Data Mining Techniques for Credit Card Fraud Detection: Empirical Study

Marwan Fahmi, Abeer Hamdy, Khaled Nagati
British University in Egypt, Faculty of Informatics and Computer Science

Abstract

Fraud detection is a crucial problem that has been facing the e-commerce industry for decades. Financial institutions throughout the world lose billions due to credit card fraud, which necessitate the use of credit card fraud prevention. Several models have been proposed in the literature, however, the accuracy of the model is crucial. In this paper four fraud detection models based on data mining techniques (Support vector machine, K-nearest neighbours, Decision Trees, Naïve Bayes) were developed and their performances were compared when applied on a real life anonymised data set of transactions (“UCSD-FICO Data Mining Contest 2009”). Four relevant metrics were used in evaluating the performance of the classifiers which are True positive rate (TPR), False Positive Rate (FPR), Balanced Classification Rate (BCR) and Matthews Correlation Coefficient (MCC).

© 2012 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Global Science and Technology Forum Pte Ltd

Keywords: Fraud detection, SVM, K-NN, Naïve Bayes, Decision Tree.

1. Introduction

In today’s world, internet-dependency has reached its peak. As a result, the use of internet based purchases using credit cards has become convenient and essential. Eventually, credit cards’ transactions have become the de facto standard for E-commerce, which is the electronically performed commercial transactions for buying/selling of products or services.

In Egypt, e-commerce has attracted millions of pounds of investments from foreign and Egyptians investors due to the high potential of growth in the Egyptian internet market with more than 40 million internet users and 15 million online buyers making it the largest market in the middle east, according to Al-Khalidi, Abdalla, Soudodi, & Syed (2015). In the same manner, the Egyptian internet economy represented 1.1% of the country’s 2011 Gross domestic product (GDP) (Chabenne, Dean, De Bellefonds, Stevens & Zwillenberg, 2012); in the study conducted by The Boston Consulting Group (BCG) and Google Egypt titled “Egypt at a Crossroad: How the internet is Transforming Egypt’s Economy”, it showed that the Egypt’s
internet economy in 2011 reached EGP 15.6 billion. Further, the internet played an important role in the 2011 revolution which created an awareness of its power and capability of changing the way of life. Speaking of the potential growth of Egypt’s e-commerce, the study added “We anticipate that Egypt's e-commerce, which now represents about 0.2 to 0.3 percent of retail spending, could grow to 0.9 (The equivalent of E£14.5 billion in nominal Egyptian pounds, or E£8.5 billion in 2011 Egyptian pounds)” (Chabenne, Dean, De Bellefonds, Stevens & Zwillingberg, 2012). Additionally, Al-Khalidi, Abdalla, Soudodi, & Syed (2015) stated that Egypt’s internet market was anticipated to reach USD 2.7 billion by 2020, almost double the value in 2014. In the light of that, the Arab world’s overall market was anticipated to reach USD 13.4 by 2020 compared with 7 billion in 2014.

Globally, the United States of America’s e-commerce market sales reached USD 231 billion in 2012 and were expected to reach USD 370 billion by 2017 passing the total amount of sales generated by businesses that use buildings and stores for operations. In addition, according to the U.S. Census Bureau, “53% of people in the USA shopped online in 2011 and it is predicted to grow to 58% in 2016” (as cited in Li, 2013).

Meanwhile, e-commerce market sales in western Europe are anticipated to increase at a quicker rate compared with the USA during the period between 2013 to 2017, from EUR 112 billion to EUR 191, respectively (Forrester Research as cited in Indvik, 2013). In the UK, the market represented 13% of the economy in 2013 and was anticipated to increase its share to 15% by 2017. While in France, the market represented only 5% of the total country’s economy in 2013 and was expected to reach 7% in 2017. Furthermore, 65% of internet users in the European Union (EU) used the internet for shopping in 2015 given that 80% of the EU citizens used the internet in the last few months of 2015 (Reinecke, 2015).

With nearly half of earth’s population online by this year, as it was predicted, digital market was estimated to reach USD 4.2 trillion in the G-20* economies. In their elaboration, Dean, et al. (2012) added “If it were a national economy, the Internet economy would rank in the world’s top five, behind only the U.S., China, Japan, and India, and ahead of Germany.”

