Unbalanced Data Classification Using *extreme outlier* Elimination and Sampling Techniques for Fraud Detection

T. Maruthi Padmaja[1], Narendra Dhulipalla[1], Raju S. Bapi[2], P.Radha Krishna[1]
1.Institute for Development and Research in Banking Technology(IDRBT), Hyderabad-500 057, India.
{tmpadmaja,dnarendra,prkrishna}@idrbt.ac.in
2. Dept of Computer and Information Sciences, University of Hyderabad (UoH), Hyderabad-500 046, India
bapics@uohyd.ernet.in

Abstract

Detecting fraud from the highly overlapped and imbalanced fraud dataset is a challenging task. To solve this problem, we propose a new approach called extreme outlier elimination and hybrid sampling technique. k Reverse Nearest Neighbors (kRNNs) concept used as a data cleaning method for eliminating extreme outliers in minority regions. Hybrid sampling technique, a combination of SMOTE to over-sample the minority data (fraud samples) and random under-sampling to under-sample the majority data (non-fraud samples) is used for improving the fraud detection accuracy. This method was evaluated in terms of True Positive rate and True Negative rate on the insurance fraud dataset. We conducted the experiments with classifiers namely C4.5, Naïve Bayes, k-NN and Radial Basis Function networks and compared the performance of our approach against simple hybrid sampling technique. Obtained results shown that extreme outlier elimination from minority class, produce high predictions for both fraud and non-fraud classes.

Keywords: Data Mining, Unbalanced dataset, *k*RNN, Hybrid Sampling, SMOTE and Fraud Detection.

1.Introduction

Fraud is pervasive in today's society. It is the illegitimate activity by a person other than eligible person. This type of fraud poses a big problem for many industries like credit card, insurance and telecommunications. Fraud can be done in various ways and each *modus operandi* is different from the other, fraud detection must involve identifying fraud as quickly as possible once it has been perpetrated.

By using the massive amounts of data on financial transactions, we can identify the fraud patterns using data mining methods. Fraud detection poses some technical and practical problems for data mining; due to poor quality of the data itself, lack of domain knowledge. And another crucial technical dilemma is due to the highly unbalanced data for fraud detection. Typically there are many more legitimate than fraudulent samples. Hence, appropriate classification approaches are needed to classify the unbalanced data, especially for fraud detection problem

Even though fraud detection is a binary classification problem, in reality it is *n*-class problem, as each fraud is different from the other. In this work, we considered fraud detection as highly overlapped unbalanced data classification problem, where non-fraud samples heavily outnumber the fraud samples. Usually, the classification algorithms exhibit poor performance while dealing with unbalanced datasets and results bias towards majority class. For this type of problems, accuracy is not an appropriate measure because the cost associated with fraud sample is predicted, as non-fraud sample is very high. So the performance measures could be used here are cost based metrics and ROC analysis.

In this paper, we propose a new method for fraud detection, which uses *extreme outlier* elimination using *k Reverse Nearest Neighbors (k*RNNs) [12] approach with hybrid sampling technique. Also we compare our approach with the hybrid sampling that does not use any outlier elimination of minority samples.

The rest of this paper is organized as follows. In section 2 we present the related work done from two perspectives. Section 3 describes the proposed approach and experimental setup. Section 4 provides the results and discussion. In section 5 we conclude the paper.

0-7695-3059-1/07 $25.00 © 2007 IEEE
DOI 10.1109/ADCOM.2007.74

2.Related Work

The fraud detection problem has been approached from two directions of *how to handle unbalanced data* (i) Fraud detection using traditional approaches (ii) Sampling techniques.

Fawcett and Provost [1] have shown that adaptability to dynamic patterns of fraud can be archived by generating the fraud detection systems automatically using data mining techniques. Clifton [2] categorizes, compares and summarizes relevant data mining based fraud detection methods. Stolfo *et al* [5,6] outlined a meta-classifier system for detecting the fraud by merging the results obtained from local fraud detection tools at different corporate sites to yield a more accurate global tool. Chan *et al* [3,7] further elaborated the meta-classifier system to a distributed data-mining model. It is a scalable, supervised black box approach, uses realistic cost model to evaluate classification models. Their results demonstrated that using stacking to combine multiple models generated by different algorithms on the subsets of fraud/non-fraud data distributions significantly improve cost savings. Neural data mining approach [15] uses generalized rule based association rules to mine the symbolic data and Radial Basis Function networks to mine the analog data. It has been found that using supervised neural networks to check the results of association rules increases the predictive accuracy. Wheeler and Aitken [8] have also explored the combination of multiple classification techniques. Clifton *et al* [4] proposed a fraud detection method, which uses stacking-bagging approach to improve cost savings.

