Bank Fraud Detection Using Support Vector Machine

1.Nana Kwame Gyamfi
Computer Science Dept.
Kumasi Technical University
Kumasi, Ghana
nkgyamfi@st.ug.edu.gh

2.Dr Jamal-Deen Abdulai
Computer Dept.
University of Ghana
jabdul@ug.edu.gh

***Abstract*—With the significant development of communications and computing, bank fraud is growing in its forms and amounts. In this paper, we analyze the various forms of fraud to which are exposed banks d data mining tools allowing its early detection data already accumulated in a bank. We use supervised learning methods Support Vector Machines with Spark (SVM-S) to build models representing normal and abnormal customer behavior and then use it to evaluate validity of new transactions. The results obtained from databases of credit card transactions show that these techniques are effective in the fight against banking fraud in big data. Experiment result from the study show that SVM-S have better prediction performance than Back Propagation Networks (BPN). Besides the average prediction, accuracy reaches a maximum when training the data ratio arrives at 0.8.**

Keywords— Support Vector Machine; Bank fraud detection; Abnormal and Normal customer's behavior; Spark
Malware; Malware detectors; Mobile Phone; Signature based

I. INTRODUCTION

Financial service organizations have a number of strategic goals including the acquisition and retention of new and existing customers through the application of various management methodologies. In view of these goals, the institutions generate large volumes of profile data, purchase and browsing history and social media data daily. Banks as a financial institutions generate huge amount of data from different sources from its customers and this has contributed to the need for Big Data according to Wikipedia, (2016). Support Vector Machine can help to reduce risk and improve the quality of service extended to customer's in order to succeed in business. Fraud has become a very important risk to facing financial institutions, credit unions and banks in particular. Combating fraud come with traditional prevention techniques such as PINs, passwords and identification systems however have become inadequate in modern banking systems [Md Delwar, Karim Mohammed and Muham-mad Azzur, 2010]. Banks faced fraud in several activities but remote use of credit appears to be the most vulnerable.

Big data applications with data mining techniques can play a major role in the fight against these types of fraud. Data mining is a set of techniques for extracting important information from large amounts of data to assist in decision-making. The Spark is big data tool used particularly in this context, due to it numerous machine learning techniques and real time streaming. It is effective to be employed by financial institutions. Spark with SVM method is the best for these kinds of fraud. Several techniques have been proposed and used by many researchers, for fraud detection, including credit cards fraud. Among these data mining techniques are, Bayesian networks, Markov chains, neural networks, linear regression, sequence alignment etc.

The objective of this work is to provide fraud detection architecture that will enable bank to detect fraudulent transactions in real time with spark based on machine learning technique support vector machine. SVM which is very powerful for face recognition, fingerprints identification, voice recognition and similar task.

Our goal is to tackle the problem of fraud in banks and it resolution through Spark with SVM techniques. We present an analysis of the banks fraud problem and for each type design, the variant of SVMs that can be used for its solution and the necessary adaptations.

The rest of the paper is organized as follows: the kinds of fraud detection will be first presented as well as indices used to discover it, then we discuss the types of big data solutions that can used. In the third section we discuss the use of support vector machine in spark to meet the needs of detection of fraud. Further, on fourth section present the validation of proposed solutions by testing them on bank databases. We conclude the article by a conclusion and recommendation.

II. FORMS OF BANK FRAUD AND THEIR INDICES.

Fraud comes in many forms at the bank; it can be internal, that is committed by employees of the bank itself or external committed by clients, persons or bodies foreign to the bank. We are interested in considering external fraud in this paper.

A. Money laundering

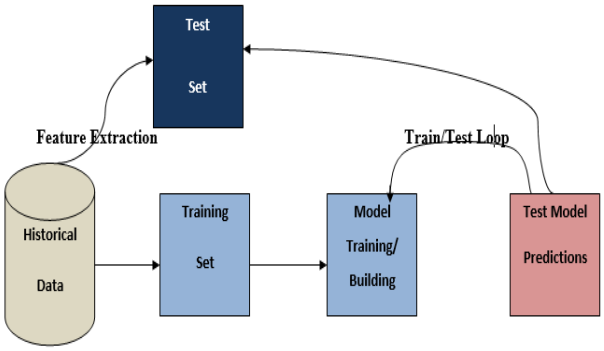
Money laundering is also a well-known form of fraud, international lute against this activity is conducted by different states to discover and prosecute criminal activities that occur. The fight against money laundering in the financial industry is based on the analysis and processing statements regarding suspicious transactions detected by financial institutions [FATF-GAFLORG, 2006]. Generally, only a few suspicious transactions are really money-laundering operations, but the number of operations to be analyzed by financial institutions require a long time. In the literature, artificial intelligence methods may use to improve the ability of financial

institutions to automatic processing of suspect data. However, the search for efficient methods for identifying suspicious transactional behaviors of money laundering remains a very active research field. Money laundering variables and indices are now easy to determine, because such unofficial activities are generated by complex social economic conditions. The following are few money bleaching indices used in the literature; the amount of transaction if it exceeds a predetermined amount by the bank, the transaction not justified, is then suspicious. For example, in Ghana the maximum amount to shop with your credit should not exceed 5000 euros. The following are further examine.

