Credit Card Fraud Detection with Unsupervised Algorithms

Maria R. Lepoivre, Chloé O. Avanzini, Guillaume Bignon, Loïc Legendre, and Aristide K. Piwele
ECE Paris School of Engineering, Paris, France
Email: {maria.lepoivre, chloe.avanzini, guillaume.bignon, loic.legendre, aristide.piwele-koumba}@edu.ece.fr

***Abstract*—According to international credit card organisms such as VISA, there are more and more credit card frauds, both in quantity and in amount. To cure the problem, an anti-fraud project is developed using a combination of two unsupervised algorithms: Principal Component Analysis and SIMPLEKMEANS algorithm. To augment model accuracy, geographic positions of the transaction and of the client are added to traditional studied data, as everybody is fully connected with smartphones nowadays and as such tendency is growing up for a near future. Good results are obtained for proposed model on created test data base by achieving the foreseeing results and getting the classification of possible frauds.**

***Index Terms*—data mining, credit card, fraud detection, principal component analysis, SIMPLEKMEANS algorithm**

I. Introduction

Credit card payment is nowadays a very common process for most financial transactions. But in parallel, the number of fraudulent operations has been increasing [1] and is demanding active surveillance for reducing its impact on economy, the more as internationalization and extreme simplicity of transactions makes more difficult the application of different security norms. In France for instance, between Nov. 1st 2013 and April 30th 2014, 532.2 billion Euros transactions have been realized by 68.4 million cards in France, a total amount of card payments increased by 4.4% in comparison with 2012 [2]. Meanwhile, the total fraud amount reached 469.9 million Euros during the same period, which represents a 4.3% raise. For a long time researches have been developed in the domain to find solution to fraud problem [3]-[15]. Typically, fraudulent operations are representing a small fraction of all transactions, leading to skewed distributions, which are also noisy due to errors from collecting devices in data sets. Another difficulty stems from data overlapping when operations may look fraudulent when legitimate and vice versa. Obviously, fraudulent techniques are changing over time so detection system ought to be adaptive to maintain its efficiency.

For these various reasons it is difficult to design a very effective fraud filter, and usual approach is to take advantage of artificial learning systems for recognizing fraudulent features when facing them in real life after adequate training which mainly consists in optimizing a cost function measuring the distance of legitimate observed real data to fraudulent ones once a convenient metrics has been set. In [16], supervised probabilistic Bayes and Bayesian Networks algorithms have been used on the following variables: operation code/ response transaction code/ transaction date (YYYYMMDD)/ hour, minute, second, transaction amount and other Boolean values. Results are obtained with an error rate between 0.92% and 0.47%. Decision tree method and different Support Vector Machine (SVM) with polynomial and sigmoid functions have been compared in [17], with the conclusion that SVM generates over-fitting and is less efficient than decision trees. Artificial neural networks and Bayesian belief networks have been taken in [18, 19], and it has been observed that an error in selecting the set of detection variables could block the system due to imbalance between legal and fraudulent transactions. Most current approaches so far are depending on relatively heavy numerical treatment which makes improvement much heavier and obscures full understanding of their development.

A different approach is followed in present study where the intention is to reach more understandable results and at the same time simplify their getting. For that it is proposed to use two very well defined unsupervised algorithms, the Principal Component Analysis (PCA) and SIMPLEKMEANS (SKM) algorithm, both exhibiting full transparency in their operating process, and to discuss their reliability when applied to fraud detection problem.

II. Data Generation

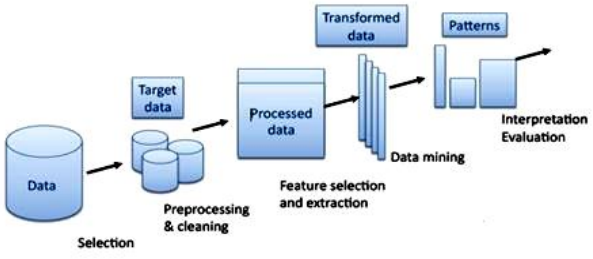

Figure 1. Schema of knowledge discovery from data process (Fayyad & Patetsky Shapiro & Smith, 1996).

