A Study on False Positive Reduction in Unbalanced Datasets through SVM Using Cost Based Learning

1D. Manjula and 2J. Thilagavathi,
1Research Scholar 2Assistant Professor
1,2Department of Computer Science, NGM College of Arts and Science, Pollachi, India.

Abstract: Fraud Detection in unbalanced Datasets victimization price primarily based Learning uses price complicated Support Vector Machines (CC-SVM) for locating the ineligible transactions. SVM could be a binary classification, therefore the transactions square measure labeled either as dishonest or legitimate. To handle the unbalanced dataset, the formula victimization totally different error price for the positive (+) and therefore the negative (-) categories is employed. SVM weight implements price sensitive learning, just like SVM, the weighted SVM is employed to maximize the margin of separation and minimize the classification error. The margin boundary is employed to separate the categories. In CS-SVM totally different weights square measure assigned to the categories. Effective call boundary is learned by adjusting the weights of the various categories. It improves the accuracy of the prediction rate.

Keywords: E-Payment, Unbalanced dataset, Fraud deduction, Cost based Learning, SVM.

I. INTRODUCTION

The process of electronic payment has been used since 1970’s. Many methods were designed to provide payments. After the internet came into being and after the incorporation of internet into many areas, the world has seen an explosive number of people using it. Payment using electronic means has started becoming a common mode of transaction during the late 90’s. It was during this period when ideas came about for cashless transactions. These served as the basis for the electronic commerce or cashless transactions. This was also the period when electronic commerce came into full view. This got researchers working on the process and many academic studies conducted under this area. Since this process also involved a lot of commercial interest, many commercial agencies were also interested in this area and hence this area of research started bustling with activity. Many ideas were proposed for carrying out these cashless transactions, some of them were even launched into the market. But many failed to reach the targeted audience and hence failed. One of such methods, The Cyber cash launched payment systems. These systems achieved quite extensive deployment but failed to generate an economic return.

At the same time many companies started up new methods of payments for B2C sector [1]. The Electronic Payment (e-payment) is a method of value exchange in electronic commerce, where the value is transferred via the Internet and communication technologies. The electronic payment systems have evolved from traditional payment systems and consequently the two types of systems have much in common. An electronic payment system denotes any kind of network service that includes the exchange of money for goods or services. E-payment is conducted in different e-commerce categories such as Business-to Business (B2B), Business-to-Consumer (B2C), Consumer-to- Business (C2B) and Consumer-to-Consumer (C2C) [2].

The use of credit cards has become one of the common tasks in the life of an average person. The usage of credit cards is one of the easier modes of transactions; hence it is also the most flexible and the easy way of payment. Detection of frauds performed using credit cards is a tedious as well as a crucial process. Since it also involves the customer’s interests and flexibility, this process should be performed carefully. This has recently become a study of great importance due to the increase in the usage of credit cards and increase in the frauds detected. Even though this process is of high priority, handling this in an efficient way is very important, since a large customer base is involved. A simple flaw in the detection process might prove to be a heavy loss[9][10]. There are many ways in which a credit card fraud takes place. In general these frauds can be categorized into two broad classifications:

A. Physical Card

This type of fraud takes place when the actual card is in possession of the adversary. The cardholder either loses the card or his card is stolen and is then used by somebody else. This is the most common type of fraud that occurs. If the actual cardholder does not realize that his card is missing, then he would experience a large amount of loss. When a fraudster gets hold of such a card, he would attempt to perform large amount of transactions and purchases with very high value. These also take place in a very short span of time. Hence it shows a large deflection in the purchase pattern and the purchase volume when compared to the actual transactions carried out by the
actual cardholder. Hence this can be easily detected with a simple analysis system.

**B. Virtual Card**

This type of fraud is more difficult to detect, since the actual physical card is in possession of the cardholder. This type of fraud occurs when the fraudster gets hold of the card’s credentials instead of the actual physical card. This is possible when the user provides the credentials online in an unsecure payment site or in a spoofed site. This can also occur when the fraudster possess counterfeit cards. This process is much more difficult to detect, since the actual user is not aware of another person using his card. If the fraudster makes purchases that are more similar to the actual user, then even an analysis system would not be able to detect the malicious transaction. This type of fraud comes to light only after the examination of the monthly statement by the actual cardholder[5][8].

One of the major problems facing the fraud detection is the lack of the availability of information regarding the credit card details and the frauds associated with them[8]. The details about credit card transactions are a set of vulnerable data, hence not many banks are willing to offer these information publicly. Even when such kind of information is offered, some significant parameters are eliminated from the dataset. Further, some parameters are substituted with dummy values for maintaining the confidentiality of the data. Only a very few publications provide valuable contributions in this field. Even these publications do not perform their experiments using the absolute data, instead, these experiments were performed on artificial data generated, that closely resembles the original data or real time datasets with dummy data as significant parameters. Hence this process does not prove to be much useful[8][9][10].