1. Sources of transfer
2. The date of transaction
3. The change of address
4. The time of transaction, transaction made at night with a large amount are suspected

B. Credit Cards Based Fraud

Credit card based fraud keeps increasing. Banks and financial companies lose sums of huge amount of money annually through fraud by credit card use. The detection of credit card fraud is often based on a number of forecast indicators that are generally concluded from transaction information retrieved from the historical database. We examine indices such as, frequent use of card, the remaining unpaid balance of each cycle, the maximum number of late days, shopping frequency, daily transaction, the largest number of frequency in historical database etc. These features are extracted for each transaction and are recorded for discovering patterns of fraudulent



transactions. Fraud detection model is shown in figure 1.

Figure1: Fraud detection Model

The proposed data model is built on the historical data already in the bank's warehouse, using the propose algorithm which consists of SVM-S to check if the outcome is fraudulent or not. A set of similar data will be used to predict the outcome for the effectiveness of the model. The model is used to evaluate a new transaction; transaction accepted by the model is executed then appended to the database to improve the model. Transactions rejected by the model are not executed but rather flagged as suspicious. If the transaction is normal, they are executed and added as already stated above.

III. BIG DATA FOR BANK FRAUD DETECTION

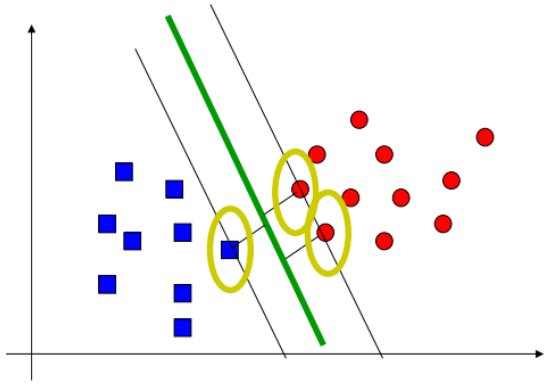
Due to rapid change and development of techniques used by fraudsters, data mining tools can no more analyze abnormal behaviors. Big data, in this context, come with machine learning techniques for fraud detection in database, which is best to fight against banking fraud.

In the literature, two forms of automatic learning are used: supervised and unsupervised. The supervised learning methods have been used for as association rules [DanielSa, Jose-Mar and LCcerda, 2009), Bayesian networks [Edgar, Freund and Girosi, 1997]. These methods assume a prior knowledge of the nature of transactions, fraudulent or genuine; the learning in this case consists of building a model separating the space into two parts according to the available examples then classifying new examples based on their membership to one of these two classes. Unsupervised method such as neural networks require no prior classification of training examples; it is rather based on the detection of strange transactions.

Big data applications, like such Spark, Hadoop, Cassandra etc. come with effective algorithms to manage structure, unstructured and semi-structured data.

A. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine-learning algorithm, which can be, used for both classification and regression challenges [Daniel et al. 2009, Bemhard et al. 2001]. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in an *n*-dimensional space (where *n* is number of features you have) with the value of each feature being the value of a particular coordinate. Figure 2, showing the hyper-plane differentiating



the two classes.

Figure 2: Hyper-plane

B. Linearly Separable Case

In the linearly separable case, there exist one or more hyperplanes that may separate the two classes represented by training data with 100%.

C. Non-linearly Separable Case

In the non-linearly separable case, it is not possible to find a linear hyperplane that separates all positives and negative examples [Rdigger, 1999]. To solve this case, the margin maximization technique may be relaxed by allowing some data point to fall on the wrong side of the margin, i.e. to allow a degree of error in the separation. Slack ε Variables are introduced to represent the error degree for each input data point.