The size of credit cards’ transactions is growing rapidly creating a higher risk of fraud occurrences to financial institutions and individuals. Despite the industry’s huge effort to fight such a serious challenge, losses are still increasing as fraudsters find ways to get the necessary information to access the merchants’ systems and obtain the money. According to The Nilson Report (2015), the total amount of fraud losses of banks and businesses around the world reached more than USD 16 billion in 2014 with an increase of nearly USD 2.5 billion in the previous year recorded losses, meaning that, each USD 100 is having 5.6 cents that was fraudulent, the report concluded.

Besides, based on samples of their transactions’ data, Sift Science † was able to identify top 25 most fraudulent countries; Egypt was listed the second most country that generates fraudulent e-commerce transactions across the world, following Latvia which is the first in the list and followed by the United States of America, according to Gillie (2013).

Although, online payment is becoming the most commonly used method of payment due to its ease of use and efficiency, it still has its challenges. The rise of e-commerce across the globe and the users’ need for the electronic method of payment using credit and debit cards have created opportunities for fraudsters. Credit card fraud is one of the major challenges for online payment, given the absence of this challenge in alternative methods of payment; such as cash on delivery. Over the coming few years, fraudsters are expected to increase

---

* G-20: An international forum for governments of the major twenty economies in the globe.
† Sift Science: A fraud detection solution for websites and mobile applications, founded June 1, 2011 in San Francisco, California
- [http://www.g20.org/English/aboutg20/AboutG20/201511/20151127_1609.html](http://www.g20.org/English/aboutg20/AboutG20/201511/20151127_1609.html)
- [https://siftscience.com/](https://siftscience.com/)
in number, as a result, the report is expecting that by 2020, fraud losses will reach over USD 35 billion, globally.

Clearly, the first task that needs to take place in order for a card fraudulent activity to occur is a thief getting credit cards information. Card Not Present (CNP) transactions are perfect for fraudsters since no PIN code, signature or the physical card’s existence are required. A CNP transaction is a transaction made through the internet, phone or mail by exchanging the credit card information. There are several ways for thieves to get this information which includes Phishing, Data breach, Malware, Skimming. Phishing is an old strategy that requires contacting the victims directly through mail, phone or any other way, asking for their card information usually as the card issuer’s representative. Data Breach is another way thieves can obtain credit cards information through online data breaches. When sensitive and protected data including credit cards information are released to untrusted environment whether intentional or unintentional. Malware requires a malware software to be installed on the victim’s personal device to get the customer credit cards information when using them on that device. Skimming requires a scanning device usually called a skimmer that can obtain and record credit card information once the card runs through it while someone is performing a legitimate transaction. These devices can be planted in places where credit cards are swiped regularly. Once the criminal has the credit cards information, they use them for stealing through internet (fraud transaction). The paper is focused on detecting such fraud transactions.

The contribution of the paper could be summarized in the following:
1. Four classifiers based on different machine learning techniques (Naïve Bayes, Decision trees, support vector machines and K-nearest neighbours) were trained on a real life data set of financial transactions for the purpose of fraud detection and their performances were evaluated and compared based on several relevant metrics for the domain.
2. The dataset was investigated and some of the features were removed which resulted in reducing the dimension of the input space.
3. Under sampling technique was applied on the dataset to deal with its imbalanced nature, in order to improve the performance of the classifiers.

This paper is organized as follows: Section 2 surveys previous fraud detection model, section 3 proposed the methodology in developing data mining based fraud detection models. Section 4 discuss the results and section 5 concludes the paper.

2. Related Work

Several machine learning techniques have been used in the literature to approach the credit card fraud detection problem. Ng & Singh, 1997 developed models based on an individual and combined machine learning techniques for handwritten digits recognition. The results showed that the classification accuracy of the combined classifier model outperformed the individual classifier model. Maes 2002 tried Artificial Neural Networks (ANN) and Bayesian Belief Networks (BBN) on a real dataset obtained from Europay International. Their experiment showed that the Bayesian Belief networks outperforms ANN in terms of classification accuracy and training time. It was found that ANN may need several hours for training while BNN takes only 20 minutes. However, the trained ANN was found to be faster in classifying new instances. Foster & Stine 2004 attempted to predict personal bankruptcy using a fully automated stepwise regression model and compared its performance to decision trees. They found out that the performance of the statistical model is competitive with decision trees. Zaslavsky & Strizkak 2006 compared ANN based model to a decision trees model, and they found out that the ANN outperforms decision trees. Sahin & Duman, 2011 applied decision
trees and support vector machines (SVM) on a dataset obtained from a real world national bank’s credit card data warehouses. They found out that decision trees outperform SVM in solving the problem. Haung, 2013 developed two models based on logistic regression and SVM. It was found that logistic regression outperforms SVM. Ehramikar, 2000 developed a fraud detection model based on the decision trees and he founded that decision trees suffer from under fitting problem in case of imbalanced data set (case of fraud detection dataset).