**II. METHODOLOGY**

The available data cannot be directly used by the SVM. SVM uses only numerical data under defined boundaries. But the data presented by the user will usually contain numerical and categorical attributes. These attributes must be converted to numerical format to be used by the SVM. Hence the preprocessing phase plays an important role in the initial phase. This cannot be performed directly by the system, since categorical attributes are involved. The user must provide corresponding values for the categorical attributes that are present in the data set. Further, SVM has the tendency to process data only within a certain defined range (-1 to +1) or (0 to 1). Hence the next process is to convert all the available numerical values into values that come under a certain range. This process is called normalization. The boundary values are defined by the SVM. All the numerical attributes are then normalized to the corresponding boundary values using the Min-Max Normalization. In our current process, we use the boundary values of -1 to +1.

Min-Max Normalization is used to perform the normalization process. Min Max Normalization transforms a value A to B which fits in the range[C,D]. It is given by the formula

\[
\text{Normalized value (B)} = \frac{A - \text{Min}}{\text{Max} - \text{Min}} \times (D - C) + C
\]

Each available data is passed on to the Normalization function to obtain its corresponding normalized value. This process converts the data into an easily understandable and easily interpretable format, hence becomes easier for the algorithms during the comparative analysis of the data.

After this process, using the normalized data, the training and testing data files are created. The SVM requires a special format for reading the data. The expected format of input for an SVM is

[label] [index1]:[value1] [index2]:[value2] ...

Where label represents the final classified label, index values range from 1,2...n and value1, value2.... represents the corresponding column values.

**Figure 1: Fraud Detection using SVM**

**Figure 2: Sample input data for SVM**

The formatted data is then divided into two different files, the training set and the test set. The training set contains most of the available data, while the test set contains some sample data of all formats.
During the training phase, the training data file is passed to the SVM. In this current implementation we use the RBF kernel for performing the training. The RBF kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. The SVM with the RBF kernel is then fine tuned by changing the two parameters C and γ. A v fold cross validation is performed for the fine-tuning process.

Algorithm:
1. Obtain transaction data
2. Preprocess data to convert categorical attributes to numerical attributes
3. Normalize the numerical data using Min-Max Normalization
4. Create training and testing data files using the SVM format
5. Supply training file to SVM
6. Set the values for C and γ
7. Obtain results using the current C and γ pair
8. Perform step 6 and 7 till satisfactory results are obtained from the training set
9. Test the accuracy using the test file

Setting the appropriate values for C and γ plays a crucial role in determining the accuracy of the system. Every type of data requires a different combination of C and γ which should be determined by the user. This makes the SVM approach a supervised methodology[3][6].

The advantages of using SVM are, by introducing the kernel, SVMs gain flexibility in the choice of the form of the threshold separating normal data from the outliers, which need not be linear and even need not have the similar practical form for all data, since its function is non-parametric and operates locally. Since the kernel implicit lycovers a non-linear transformation, no assumptions about the functional form of the transformation, which makes data linearly separable, is necessary. The transformation occurs implicitly on a robust theoretical basis and human expertise judgment beforehand is not needed. SVMs provide a good out-of-sample generalization, if the parameters C and γ are appropriately chosen. This means that, by choosing an appropriate generalization grade, SVMs can be robust, even when the training sample has some bias. SVMs deliver a unique solution, since the optimality problem is convex. This is an advantage compared to Neural Networks, which have multiple solutions associated with local minima and for this reason may not be robust over different samples. SVMs can produce accurate and robust classification results on a sound theoretical basis, even when input data are non-monotone and non-linearly separable. So they can help to evaluate more relevant information in a convenient way. Since they linearize data on an implicit basis by means of kernel transformation, the accuracy of results does not rely on the quality of human expertise judgment for the optimal choice of the linearization function of non-linear input data.

III RESULTS AND DISCUSSION

In the initial detection Model, the unsupervised approach is used. The unknown frauds are easily found by using this approach. The models based on supervised approach must have the labeled data for both normal data and anomalies. It is only able to detect frauds of a type which has previously occurred. In contrast, unsupervised methods don’t make use of labeled records. It detects the changes in behavior or unusual Transactions. Unsupervised learning is a feasible method to learn the large and more complex model.

Applying the algorithms in the dataset, will reduce the number of nearest neighbor searches and number of reach ability distance computation. This model helps detect fraudulent transactions in an efficient manner.

The costs for positive class is always 1 and for the negative class is ratio of negative samples over positive samples are used. Asymmetric error function is used to control the tradeoff between the negative and positive instances. Training samples with different class ratio are used in this work. Increasing the size of the negative classes will not affect the result. Moreover all different samples predict the similar accuracy rate.

CONCLUSION

This provides associate degree economical means that to find frauds in a very customer's transactions. The presently projected system has the power to find frauds performed by each physical and virtual kind of frauds. While this approach is effective in many ways, there exist certain issues that act as a downside to this approach. They are, an SVM can be applied to only two-class tasks, uncertainty is taxonomies exist and the factors of the explained model are difficult to implement. Since the problem at hand demands high accuracy in determining fraudulent transactions, a better approach can be adopted for obtaining highly accurate results. A common disadvantage of non-parametric techniques such as SVMs is the lack of transparency of results. Further SVMs have a very high false positive rates. This acts as an important issue, because marking a legal transaction as illegal and acting on it proves to be more costly than marking a fraudulent transaction as legitimate. Since customer’s feelings are involved, one wrong step might prove to be a disaster for the organization. Hence the unreliable nature of the SVM
requires a better verification mechanism to deal with these issues to provide better results.

References


