ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 7, July 2013

Anomaly Detection using Support Vector Machine

Dharminder Kumar[1], Suman[2] , Nutan[3]
[1] GJUS&T Hisar
[2] Research Scholar, GJUS&T Hisar
HCE Sonepat
[3]HCE Sonepat

Abstract: Support vector machine are among the most well known supervised anomaly detection technique, which are very efficient in handling large and high dimensional dataset. SVM, a powerful machine method developed from statistical learning and has made significant achievement in some field. This Technique does not suffer the limitations of data dimensionality and limited samples. In this present study, We can apply it to different domains of anomaly detection. Support vectors, which are critical for classification, are obtained by learning from the training samples. Results of SVM achieved high Accuracy and low false positive rate. Theoretically we compared our approach with neural network and clustering technique

Keywords: **anomaly detection, Data Mining, fraud, support vectors**

1. INTRODUCTION

Anomaly Detection is the process where to localize objects that are different from other objects. It is a technique for improving the analysis of typical data objects [9].These anomalous objects are exceptional in some sense. It lie far away from other data points (outliers) and it have attribute values that deviate significantly from the expected or typical attribute values that indicate errors in data. Data mining is a field that will helps us to detect anomalies. The overall goal of data mining process is to extract information from a data set.

Various data mining techniques are discussed in [1, 3, 4]. Support vector machine is a classification technique [6] of data mining which is used for anomaly detection. Several unique features of these algorithms make them especially suitable for binary classification problems like fraud detection [8] which is application of anomaly detection. Fraud can arise from potentially undetected fraud transactions. Our study is based on real-life data of transaction from an international credit card operation. SVMs are linear classifiers that work in a high-dimensional feature space that is a non-linear mapping of the input space of the problem at hand. An advantage of working in a high-dimensional feature space is that, in many problems the non-linear classification task in the original input space becomes a linear classification task in the high-dimensional feature space [12].

The remainder of the paper is organized as follows. Section 2 reveals the existing literature for different types of credit card fraud detection. Next section represents the anomaly detection

methods. Section 4 represents the different data mining techniques (SVM, Clustering and Neural Network). Section 5 will discuss about the results, which we have obtained. Next Section gives the theoretical comparison between data mining techniques. . At the last section of the paper we derive some conclusion.

2 LITERATURE SURVEY

At the early stage, the research focus lies in using rule-based expert systems and statistical approaches. But when encountering larger datasets, the results of rule-based expert systems and statistical approaches become worse. Thus, many data mining techniques have been introduced to solve the problem. Among these techniques, the Artificial Neural Network (ANN) is widely used and has been successful in solving many complex practical Problems [15]. From the work of view for preventing credit card fraud, more research works were carried out with special emphasis on data mining and neural networks. Sam and Karl [16] suggest a credit card fraud detection system using Bayesian and neural network techniques to learn models of fraudulent credit card transactions. Kim and Kim have identified skewed distribution of data and mix of Legitimate and fraudulent transactions as the two main reasons for the complexity of credit card fraud detection [17].

3 ANOMALY DETECTION

Anomaly detection refers to detecting patterns in a given dataset that do not conform to an established, normal behavior. The patterns thus detected are called anomalies, and often translate to critical and actionable information in several application domains [11]. Anomalies are also referred to as outliers, surprise, aberrant, deviation, peculiarity; etc. Anomaly detection is used in a variety of domains, such as intrusion detection, fraud detection and system health monitoring [9, 10]. Three broad categories of anomaly detection techniques exits:

1. Supervised anomaly detection techniques learn a classifier, using labeled instances belonging to normal and anomaly classes, and then assign a normal or anomalous label to a test instance.
2. Semi-Supervised anomaly detection techniques construct a model representing normal behavior from a given normal training data set and then test the likelihood of a test instance to be generated by the learnt model.
3. Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set.

Our method for anomaly detection is a supervised method based on support vector machine, so firstly we should explain it.

4 TECHNIQUES FOR FRAUD DETECTION

There are various techniques used for fraud detection. Few of them are explained below.

4.1 Support Vector Machine

The Support Vector Machine (SVM) was first proposed by Vapnik and has since attracted a high degree of interest in the machine learning research

community. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. Sims have been employed in a wide range of real world problems such as text categorization, hand-written digit recognition, tone recognition, image classification and object detection, micro-array gene expression data analysis, data classification[1,5,7]. It has been shown that Sims is consistently superior to other supervised learning methods. However, for some datasets, the performance of SVM is very sensitive to how the cost parameter and kernel parameters are set. As a result, the user normally needs to conduct extensive cross validation in order to figure out the optimal parameter setting.