Support vector machine would be used to classify the features of credit cards by each customer. The binary SVM solves the problem of separating two classes represented by n examples of m attributes each. Consider the following problem:

$$\{(x_1, y_1) \ldots\ldots\ldots (x_n, y_n)\}, x_i \in \mathcal{R}m, y_i \in \{-1, +1\}$$

Where x_i are learning examples and yi their respective classes. The objective of the SVM method is to find a linear function *f* (equation 1) called hyperplane, which allows separating the two classes:

$$F(K) = (x \cdot w) + b$$

(Eqn.1)

Where *x* is an example to classify, *w* is a vector and *b* is a bias. We must therefore find the widest margin between the two classes, which is equivalent to minimizing $1/2\, w^2$.

1. Kernel (trick)

The kernel trick allow the computation of the vector product $\phi(X_i)^T \phi(X_j)$ in the lower dimension input space.

From Mercer's theorem, there is a class of mapping Φ such that $\Phi(x)^T \Phi(y) = K(x, y)$ where *K* is a corresponding kernel function. Being able to compute the vector products in the lower dimension input space while solving the classification problem in the linearly separable feature space in a major advantage of SVMs using a find α that maximizes

$$\sum_i \alpha_i - 1/2 \sum_i \sum \alpha_i \alpha_j y_i\, y_j K(x_i, x_j)$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

Subject to $0 \leq \alpha_i \leq C,\ \forall i$ and the resulting SVM takes the form:

$$f(x) = w^T \Phi(x_i) + b = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b$$

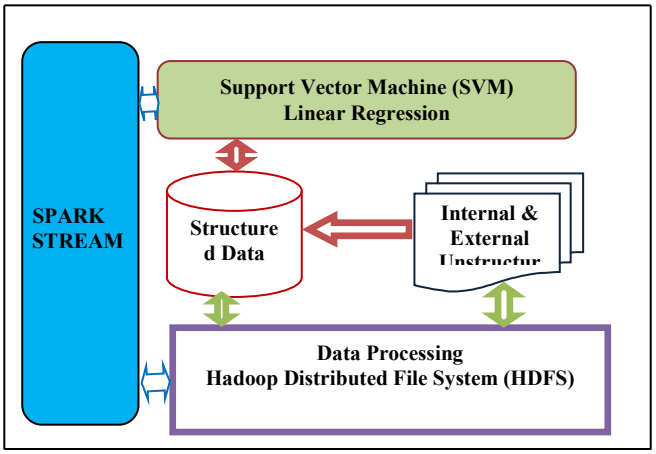


Figure 3: SVM-S Architecture for Big Data Fraud Detection

Figure 3, shows the SVM-S Architecture for the Fraud detection Framework. The spark application does not come with database; it always deserves on external database for it operations. The HDFS, which is popular when it comes to managing big data, is used as it database. With the machine learning techniques, we deploy the SVM to train the dataset for predictions.

2. Back Propagation Network: Using BPN to train data need to set some parameters. Among the most important parameters are the numbers of hidden layer N_i, hidden node N_j, and training epochs N_k, learning rate R_j, and momentum rate R_m. Further, setting of the parameter values remains as an art rather than a science. Complex problems can be incrementally better modelled by increasing hidden layers, but the improvement is generally accompanied by an associate cost in terms of training time and data overfitting.

3. Test and Result

a. Used Data: Due to confidential nature of data, it was very difficult to obtain real data that describe the behavior of bank customer. However, there are standard databases used in the literature to test fraud detection methods. To test our proposition of hybridization, we used the German and Australian databases of credit cards using.

Dataset for the implementation were called into memory for processing; the dataset was .csv file extension. These data created were used as the historical data for training, i.e. it is using supervised algorithm. The name of the file was called "creditcard.csv" for the first data set and "creditcard.csv1" for the second data set. Spark's primary abstraction is a distributed collection of items called a Resilient Distributed Dataset (RDD). In that case, the files were transformed to RDD for action and transformation.

The dataset was classified using the SVM and trained with the linear regression and logic regression for detecting anomaly using the credit card features. Combining these three methods in our model increase the accuracy of detections. The Table 1 shows the parameters set for the data set.

Data Set	**Classifier Model**	**Parameters**	**Definition**	**Source**
1st Data Set	SVM	C=10 RBF ε = 0.1		Djeffal et al. 2014
2nd Data Set	BPN	C=10 RBF ε = 0.1		Soltani et al. 2014

Table 1: Input parameters for the SVM

SVM-S model built for the two data sets used the same kernel function. There are two parameters associated with RBF kernel: *C* and *ε*. The regularization parameter *(C*) controls the trade-off between maximizing the margin and minimizing the training error term. Increasing the value improves the classification accuracy (reduce the regression error) for the data training. For an SVM-S, the value of ε in the ε-insensitive loss function should also be selected. The ε has an effect on the smoothness of the SVM's response and it affects the number of support vectors, so both the complexity and the generalization capability of the network depends on its value.

