methods, both cost-insensitive and cost-sensitive with fixed misclassification cost ratios, such as traditional decision tree algorithms, ANN and SVM. The results show that this cost-sensitive decision tree algorithm outperforms the existing well-known methods on our real-world data set in terms of the fraudulent transactions identified and the amount of possible losses prevented.

In credit card fraud detection, the misclassification costs and the priorities of the frauds to be identified differ depending on the individual records. As a result, the common performance metrics such as accuracy, True Positive Rate (TPR) or even Area Under Curve are not suitable to evaluate the performance of the models because they accept each fraud as having the same priority regardless of the amount of that fraudulent transaction or the available usable limit of the card used in the transaction at that time. A new performance metric which prioritizes each fraudulent transaction in a meaningful way and checks the performance of the model in minimizing the total financial loss should be used. Fraudsters generally deplete the usable limit of a credit card once they get the opportunity of committing fraudulent transactions using the card. Accordingly, the financial loss of a fraudulent transaction can be assumed as the available limit of the card before the transaction instead of the amount of the transaction. So, the performance comparisons of the models over the test set are done over the newly defined cost-sensitive performance metric Saved Loss Rate (SLR) which is the saved percentage of the potential financial loss that is the sum of the available usable limits of the cards from which fraudulent transactions are committed. To show the correctness of our argument, True Positive Rate (TPR) values for the performance of the models are also given in performance comparisons of the models.

The rest of this paper is organized as follows: Section 2 gives a review of the cost-sensitive approaches in machine learning and Section 3 gives some insights to the structure of credit card data. Section 4 gives the details of the newly developed cost-sensitive decision tree algorithm. Section 5 gives the results and a short discussion about the results and Section 6 concludes the study.

2. Cost sensitive approaches in machine learning

There are different methods used to take cost sensitivity into account while building up classifier models. The first one builds up cost-sensitive classifier models by changing the training data distributions by oversampling or undersampling so that the costs of the data in the set are conveyed by the appearance of the examples. Some studies tried to overcome misclassification cost problem by stratification; and duplicating or discarding examples when the data set is imbalanced (Japkowicz, 2000; Kubat & Matwin, 1997). However, these researchers assume that the cost matrix entries are fixed numbers instead of record-dependent values. Researchers such as Domingos tried to build up mechanisms like MetaCost to convert cost-insensitive classifiers to cost-sensitive ones (Domingos, 1999; Elkan, 2001.

According to some studies, oversampling is effective in learning with imbalanced data sets (Japkowicz & Stephen, 2002; Japkowicz et al., 2000; Maloof, 2003). However, oversampling increases the training time and because it makes copies of examples of the minor class/classes, it may result in overfitting problem (Chawla, Bowyer, & Kegelmeyer, 2002; Drummond & Holte, 2003). Unlike oversampling, undersampling tries to decrease the number of examples of major class/classes so that a balance is achieved in the distribution of training set data with respect to classes. Some studies have shown that undersampling is good at handling the imbalanced data problem (Drummond & Holte, 2003; Japkowicz & Stephen, 2002; Japkowicz et al., 2000; Maloof, 2003).

The second method to take cost sensitivity into account while building up classifier models is adjusting the threshold toward inexpensive classes to make misclassification of expensive class