Manuscript received July 30, 2015; revised September 8, 2015.

doi: 10.12720/jait.7.1.34-38

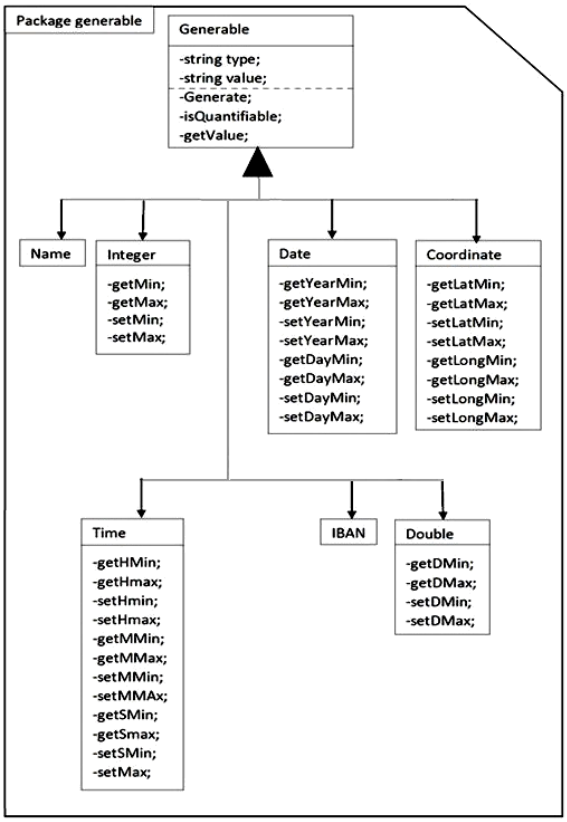

Figure 2. Model used in the program, showing the fields and associated methods.

Knowledge discovery process from data base requires five steps as showed in Fig. 1. First of all, an Extraction Transfer Loading is necessary. This stage consists in extracting data from different sources (data base, files, applications...), to be transformed and to be regrouped in a same data base. Afterwards, this one has to be cleaned, detecting and correcting corrupted or missing data. Subsequently and just to simplify the number of operations and get a faster system, only the most relevant attributes of data are going to be considered.In this way, the fulfilment of the attribute selection step is achieved. As third stage, data have to be transformed, building new attributes or changing their own format to obtain easier future manipulations. Previously selected data mining algorithms are thus able to be applied to the abovementioned data. To finalize the process, all the results obtained by the system have to be analyzed and interpreted [20].

In that way, to be able to test the efficiency of Credit Card Fraud Detection System, obviously data are required. Nevertheless, collection of real data from different banks is usually unsuccessful, because it is often related to sensitive financial transactions kept confidential for elementary privacy reasons. So randomized and forged data have to be generated for the purpose. These data are created so that data mining methods can be directly applied without having to clean and treat them. As it is necessary to deal with a large range of data types (coordinates, IBAN, dates, times…), they are generated regarding different algorithms such as a simple (just put a random value in each field) or a more complex one (fields linked to one another in order to simulate several transactions for a same person, for example). JAVA and JEE languages have been used to benefit from a web interface and easier implementation.

The generator allows the user create a structure with desired fields. As showed in Fig. 2, several fields are displayed from which the user can define different limits according to his needs. All choices can be modified by user request, except during data generation. After choosing the number of desired entities, the generator calls the Generable class which produces randomized data. Then a CSV file is created, containing all generated data according to user choices. This file can be used subsequently by the fraud detector program without any further needed modifications.

III. PCA AND SKM ALGORITHM

PCA is a powerful tool which allows us, with only some calculations, the obtaining a wide view of relationships among different credit card transaction characteristics. Its flexibility is demonstrated by the fact that it can be applied to very large data sets, independent of contents and size, an essential point for this problem. Afterwards, SKM algorithm will make an easier and faster identification of fraudulent or legal transactions. In other words, the following matrix is built:

$$T = \begin{bmatrix} t_{11} \cdots t_{1p} \\ \vdots \ddots \vdots \\ t_{n1} \cdots t_{np} \end{bmatrix} \quad (1)$$

T represents all the transactions of a bank account and each transaction $T_j = \{t_{j1}, t_{j2}, \ldots t_{jp}\}$ is described by p characteristics. T contains both legal TL and fraudulent TF transactions, T = {TL,TF} and the problem here is to exactly and only detect second ones by successive application of adapted filters $\{\ _k\}$. Best ones are such that with minimum number of most transparent operations (so the choice of best filtering set is made difficult by the interaction between operations belonging to different successive filters). A test of full filter efficiency is the distance $\Delta = |\Pi_k\{\Phi_k\}T - TF|$ measured with adapted metrics when tested on a representative base set $\mathcal{T}$ of possible transactions T. As indicated above other important elements in the choice of filtering set $\{\Phi_k\}$ are calculation simplicity and operation transparency. Here filters are Φ_1 = PCA and Φ_2 = SKM algorithms, and after their application two sets are obtained: QL (transactions classify as legal) and QF (transactions classify as fraudulent).