There have been several sampling approaches for coping with unbalanced datasets. Kubat and Matwin [9] did selective under-sampling of majority class by keeping minority classes fixed. They categorized the majority samples into some noise overlapping, the positive class decision region, borderline, redundant and safe samples. Excluding safe samples, they used all other categories of negative samples for under-sampling. Ling and Li [14] combined over-sampling of the minority class with under-sampling of the majority class. They used lift index instead of accuracy to measure a classifier's performance. But this combined approach did not show much improvement in lift index than the under-sampling approach, exhibited good lift index at equal class distributions for direct marketing problem. Chawla et al [10] proposed Synthetic Minority Over-sampling TEchnique (SMOTE). It is an over-sampling approach in which the minority sample is over-sampled by creating synthetic (or artificial) samples rather than by over-sampling with replacement. The minority class is over-sampled by taking each minority class sample and introducing synthetic samples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. This approach effectively forces the decision region of the minority class to become more general. Since this technique creates new minority instances artificially, they found this technique to be more useful than simple random over-sampling. Batista *et al* [17] proposed two hybrid sampling techniques for overlapping datasets namely SMOTE+TOMEK Links and SMOTE+ENN for better-defined class clusters among majority and minority classes. Estabrooks *et al* [11], combined different expressions of the resampling approaches in order to get optimal class distribution from the original unbalanced class distribution.

3.Proposed Approach

Here our basic motivation is to improve the fraud recognition from the fraud dataset, since their presence is less and sparse. For this, we used hybrid sampling technique proposed by Chawla et. al [10].The hybrid sampling technique is a combination of SMOTE and Random Under-sampling technique. As mentioned in Section 2, SMOTE is used for generating the synthetic samples across the minority class. Random under-sampling is used for balancing the training data, class distribution.

If we use SMOTE on the entire fraud class samples, fraud regions may not be well defined, because of sparsely located fraud samples. The blind generation of synthetic samples along the sparse fraud samples, resulting a greater chance of class mix. Hence the minority samples may not recognize well. To avoid this class mix in training data distribution, we use *extreme outlier* elimination from the minority class by using kRNN concept as a data cleaning method. The kRNNs cardinality value represents whether the point located in a sparse region or in a dense region. We called the points that are very sparsely located are *extreme outliers*. By eliminating the *extreme outliers*, we are ignoring the points that are far from the minority decision boundary for doing SMOTE.

In the following subsection, we demonstrate the application of kRNN in addressing the problem of extreme *outlier* elimination for unbalanced datasets.

3.1 Notations and Definitions

Following are the notations used in this paper.

X: *d*-dimensional dataset, **n**: size of dataset
$\mathbf{X}=\{x_1,x_2,x_3,\ldots x_i,x_j,\ldots,x_n\}$
$\mathbf{x_i,x_j}$: Any two points(records) in X
$\boldsymbol{k}\mathbf{NN}(\boldsymbol{x_i})$: set of k-nearest neighbors of $x_i \triangleright \psi$
$\mathbf{d_{ij}}$: distance between two points x_i and x_j.
k **nearest neighbor set** - kNN(x_i) is defined as $\{ x_j \mid d_{ij} <$ k-th nearest distance of $x_i \}$. For given point x_i, the k-th smallest distances after sorting all the distances from x_i to the remaining points.
$\boldsymbol{k}\mathbf{RNN}(\boldsymbol{x_i})$: set of k-reverse nearest neighbors of x_i . A point x_j belongs to kRNN(x_i) iff $x_i \psi \in \leftarrow k\text{NN}(x_j)$.
k **reverse nearest neighbor set** kRNNs(x_i) -- defined as $\{ x_j \mid x_i \in k\text{NN}(x_j) \}$

kRNNs set defines influence around a point in a dataset and these are used to capture the neighborhood of a point. Note that, in case of *k*NNs, for a given *k* value, each point in the dataset will have atleast *k* nearest neighbors (> k in case of ties), but the *k*RNNs set of a point could have zero or more elements. The *k*RNNs set of point x_i gives a set of points that consider x_i as their *k*-nearest, for a given value of *k*. If a point x_i has higher number of *k*RNNs than another point x_j, then we can say that x_i has a denser neighborhood than x_j. Lesser the number of *k*RNNs, the farther apart are the points in the dataset to x_j, i.e. the neighborhood is sparse.