SVMs are set of related supervised learning methods used for classification and regression they belong to a family of generalized linear classification. A special property of SVM is, SVM Simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM called Maximum Margin Classifiers. SVM is based on the Structural risk Minimization (SRM). SVM map input vector to a higher dimensional space where a maximal separating hyper plane is constructed. Two parallel hyper planes are constructed on each side of the hyper plane that separate the data. The separating hyper plane is the hyper planes that maximize the distance between the two parallel hyper planes. An assumption is made that the larger the margin or distance between these parallel hyper planes the better the generalization error of the classifier will be .We consider data points of the form

$\{(X_1, Y_1), (X_2, Y_2), (X_3\ Y_3), (X_4,Y_4)\ldots\ldots\ldots, (X_n, Y_n)\}$.

Where Y_n=1 / -1, a constant denoting the class to which that point Xn belongs.
n = number of sample. Each X_n is P - dimensional real vector. The scaling is important to guard against variable (attributes) with larger variances. We can view this training data, by means of the dividing hyper plane, which takes
W. X + b = O ----- (1) Where b is scalar and W is p-dimensional Vector. The vector W points perpendicular to the separating hyper plane. Adding the offset parameter b allows us to increase the margin. Absent of b, the hyper plane is forced to pass through the origin, restricting the solution. As we are interesting in the maximum margin, we are interested SVM and the parallel hyper planes [11]. Parallel hyper planes can be described by equation

W.X + b = 1
W.X + b = -1

If the training data are linearly separable, we can select these hyper planes so that there are no points between them and then try to maximize their distance. By geometry, we find the distance between the hyper planes is $2 / |w|$. So we want to minimize $|w|$. To excite data points, we need to ensure that either W. $(X_I - b \geq 1)$ or $(W. X_I - b \leq -1)$.
This can be written as $Yi\ (W. X_I - b) \geq 1$, $1 \leq i \leq n$ ------ (2)

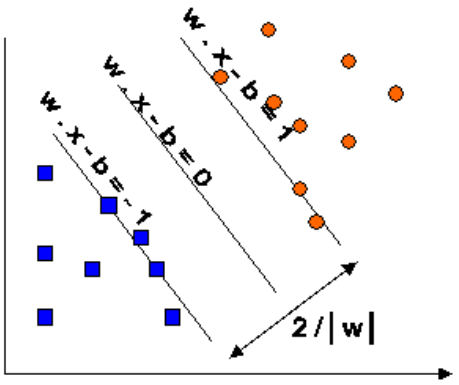

Fig.1 Maximum margin hyper planes for a SVM trained with samples from two classes

A separating hyper plane with the largest margin defined by M = 2 / $|w|$ that is specifies support vectors means training data points closets to it.

$Y_j [W^t . X j + b] = 1,$ i =1 ----- (3)

Optimal Canonical Hyper plane (OCH) is a canonical hyper plane having a maximum margin. For all the data, OCH should satisfy the following constraints

Yi $[W^t. Xj + b] >=1;$ -------- (4)

i=1, 2 ….n In order to find the optimal separating hyper plane having a maximum margin, A learning machine should minimize $\|w\|^2$ subject to the inequality.

4.2 Neural Networks

Neural networks have been widely used in fraud detection. Neural Networks can actually calculate user profiles in an independent manner, thus adapting more elegantly to the behavior of the various users. Neural networks are claimed to substantially reduce operation costs [13]. Neural networks are specialized computer software that produces non-linear models of complex problems in a fundamentally different way. From a large database, neural networks can develop a set of rules to recognize and predict certain conditions. Neural Networks works best at recognizing, predicting and controlling patterns, such as in fraud detection, payment reviews and other areas where large amounts of data are gathered.

4.3 Clustering

Clustering helps in grouping the data into similar clusters that helps in uncomplicated retrieval of data [14]. Cluster analysis is a technique for breaking data down into related components in such a way that patterns and order becomes visible. This model is based on the use of the parameters' data cauterization regions. In this system 24 parameters of transactions are used for classification

5. RESULTS

This section presents the results of our approach. The Various performance measures are defined with respect to the Confusion matrix given below. We use True positive Rate, False positive rate and accuracy as performance measure.

Table 1: Confusion matrix

	Predicted positive	**Predicted negative**
Actual positive	True positive (TP)	False negative (FP)
Actual negative	False positive(FP)	True negative (TN)

True positive Rate represents the fraction of deceitful transaction correctly

identified as deceitful and genuine transactions correctly identified as genuine. False positive Rate represents the fraction of genuine transaction identified as deceitful and deceitful transactions identified as genuine. Accuracy denotes the closeness of computations or estimates to the exact or true values.

True Positive Rate=TP/TP+FN

Accuracy=TP+TN/TP+TN+FP+FN

False Positive rate=FP/FP+TN

Fig 2 shows different types of data having different fraud rates.data1, data2, data3 and data4 having transactions 1000, 1500, 2000, and 2500.As higher the data, there is more accuracy, tp rate and less fp rate.