$$\Pi_k\{\Phi_k\}T = TF \quad (2)$$

$$T = \begin{pmatrix} T_1 \\ T_2 \\ \cdots \\ T_n \end{pmatrix} \text{ with } T_j = [\, t_{j1}, t_{j2}, ..., t_{jp}] \Rightarrow PCA,SKM\} \Rightarrow$$

$$\left\{ \begin{matrix} QL=\{Ti\ldots Tj\} \\ QF=\{Tk\ldots Tl\} \end{matrix} \right\} \quad (3)$$

where $i \neq j \neq k \neq l \in [1,n]$

A. PCA

PCA is a data analyzing method which transforms correlated variables into uncorrelated ones. In present case this method aims at representing transactions described by different attributes (transaction amount, date …) in a smaller subspace than initial one, and so that the least possible information is lost.

For each bank account, the matrix representing "*n*" transactions with their "*p*" respective attributes is built up. After centralized each value, variance-covariance matrix $\Sigma = N^{-1}(X^TX)$ is also constructed, representing the difference between the value and it respective estimation.

$$X_{(n,p)} = \begin{bmatrix} x_1^1 & \cdots & x_1^j & \cdots & x_1^p \\ x_2^1 & \cdots & x_2^j & \cdots & x_2^p \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_i^1 & \cdots & x_i^j & \cdots & x_i^p \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_n^1 & \cdots & x_n^j & \cdots & x_n^p \end{bmatrix} \quad (4)$$

The solutions of error minimization are the different eigenvectors of matrix Σ obtained by application of Gram-Schmidt method. After having deduced the respective eigenvalues, the dimension *d* of new space is chosen following cumulated variance percentage technique. The new space is built from the first *d* eigenvectors related to the *d* higher eigenvalues. The last step consists in projecting each transaction in this new space.

B. SKM Algorithm

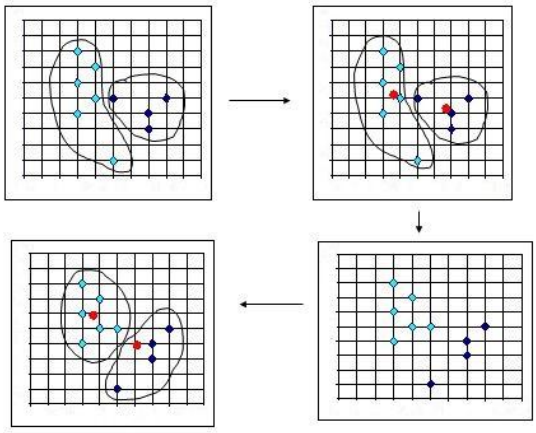

Figure 3. Application scheme of SIMPLEKMEANS algorithm.

After PCA, SIMPLEKMEANS unsupervised classification scheme [21] has been applied to classify the transactions. This algorithm consists in picking up randomly *k* initial points (cluster center), assigning then each point to the closest cluster, reevaluating the center of each cluster and reassigning points to their closest cluster, see Fig. 3. This cycle is repeated until the different sets become stable.

IV. Results

The model has been applied to manually implemented data containing on five bank accounts. The first one contains 8 transactions in which there are 2 fraudulent and 6 legal ones. In the second bank account, there are 2 legal transactions and 1 fraudulent one. The third bank account contains 3 legal transactions. The fourth bank account contains 20 transactions in which 15% of them are fraudulent and finally the last bank account contains 15 transactions with 33.33% of fraud.

TABLE I. Results for 5 Different Bank Accounts

	Bank Account n°1	Bank Account n°1	Bank Account n°2	Bank Account n°2	Bank Account n°3	Bank Account n°3
\Estimate Reality\	Fraud	No Fraud	Fraud	No Fraud	Fraud	No Fraud
Fraud	2	0	1	0	0	0
No Fraud	0	6	0	2	1	2

	Bank Account n°4	Bank Account n°4	Bank Account n°5	Bank Account n°5
\Estimate Reality\	Fraud	No Fraud	Fraud	No Fraud
Fraud	3	0	5	0
No Fraud	0	17	0	10