According to *k*RNN concept [12], *outlier point* is defined as follows:
An outlier point is a point that has less than *k* number of *k*RNNs, i.e., $|k\text{RNNs}(x_i)| < k$. Lesser the number of *k*RNNs, the more distant the point from its neighbors.

For overlapped and unbalanced datasets, many minority samples may have less than *k* number of *k*RNNs. In order to find out the minority samples that are far from the minority decision regions, in this work, we define the concept called *extreme outlier*.

An *extreme outlier* point is a point that has number of *k*RNNs far less than *k,* when *k* value is increased from 1 to 3*n/4, i.e., the points which are distinct from the rest of all points.

For example, we can say a point x_i as *extreme outlier* if its $|k\text{RNNs}(x_i)|=3$ at k=75 for n=250.

After identification of *extreme outliers* from the fraud class, we eliminated the *extreme outliers* from the original fraud data. Then we applied hybrid sampling approach on entire training dataset for fraud detection. The hybrid sampling approach is a combination of random under-sampling and over-sampling. It mainly works based on determining how much percentage of fraudulent samples (original fraud + artificial fraud samples) and non-fraud samples to add to the training set such that a classifier can achieve best True Positive (*TP*) and True Negative (*TN)* rates, i.e., to get an optimal class distribution. Here, *TP* is the number of fraud samples correctly classified and *TN* is the number of non-fraud samples correctly classified

.

4.Experimental Results and Discussion

Experiments are conducted on an insurance dataset [4]. The data set pertains to automobile insurance and it contains 15421 samples, out of which 11338 samples are from January-1994 to December-1995 and remaining 4083 samples are form January-1996 to December -1996. There are 30 independent attributes and one dependent attribute (class label). Here, six are numerical attributes and remaining are categorical attributes. The class distribution of non-fraud and fraud is 94:6, which indicates that data is highly unbalanced. We discarded the attribute *PolicyType*, because it is an amalgamation of existing attributes *VehicleCategory* and *BasePolicy*. Further, we created three attributes namely *weeks_past*, *is_holidayweek_claim* and *age_price_wsum* to improve the predictive accuracy of classifiers as suggested in [4]. So the total numbers of attributes used are 33. All the numerical attributes are discrete in nature and thus converted into categorical attributes. While doing SMOTE, we used Value Difference Metric (VDM) distance measure, which is suitable for categorical attributes in finding the nearest neighbors.

Attributes with two category values like *witness present*, *agent type* and *police report filed* have highly skewed values where majority class samples account for more than 97% of total samples. The attribute *make* has a total of 19 possible attribute values of which claims from 5 attribute values account for almost 90% of total samples. Average number of claims per month is 430.

We implemented the *extreme outlier* detection using *k*RNNs and SMOTE in MATLAB7.0 and used Weka3-4[16] toolkit for experimenting with the classifiers. Figure 1 shows the process of generating samples for training the classifier. Here, we have used Original Fraud Data Percentage (OFD percentage or rate) for identifying the optimal class distribution from original data. Initially, fraud and non-fraud samples are separated from the dataset and we eliminate *extreme outlier*s in the fraudulent samples using the method described in section 3. For the dataset under consideration, total number of minority samples is 922. So we varied *k* from 10,20,30,50,100,200,300 to 400 and found that 92 points qualify as *extreme outlier*s and eliminated them from the dataset. Then applied SMOTE on new set of fraudulent samples for the given level of SMOTE factor. For example, if we specify SMOTE factor as 5 and input fraud samples are x, then artificial fraud samples generated after SMOTE are 5x.