Samples		**# of Records**	
		Training Set	**Test Set**
Set -1F-To-1N (SVM-S)	Normal	100	86
	Fraud	25	20
Set -1F-To-1N (BPN)	Normal	150	125
	Fraud	25	12

Table 2: Training and Test Set Sizes for the Samples

Two different data sets were used to test the each algorithm model, with different conditions. The first data set has one normal transaction for each fraudulent one. Whiles the second set, has four normal for each fraudulent one. For each sample data, 70% of the data, both 70% of the normal transaction and 70% of the fraudulent, are taken as the training set for the model; while 30% of the data is taken as the training set to evaluate the performance of the model deployed. The training and test set size are given in Table 2, above.

b. Result: The performance of the SVM-S architecture is presented in Table 3, the left column shoes the method used to build the architecture models, the column named as "Train" indicates the prediction accuracy of the proposed architecture on the training data set of the given samples, the columns labelled "Test" shows the prediction accuracy of the architecture model on the testing data set of the given samples. The column named as "Build Time" show the time elapse for building the given architecture model over the given sample and columns "Frauds" show the number of fraudulent transaction in the data set assigned as fraud (True Positive) by the architecture over the given samples.

c. Comparison of Predicted Result between BPN and SVM-S

The table 3, below show the comparison of predicted result that exist between BPN and SVM-S, from the experiment.

Model/ Data Set	**Set-1F-To-1N**				**Set-1F-To-4N**			
	Train	**Test**	**Build Time**	**Fraud**	**Train**	**Test**	**Build Time**	**Fraud**
D1: SVM	98.78%	84.37%	<20m	20	96.34%	82.54%	<20	20
D2: BPN	99.86%	85.58%	<25m	13	97.346%	84.67%	<25	10

Table 3: Performance Accuracy over Training and Test Sets

From Table 3, it is clear that the SVM-S architecture, with SVM performance is very efficient. However, accuracy shows the rate of true assignment regardless of whether it a true fraud assignment or true normal assignment. Nevertheless, the number of fraudulent transactions assigned as fraudulent by the model, were accurate for the model. The fact that the expected outcome of the number of fraudulent transaction injected with each data set, SVM-S outcome was equal to task as compare with BPN outcome. This clearly shows that, SVM-S is more reliable and accurate than BPN. With time complexity, SVM-S use few time for predicting anomalies as compare with other algorithm such as BPN.

IV. CONCLUSION

We studied in this context, two cases of fraud in banks: credit card fraud and money laundering.

The performance of the proposed system was tested on the benchmarks General Ledger, Payables Data, created as similar to bank database. The precision obtained for the single class SVM method, was of about 80%, which represents a significant improvement in comparison to similar works reference. For the method, the slight improvement on credit scoring databases was because of the difficulty of obtaining real databases. The results can be improved by studying the influence of various parameters used by the SVM-S architecture.

REFERENCES

[1] Bernhard Scholkopf and Alexander J Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. The MIT Press, 2001.

[2] DanielSa nchez, A ila, Cerda,and osear aSerrano.Association rules applied to credit card fraud detection. Expert Systems with Applications, 36(2):3630–3640, 2009.

[3] FATF-GAFI.ORG. Financial action task force on money laundering. Rapport 1996-1997 sur les typologies du blanchiment de l'argent, Groupe d'Action Financire (GAFI), Fvrier 1997.

[4] Md Delwar Hussain Mahdi, Karim Mohammed Rezaul, and Muham- mad Azizur Rahman. Credit fraud detection in the banking sector in uk: a focus on ebusiness. In Digital Society, 2010. ICDS'10. FourthInternational Conference on, pages 232–237. IEEE, 2010.

[5] Rdiger W Brause, T Langsdorf, and Michael Hepp. Neural data mining for credit card fraud detection. In Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on, pages 103–106. IEEE, 1999.

[6] Suvasini Panigrahi, Amlan Kundu, Shamik Sural, and Arun K Majum- dar. Credit card fraud detection: A fusion approach using dempster– shafer theory and bayesian learning. Information Fusion, 10(4):354–363, 2009.

[7] Wen-Fang Yu and Na Wang. Research on credit card fraud detection model based on distance sum. In Artificial Intelligence, 2009. CAI'09. International Joint Conference on, pages 353–356. IEEE, 2009.

[9] Wikipedia.com "The Big Data" Retrieved on 26th May, 2016. From www.wikipedia.com/bigdata/