According to different tests, proposed present model gives good results. Transactions of bank accounts N°1, 2, 4 and 5 have been correctly classified, with 100% precision, see Table I (a diagonal matrix is obtained). An error has been detected in the third bank account where a legal transaction has been considered as fraudulent. This result could be explained by the fact that the number of clusters in SKM algorithm is fixed to 2 and that all transactions are forced to belong to one of these clusters even in indeterminate cases. The problem would also appear in the other extreme case of 100% fraudulent transactions. Nevertheless, even with first plain iteration $\Phi = \Phi_1\Phi_2$ of filters PCA and SKM, results from proposed present model are attractive. With basic undifferentiated Euclidian metrics distance, measurement error is $\Delta = 1/(707)^{1/2}$, a figure which can be significantly reduced by iterating Φ_2 several times without too much numerical involvement, see Fig. 4 which exhibits the remarkable reliability of proposed filter for fraud detection above some critical percentage. This suggests a reductive step

by step procedure to eliminate as much legal transactions as possible to end up within absolute reliability interval as it will be discussed elsewhere.

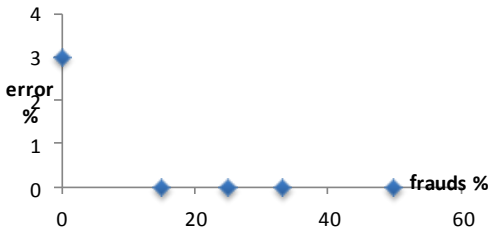

Figure 4. % Error vs % Frauds

V. Conclusion

The proposed model has been developed to satisfy the two conditions of calculation simplicity and operation transparency. An interesting trade-off, composed of two unsupervised algorithms (Principal Component Analysis and SIMPLEKMEANS algorithm) which considers geographic position of both transactions and clients, has various advantages. It directly classifies the transactions with a good precision and it can detect new fraudulent behaviors. Principal Component Analysis offers a complete view of relations among different attributes and at the same time, it is more flexible. Nevertheless, the risk remains to achieve a 'local' optimum instead of a general one. This risk could be reduced by repeating the "k means" process several times with different initial clusters to the expense of increasing the execution time.

Acknowledgment

The authors are very much indebted to ECE Paris School of Engineering for having provided the environment in which the study has been developed, to Pr H. Mechkour for guidance in the course of the research and Pr. M. Cotsaftis for help in preparing the manuscript.

References

[1] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical Science*, vol. 17, no. 3, pp. 235-255, 2002.
[2] C. Noyer, *The 2013 Annual Report of the Payment Card Security Observatory*, French Bank Governor and Pres. of the Payment Card Security Observatory, 2013.
[3] I. F. W. Silvaz, "Minería de Datos para la Predicción de Fraudes en Tarjetas de Crédito," *Vínculos. Colombia*, vol. 7, no. 2, pp. 58-69, 2010.
[4] U. Murad and G. Pinkas, "Unsupervised profiling for identifying superimposed fraud," *Principles of Data Mining and Knowledge Discovery, Lecture Notes in Artificial Intelligence*, vol. 1704, pp. 251-261, 1999.
[5] P. L. Brockett, R. Derrig, L. Golden, A. Levine, and M. Alpert, "Fraud classification using principal component analysis of RIDITS," *The Journal of Risk and Insurance*, vol. 69, no. 3, pp. 341-371, 2002.
[6] E. Duman and M. H. Ozcelik, "Detecting credit card fraud by genetic algorithm and scatter search," *Expert Systems with Applications*, vol. 38, pp. 13057-13063, 2011.
[7] P. K. Chan, S. J. Stolfo, D. W. Fan, W. Lee, and A. L. Prodromidis, "Credit card fraud detection using meta learning: Issues and initial results," *Working Notes of AAAI Workshop on AI Approaches to Fraud Detections and Risk Management*, 1997.
[8] P. K. Chan and S. J. Stolfo, "Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection," in *Proc. 4th Intern. Conf. on Knowledge Discovery and Data Mining*, 1998, pp. 164-168.
[9] T. Fawcett and F. Provost, "Adaptive fraud detection," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, 1997.
[10] R. Bhowmik, "Data mining techniques in fraud detection," *J. Digital Forensics, Security and Law*, vol. 3, no. 2, pp.35-54, 2008.
[11] R. Brause, T. Langsdorf, and M. Hepp, "Neural data mining for credit card fraud detection," in *Proc. 11th IEEE Intern. Conf. on Tools with Artificial Intelligence*, 1999.
[12] P. Chan, W. Fan, A. Prodromidis, and S. Stolfo, "Distributed data mining in credit card fraud detection," *IEEE J. Intelligent Systems*, vol. 14, pp. 67-74, 1999.
[13] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: Classification of skewed data," *SIGKDD Explorations*, vol. 6, no. 1, pp. 50-59, 2004.
[14] M. Syeda, Y. Zhang, and Y. Pan, "Parallel granular neural networks for fast credit card fraud detection," in *Proc. 2002 IEEE Intern. Conf. on Fuzzy Systems*, 2002.
[15] R. Wheeler and S. Aitken, "Multiple algorithms for fraud detection," *Knowledge-Based Systems*, vol. 13, no. 3, pp. 93-99, 2000.
[16] F. M. S. Soto, "Modelo de Aprendizaje Automático para la Detección de Fraudes Electrónicos en Transacciones Financieras," A. I Magister Scientiarum Thesis, Lisandro Alvarado University, Barquisimeto, Venezuela, 2011.
[17] Y. Sahin and E. Duman, "Detecting credit card fraud by decision trees and support vector machines," in *Proc. Intern. Multi-Conference of Engineers and Computer Scientists (IMECS)*, Hong Kong, 2011, vol. 1, pp. 442-447.
[18] S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick, "Credit card fraud detection using Bayesian and neural networks," in *Proc. 1st Intern. NAISO Congress on Neuro Fuzzy Technologies*, 2002.
[19] J. R. Dorronsoro, F. Ginel, C. Sanchez *et al*., "Neural fraud detection in credit card operations," *IEEE Trans. on Neural Networks*, vol. 8, no. 4, 1997.
[20] E. Barse, H. Kvarnstrom, and E. Jonsson, "Synthesizing test data for fraud detection systems," in *Proc. the 19th Annual Computer Security Applications Conference*, 2003, pp. 384-395.
[21] Provided by WEKA library (Waikato Environment for Knowledge Analysis).