Generally, the choice of optimal SMOTE factor is data dependent. Since the dataset has 94:6 classes unbalance distribution, for experiments, we chose 5,7,9,11 and 13 are SMOTE factors. Similarly for each SMOTE factor, we varied the OFD rate, i.e., number of original fraud samples to be added to training ranging from 0 to 75. Then non-fraud samples are randomly under sampled from non-fraud dataset in such a way that class distribution for training becomes 50:50. Thus, the training dataset is an amalgamation of artificially generated fraud samples, original fraud samples and non-fraud samples. In this work, we conducted experiments on four classifiers namely C4.5, Naïve Bayes, *k*-NN and Radial Basis Function networks.

We computed TP rate and TN rate by varying OFD percentage for each SMOTE factor. Here, testing of each experiment was done against entire dataset. We set *k*, number of nearest neighbors to be selected while doing SMOTE, as 7, 9, 11 and 13 for each SMOTE factors 5,7,9,11,13 and 15 respectively. Total experiments conducted are 100; 25 for each classifier by doing different levels of SMOTE and varying OFD rate.

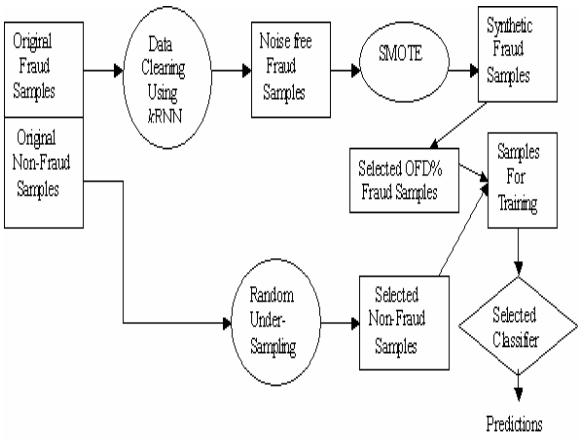

Figure 1: Generation of samples for training

In this work, we compared two methods, (Method-a): *k*RNN based *extreme outlier* elimination combined with hybrid sampling and (Method-b): Hybrid sampling. Figures 2 to 9 show the results obtained in the two methods. Here the dashed lines indicate the results obtained with method-a and solid lines indicate the results obtained with method-b. Each line with different tick mark symbol represents unique SMOTE factor i.e. number of artificial fraud samples added to the training set.

Our observations from the experiments conducted on Method-a are as follows: C4.5 has shown good fraud catching rate and non-fraud catching rate in all the 25 experiments and these rates increased with increase in SMOTE factor and OFD rate (Figures 2 and 3). In all the experiments, it has given above 90% fraud catching rate. We observed best fraud catching rate of 99.9% with *k*-NN classifier (Figure 6). This may be due to the inherent use of *k*-NN, while doing SMOTE to generate the artificial fraud samples. But the non-fraud catching rate of this classifier is below 85% (Figure 7), which is not an effective value. For Naïve Bayes classifier, fraud catching rate is good which is about 85%, but it has not shown much improvement with increase in SMOTE factor and OFD rate (Figure 4). Non-fraud catching rate of this classifier is also found to be not much effective and the value is below 80% (Figure 5). RBF has recorded good fraud catching rate of above 85% in all the experiments (Figure 8). This rate is increased with increase in SMOTE factor and OFD rate in most cases. For this classifier, non-fraud catching rate is also equally good (Figure 9).

By comparing Method-a, Method-b, we observed the following: In catching the fraudulent samples, method-a performed well compared to method-b on all classifiers except *k*-NN. C4.5 has recorded good improvement in fraud catching rate for method-a compared to method-b (Figure 2). This classifier has shown good fraud catching rate of 90% at lower SMOTE factor values and OFD rates itself. Naïve Bayes classifier has suffered from low fraud catching rate in method-b (Figure 4).