3. Methodology

Fraud detection is a binary classification task in which any transaction will be predicted and labeled as a fraud or legit. In this paper state of the art classification techniques were tried for this task and their performances were compared. The following subsections briefly explain these classification techniques, data set and metrics used for performance measure.

3.1. Classification techniques

Naïve Bayes

Naïve Bayes is based on two assumptions. Firstly, all features in an entry that needs to be classified are contributing evenly in the decision (equally important). Secondly, all attributes are statistically independent, meaning that, knowing an attribute’s value does not indicate anything about other attributes’ values which is not always true in practice.

The process of classifying an instance is done by applying the Bayes rule for each class given the instance. In the fraud detection task, the following formula is calculated for each of the two classes (fraudulent and legitimate) and the class associated with the higher probability is the predicted class for the instance.

Decision Tree

There are two categories of decision trees, classification trees and regression trees. The decision tree learning is the construction of a decision tree from class-labeled training tuples. A decision tree consists of nodes that forms a tree structure; the topmost node is called the root node. Each non-leaf node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf node holds a class label. Leaf nodes represent classes that are returned if reached as the final prediction by the model. As Zaki&Meira (2014) elaborated, given an instance with its features’ values, the model is able to classify the instance by traversing the decision tree. There are several decision tree algorithms including: ID3 (Iterative Dichotomiser 3), C4.5 (successor of ID3) and CART (Classification and Regression Tree)

k-Nearest Neighbors

The k-Nearest Neighbors (KNN) algorithm is a simple instance-based algorithm that plots all training instances and classify unlabelled instances based on their closest neighbours. In instance-based learners instances themselves are used to represent the model unlike the decision tree algorithms that use instances to develop a tree and that tree represents the model. However, it is argued that all learning algorithms are instance-based since they all use instances of the training set to construct models. In the KNN technique, an
unlabelled instance is classified by calculating the distances between the instance and surrounding instances based on a determined distance metric and the majority class is assigned to the unlabelled instance.

Support vector machines

Support vector machines (SVM) were first introduced by Vapnik, 1992 to solve binary classification problems, then they are extended to nonlinear regression problems. SVMs are based on structural risk minimization unlike ANNs which is based on empirical risk minimization. They use a nonlinear mapping to transform the input data into a multidimensional feature space. After this transformation the SVM finds the best hyper plane inside the feature space. The nonlinear mapping depends on what so called a kernel function.

3.2. Data Set

The dataset used in this research, is the one used in “UCSD-FICO Data Mining Contest 2009”[]. The competition was organized by University of California, San Diego (UCSD) and FICO a major firm of analytics and decision support in 2009. The data set contains a real-life data of financial truncations of an e-commerce organization.

The competition’s organizer provided two different versions of the data set for the participants (easy and hard); in this research, the hard version of the data set is used in the paper. The data set is composed of 100,000 instances and each instance is composed of 20 features. This data set is highly imbalanced with a ratio of approximately 97:3 towards legitimate transactions, meaning that 3% of the transactions are fraud while the other 97% are legit. Furthermore, features of the data set are anonymised strong enough not to be deduced since it is of course sensitive information.

3.3. Data Pre-processing

The dataset was pre-processed for the purpose of improving the performance of the classifiers and reducing their training and operating time. The pre-processing includes investigating the dataset feature space and handling the imbalanced nature of the dataset.

The data set has 20 anonymized features however it was observed that some of the features are redundant features (they represent the same information), for example the features named state1 and zip1 represent the same information which is the location. It was decided to remove redundant features and consequently the feature space was reduced to 16 features.

Random under sampling technique was applied on the data set to achieve a balanced class distribution in the data. Basically, the process in the random under sampling is to collect random samples from the majority class so that it will have equally number of instances as the minority class. The data set had 100,000 instances; 97346 of them are labelled legit while the other 2654 are labelled fraud. After applying the under sampling technique, the data set has an equal number of instances for each class of 2654. This way classifiers will produce unbiased predictions towards any of the classes since both have an equal number of instances.