Maria R. Lepoivre was born in Tenerife, Spain, in 1992. She obtained her Scientific High School diploma with highest honors in France in 2010. After her graduate certificate from Louis Le Grand and Claude Bernard Schools (Paris, France), she has joined ECE-Paris Engineering School (France). At the present time, she is in her last year of engineering with financial specialization in the ECE as well as Master 2 of Mathematic in Insurance, Economy and Finance (MASEF) in Paris Dauphine University (France).During the summer 2015, she achieved in Spain, the internship as assistant in production and budget following for the International Riu Group especially for worldwide hotel rebuilding.

Chloé Avanzini was born in Marseille, France, in 1992. She received her French Scientific High School diploma in France in 2010 with merit. After her graduate certificate from Dumont d'Urville School in Toulon, she has joined the ECE-Paris Engineering School. At the present time, she is in her last year of the ECE Engineering School, Paris, France with financial specialization as well as the Master ISIFAR (Statistic and Financial,

Insurance and Risk Computing Engineering) in the University Paris Diderot, France. In the summer 2015, she worked as an intern at Natixis as a financial controller and her current interest is the banking security.

Guillaume Bignon was born in Paris, France in 1994. He obtained his High School Diploma with honors in 2011, in Limoges, France. Then, he decided to study sciences in a two year intensive foundation degree located in Limoges, France. In 2013, he was accepted in ECE Paris Engineering School in which he obtained his bachelor degree. He started a Master in Computer Systems and Networks and expects to be graduated in 2016. He was the responsible of information systems in a student professional association called JEECE. He did also an internship in the company KelChauffeur as Web and Mobile Developer. He is actually studying Data Mining in Stockholm University.

Lo ï c Legendre was born in Ollioules, France in 1993. He earned his Scientific High School Diploma with distinction in 2011 in Toulon, France. After a two year intensive foundation degree, he has joined the ECE- Paris Engineering School where he studied Information Systems and he will be graduated in 2016. He has done a six weeks internship in POMONA Terre-Azur among the commercial team in 2014 and a fourth month internship in EcoCO2 among the web-developer team in order to develop a web site in 2015. He is actually carrying on his studies.

Aristide Kenneth Piwele was born in Cameroon in 1995 and aims to become a certified Data Analyst. He obtained his Scientific High School Diploma in 2011 in Cr é teil, France. After two years of preparatory classes for engineering school, he obtains the Bachelor's Degree in mathematics in the University Paris 6 in 2014. At the same time, he has joined the ECE- Paris Engineering School where he studies finances and mathematics. On the other hand, he completed in summer 2014 an internship at Bancel BTP, one of a major construction firm in France as a financial software developer during 4 months. In 2015, he worked as a Data Analyst junior/Statistician at Suez-Environnement, Paris La D é fense, an industrial services and solutions company specializing in securing and recovering resources.