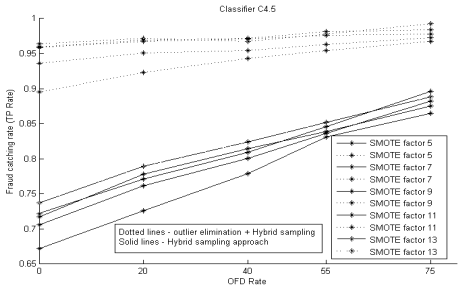

Figure 2: TP rate Vs OFD rate for

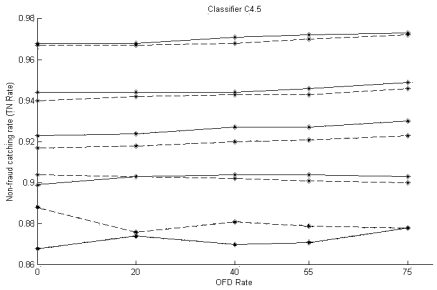

Figure 3: TN rate Vs OFD rate for C4.5

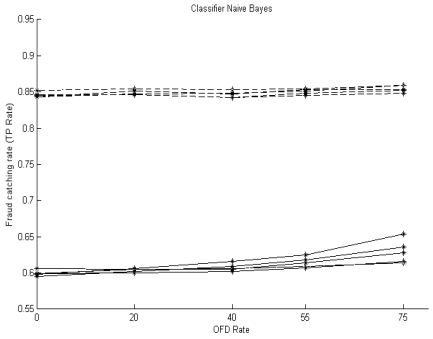

Figure 4: TP rate Vs OFD rate for Naïve Bayes

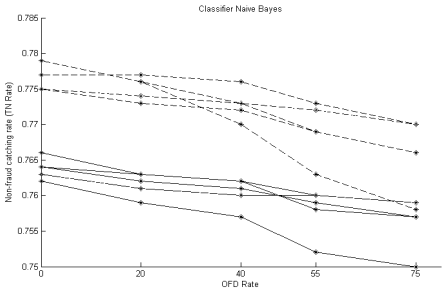

Figure 5: TN rate Vs OFD rate for Naïve Bayes

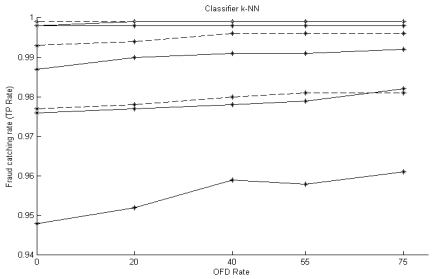

Figure 6: TP rate Vs OFD rate for k-NN

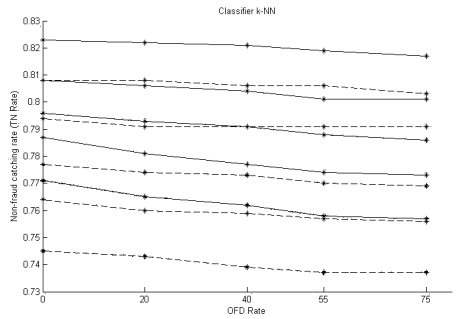

Figure 7: TN rate Vs OFD rate for k-NN

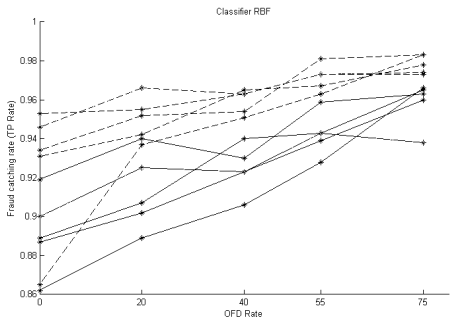

Figure 8: TP rate Vs OFD rate for RBF

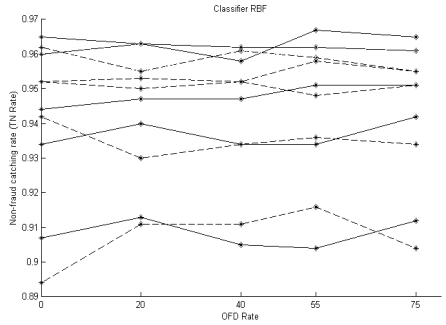

Figure 9: TN rate Vs OFD rate for

But with the proposed approach, it has recorded high fraud catching rate. Moreover, for this classifier, there is reasonable improvement of non-fraud catching rate with method-a (Figure 5). We observed that k-NN performed slightly well for method-b (Figures 6 and 7), however there is not much reduction in performance. RBF has also shown good improvement in fraud catching rate with our approach compared to method-b (Figure 8), but there is no much impact on non-fraud catching rate (Figure 9). Overall, the behavior of each classifier is the same for two methods with variation of SMOTE factor and OFD rate.