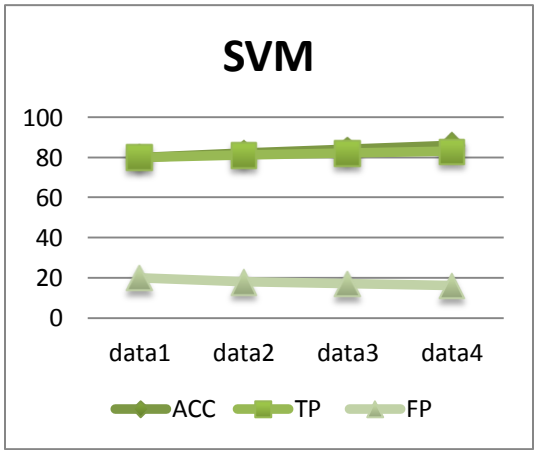

Fig 2. Fraud rates using different data by SVM

6. COMPARISON

In Neural network and Clustering a number of parameter has to be set before any training can begin. However, there are no clear rules how to set these parameters, while SVM has no need to set any parameters. SVM is domain independent. Its works on high dimensional feature space, while others techniques cannot do so. Cost of clustering is very high, due to any samples of deceitful and genuine transactions, to classify new transactions as deceitful or genuine its takes more processing steps and time, while SVM takes less processing steps and time. So, SVM is less expensive than clustering and Neural Network.

7. CONCLUSION

This paper analyzes the feasibility of credit card fraud detection based on anomaly detection. Today anomaly detection methods are of major interest to the world and are used in very different and various domains like computer intrusion detection, credit card and telephone fraud detection. And finally this method proves accurate in predicting fraudulent transactions through support vector machine emulation experiment from an international credit card operation. If this algorithm is applied in to bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions by the banks. Here, we have introduced a new supervised method for anomaly detection, Support vector machine that fuse information from various sources.

References

[1] Mrs. Bharati M. Ramageri, “data mining techniques and applications”, Indian Journal of Computer Science and Engineering Vol.1No. 4 301-305, 2010.

[2] Julie M. David1, Kannan Balakrishnan2,“Prediction of Learning Disabilities in School Age Children using SVM and Decision Tree”, International Journal of

Computer Science and Information Technologies, Vol. 2 (2) , 2011.

[3] Sarojini Balakrishnan, Ramaraj Narayanaswamy, "An empirical study on the performance of integrated hybrid prediction model on the medical datasets", International journal of computer applications (0975 – 8887) Volume 29– No.5, September 2011.

[4] Sonika Jalhotra, "Mining Techniques Used for Financial Organization", International Journal of Engineering Research and Development, Volume 2, Issue 12, August 2012.

[5] Khalid raza, "application of data mining in bioinformatics", Indian journal of computer science and engineering vol 1 no 2, 114-118, 2009.

[6] V.Vapnik, Statistical Learning Theory, Wiley, New York, 1988.

[7] N.Cristianini, j.Shawe-Taylor, An Introduction to support vector machines and other kernel based learning methods, Cambridge university press, 2000.

[8] Siddhartha Bhattacharyya, Sanjeev jha et al.,"data mining for credit card fraud: A Comparitive study", Decision support System 50(2011) 602-613.

[9] Greensmith,et al.,"Information Fusion for anomaly detection with the dendritic cell algorithm",Information Fusion(2007).

[10] M.Lotfi Shahreza, et al.,"anomaly Detection using a self-organization map and particle swarm optimization" Sharif University of Technology, 2011.

[11] Durgesh K. Srivastava, Lekha Bhambhu,"Data Classification using support vector machine", Journal of Theoretical and Applied Information Technology,2009.

[12] Chih-Wei Hsu, Chih-chung chang,et al.,"A Practical Guide to Support Vector classification".Deptt. of computer science, national Taiwan University,Taipei,106,2007.

[13] Yufeng Kou, et al,"Survey of Fraud detection Techniques" International Conference on Networking, sensing & Control, Taipel,Taiwan,March21-23,2004.

[14] Dr.R.Dhanapal, "An intelligent information retrieval agent", Elsevier International Journal on Knowledge Based Systems 2008.

[15] Wang, Gang, Hao, Jinxing, Ma, Jian and Huang, Lihua ''A new approachto intrusion detection using artificial neural networks and fuzzy clustering'', Original Research Article Expert Systems with Applications, 37(9),pp. 6225–6232 (2010).

[16] Sam Maes, Karl Tuyls, Bram Vanschoenwinkel, Bernard Manderick. Credit Card Fraud Detection Using Bayesian and Neural Networks First International NAISO Congress on Neuro Fuzzy Technologies, Havana, Cuba. 2002.

[17] M.J. Kim and T.S. Kim, "A Neural Classifier with Fraud Density Map for Effective Credit Card Fraud Detection,"Proc. International Conference on Intelligent Data Engineering and Automated Learning, Lecture Notes inComputer Science, Springer Verlag, no. 2412, pp. 378-383, 2002.