3.4. Performance metrics

A number of performance metrics could be used to report the performance of the fraud detection classifiers including the confusion matrix, Sensitivity, Specificity, false positive rate, Balanced classification Rate and Matthews correlation coefficient.
Confusion matrix

A confusion matrix of a binary classifier is a table that shows the number of instances classified correctly/incorrectly in each class. Figure 1 illustrates the confusion matrix of a binary classifier. In the fraud detection problem, positive represents the legitimate transactions and negative represents fraudulent transactions. True positive (TN) represents the number of fraudulent transactions correctly classified as fraudulent. True negative (TP) represents the number of legitimate transactions correctly classified as legitimate. False positive (FP) represents the number of legitimate transactions misclassified as fraudulent. False negative (FN) represents the number of fraudulent transactions misclassified as legitimate.

Table 1. Confusion Matrix

<table>
<thead>
<tr>
<th>Actual classification</th>
<th>Positive (Legit)</th>
<th>Negative (Fraud)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Positive (Legit)</td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td>Predicted Negative (Fraud)</td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>

True Positive Rate

Also known as Sensitivity, it represents the portion of positives which are classified as positives. In fraud detection, it denotes the fraud catching rate; fraud transactions classified as fraud.

\[
TPR = \frac{TP}{P}
\]  

(1)

False Positive Rate

False positive rate represents the portion of negatives which are classified as positives. In fraud detection, it denotes the false alarm rate; legal transactions classified as fraud.

\[
FPR = \frac{FP}{N}
\]  

(2)

Balanced Classification Rate (BCR)
Balanced classification rate represents the average of sensitivity and specificity which is the portion of negatives which are classified as negatives.

\[ BCR = \frac{1}{2} \left( \frac{TP}{P} + \frac{TN}{N} \right) \]  

(3)

**Matthews Correlation Coefficient (MCC)**

Matthews correlation coefficient is an evaluation metric for binary classification problems. MCC is used mainly used with imbalanced data sets because it considers in its evaluation TP, FP, TN and FN. The MCC’s value is somewhere between -1 and +1; +1 represents excellent classification and -1 represents total distinction between classification and observation.

\[ \text{Specificity} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \]  

(4)

**4. Experiments and Results**

The four fraud detection models were trained and tested using Weka. Weka is an abbreviation for “Waikato Environment for Knowledge Analysis”. Weka is a workbench for machine learning that implements the majority of data mining techniques and data pre-processing and filtering techniques. Weka tool was developed in Java language by University of Waikato in New Zealand.

A 10-fold cross validation was used in the process of training and testing the different models. In 10-fold cross validation the data set is divided into 10 subsets; one of them is used as the testing set and the others are used as the training set. This process is repeated taking a different subset as the testing set. The average performance results are then recorded. This methodological approach ensures that all data were represented once as a test data and several times as a training data producing accurate results.

Regarding to the SVM based model Gaussian radial basis function was used as the kernel function which is a general-purpose kernel with good performance results. The cost parameter C and the kernel parameter \( \sigma \) were set to values 10 and 0.1 respectively. These values were selected after experimenting with different combinations of these values.

C4.5 decision tree algorithm was adopted in this paper to develop the decision tree based model. Figure 1 shows the performance of the four trained models. It could be observed that the K-NN based model outperformed other models in terms of the TPR, FPR and BCR; while only the naïve Bayes model outperformed the K-NN in the MCC metric. Naïve Bayes model is the second best one in terms of the four matrices. SVM outperformed the C4.5 model (J48 algorithm is an implementation of the C4.5 in java) in terms of TPR, FPR, BCR. While C4.5 outperformed SVM in terms of MCC. Maybe using other SVM kernels or different settings to the parameters lead to better performance.
5. Conclusion

SVM, Naive Bays, Decision trees and K-nearest neighbours were used in developing four fraud detection models to classify a transaction as fraudulent or legitimate. Four metrics were used in evaluating their performances. The results showed that there is no data mining technique that is universally better than others. Performance improvement could be achieved through developing a fraud detection model using a combination of different data mining techniques (ensemble).

References


http://www.cmo.com/features/articles/2014/5/6/Mind_Blowing_Stats_Online_Shopping.html


http://www.dailynewsegypt.com/2015/05/12/egyptian-e-commerce-market-to-reach-2-7bn-by-2020/