5.Conclusion

This paper, demonstrates the significance of eliminating *extreme outliers* from the minority samples for highly skewed imbalanced data sets. For defining

*extreme outliers, k*RNNs *influence set* concept have been used in this work. After eliminating extreme outliers, we applied hybrid sampling technique, which is a combination of Random under-sampling and SMOTE on the training dataset. Experiments were carried out on insurance fraud dataset for four classifiers namely C4.5, Naïve Bayes, *k*-NN and Radial Basis Function networks. The results shown that the proposed approach is efficient on C4.5 classifier with improved fraud catching rate (TP) and non-fraud catching rate (TN) against simple hybrid sampling technique. The results also indicated that C4.5 classifier predicting the fraud class well at small values of SMOTE factor. Thus the *extreme outlier* elimination from minority samples before hybrid sampling, improved the fraud catching rate and non-fraud catching rate of a classifier. Though our approach is implemented for insurance domain, it can be applied to other domains for fraud detection

References

[1] T. Fawcett, and F. Provost, "Adaptive fraud detection,"*Data Mining and Knowledge Discovery,* Vol.1, I(3), 1997, pp.1-28.

[2] Clifton Phua , V. Lee, K. Smith, and R. Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research,*" Artificial Intelligence review*, 2005

[3] P.Chan, W. Fan, A. Prodromidis, and S.Stolfo, "Distributed Data Mining in Credit Card Fraud Detection," *IEEE Intelligent Systems* ,Vol. 14, 1999, pp. 67-74,

[4] Clifton Phua, A.Damminda, and V. Lee, "Minority Report in Fraud Detection: Classification of Skewed Data," *ACM Sigkdd Explorations: Special Issue on Imbalanced Data Sets*, **6**(1), 2004, pp. 50-59,

[5] S.J. Stolfo, D. W. Fan, W. Lee, and A.L Prodromidis, "Credit card fraud detection using meta-learning: Issues and initial results," AAAI *Workshop on AI Approaches to Fraud Detection and Risk Management ,AAAI Press, Menlo Park, CA* 1997, pp. 83-90.

[6] S. Stolfo, A.L. Prodromidis, Shelley Tselepis, Wenke Lee, and D.W. Fan, "JAM: Java agents for meta-learning over distributed databases," *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management,AAAI Press, Menlo Park, CA.*, 1997, pp. 91-98.

[7] S.J. Stolfo, Wei Fan, Wenke Lee, A.L. Prodromidis, and Phil Chan, "Cost-based modeling for fraud and intrusion detection: Results from the JAM Project," *In Proceedings of the DARPA Information Survivability Conference and Exposition 2,IEEE Computer Press. New York,* 1999, pp. 130-144.

[8] R.Wheeler, and S.Aitken, "Multiple algorithms for fraud detection. Knowledge-Based Systems," I3(2/3), 2000,pp. 93-99.

[9] M.Kubat, and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One Sided Selection," *In Proceedings of the Fourteenth International Conference on Machine Learning, Nashville, Tennesse. Morgan Kaufmann*, 1997, pp. 179-186.

[10] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research 16*, 2004, pp. 324-357.

[11] A. Estabrooks, T. Jo, and N. Japkowicz, "A Multiple Resampling Method for Learning from Imbalances Data Sets," *In Computational Intelligence*, Vol. 20, No. 1, 2004

[12] Soujanya Vadapalli, Satyanarayana R. Valluri, Kamalakar Karlapalem, "A Simple Yet Effective Data Clustering Algorithm," *Sixth International Conference on Data Mining (ICDM'06)*, 2006, pp. 1108-1112.

[13] Randall Wilson, and Tony, " Improved Heterogeneous Distance Functions," *Journal of Artificial Intelligence Research* 6,1997, pp. 1-34.

[14] C. Ling, and C. Li, "Data Mining for Direct Marketing Problems and Solutions," *In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), AAAI Press , New York,* 1998, pp. 73-79.

[15] R Brause, T.Langsdorf and M. Hepp, "Neural Data Mining for Credit Card Fraud Detection," *In Proceedings of 11th IEEE International Conference on Tools with Artificial Intelligence, Illinois, USA*, 1999, pp. 103-106.

[16] I. Witten, E.Frank, *Data Mining: Practical Machine Learning tools and Techniques with Java Implementations,* Morgan Kaufmann Publishers, 2000.

[17] G.Batista, M. Prati, and M.Monard, "A Study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations: Special Issue on Imbalanced Data Sets* 6(1), 2004, pp. 20–29.