2018 2nd Cyber Security in Networking Conference (CSNet)

Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for credit card fraud detection

Ibtissam Benchaji
IPSS. Faculty of Science
University Mohammed V
Rabat, Morocco
b.ibtissam@gmail.com

Samira Douzi
IPSS. Faculty of Science
University Mohammed V
Rabat, Morocco
samiradouzi8@gmail.com

Bouabid El Ouahidi
IPSS. Faculty of Science
University Mohammed V
Rabat, Morocco
bouahidi@yahoo.fr

Abstract—With the growing usage of credit card transactions, financial fraud crimes have also been drastically increased leading to the loss of huge amounts in the finance industry. Having an efficient fraud detection method has become a necessity for all banks in order to minimize such losses. In fact, credit card fraud detection system involves a major challenge: the credit card fraud data sets are highly imbalanced since the number of fraudulent transactions is much smaller than the legitimate ones. Thus, many of traditional classifiers often fail to detect minority class objects for these skewed data sets. This paper aims first: to enhance classified performance of the minority of credit card fraud instances in the imbalanced data set, for that we propose a sampling method based on the K-means clustering and the genetic algorithm. We used K-means algorithm to cluster and group the minority kind of sample, and in each cluster we use the genetic algorithm to gain the new samples and construct an accurate fraud detection classifier.

Keywords—Fraud Detection, Imbalanced dataset, K-means clustering, Genetic Programming, Autoencoder.

I. Introduction

Credit card fraud is a widely increasing problem in the credit card industry, particularly in the online sector. These illegal activities that aim to obtain goods without paying, or to gain illegitimate funds from an account, have caused severe damage to the users and the service provider.

Moreover, the main challenge of credit card fraud problem is to detect frauds in a huge dataset where the legal transactions are more and the fraudulent transactions are minimum or close to negligible. Hence, Developing Fraud Detection System (FDS) based on imbalanced datasets is one of the major challenges in machine learning.

A dataset is called imbalanced when the number of negative (majority) instances outnumbers the amount of positive (minority) class instances. Such imbalanced sets require additional precautions because the prediction accuracy of standard machine learning techniques, especially for the minority class which is the class of interest, tends to be lower. In real world applications, classification accuracy of the smaller class is critically important because the minority class usually is the class of great interest such as cases of fraud. Thus, misclassifying the minority class has much higher cost compared to misclassifying a majority class instance. Notice that, when the system predicts a transaction as fraudulent when in fact it is not (false positive), the financial institution has an administrative cost, as well as a decrease in customer satisfaction. On the contrary, when the system does not detect a fraudulent transaction (false negative), the amount of that transaction is lost.

In this work, we propose the generation method of imbalanced data set's minority class, by using K-means clustering and genetic algorithm, new samples can be obtained through crossover on the basis of cluster as the complement of the minority class samples.

The remainder of this paper is organized as follows. Section II introduces the imbalanced data sets challenge faced by detection systems. Section III presents some required concepts and gives all the details of our proposed approach. Finally, Section IV presents conclusions and suggestions for future work.

II. Classification Of Imbalanced Datasets

First, As introduced in Section I, the major challenge to be addressed when designing a FDS is handling the class imbalance, since legitimate transactions far outnumber the fraudulent ones. In fact, Class distribution is extremely unbalanced in credit card transactions, since frauds are typically less than 1% of the overall transactions, as shown in [1].

In this regard, various approaches have been proposed to deal with imbalanced datasets issue and improve the performance of predictive modeling. These approaches could be mainly divided into two categories. The first category of methods solve the problem at data level, that is, data resample techniques are performed to directly alter the original dataset which is not balanced by removing samples from the majority class (undersampling) or replicating training samples of the minority class (oversampling)[2]. Advanced oversampling methods like SMOTE [3] generate synthetic training instances from the minority class by interpolation, instead of sample replication. However, these techniques present significant drawbacks, such as undersampling may lose some potential information, and oversampling may lead the overfitting.

The second main type of approaches that considers classification of imbalanced data sets at algorithm level is known as ensemble learning. Ensemble methods include bagging and boosting strategies that aim to combine multiple classifiers in order to reduce the data variance. The ideas of Bagging (Bootstrap Aggregating) [4] is to split the majority class into multiple sub sets, and for each sub set, training a basic weak classifier together with the minority class, then integrate these classifiers into a strong classifier. Bagging is proposed to use random sampling technique to generate empirical distribution so as to approximate the real distribution of the data set. The AdaBoost (Adaptive Boosting)[5] talks about the misclassified samples that generated by the former basic classifier will be augmented by assigning weights. The weighted data sets will be send to next basic classifier. AdaBoost (Adaptive Boosting) focus on the misclassified samples that generated by the former

978-1-5386-7045-3/18/$31.00 ©2018 IEEE

basic classifier to be augmented by assigning weights. The weighted data sets will be send to next basic classifier to reduce the total bias error. The weighting strategy of AdaBoost is equivalent to resampling the data space [6], which are applicable to most classification systems without changing their learning methods. Besides, it could eliminate the extra learning cost for exploring the optimal class distribution and representative samples [7]. Moreover, compared with the method of eliminating samples from data set, it reduces the information loss, overfitting risk and bias error of a certain classification learning method [8].

III. Proposed Methodology

In an effort to improve the performance of imbalanced data classification when designing Fraud Detection System, a new oversampling strategy using K-means cluster and genetic algorithms to create synthetic instances from the minority class is proposed in this paper. It is designed to overcome the limitations of existing sampling methods such as information loss or overfitting. Figure 1 shows the big picture of the proposed method. Our suggested approach generates new minority class instances in each cluster and merges them with the original dataset to gain new training sets in order to construct an accurate fraud detection classifier.

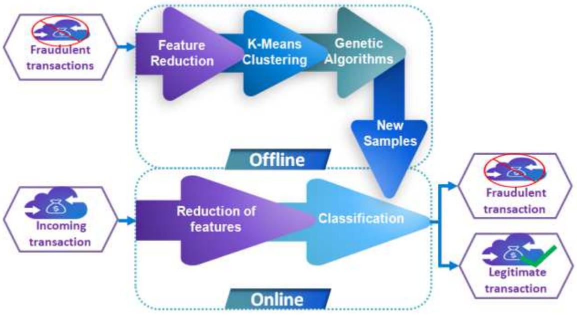

Fig. 1. The illustration of credit card fraud detection system

A. *Financial fraud data and Feature selection by using Autoencoder*

Creating domain expertise features has proven to be an integral part of building predictive models in detecting card fraud; banks collect a large amount of information about card accounts, account holders, and transactions. However, not all of this data directly specifies important predictors such as consumer spending habits throughout time, or variance in predictors, etc.

The basis of credit card fraud detection lies in the analysis of cardholder's spending behavior. This spending profile is analyzed using optimal selection of variables that capture the unique behavior of a credit card and detect very dissimilar transactions within the purchases of a customer. Also, since the profile of both a legitimate and fraudulent transaction tends to be constantly changing, optimal selection of variables that greatly differentiates both profiles is needed to achieve efficient classification of credit card transaction.

Therefore Feature selection is a fundamental preprocessing step in fraud detection systems, to select the optimal subset of relevant features by removing redundant, noisy, and irrelevant features from the original dataset, and decrease the computational cost without a negative effect on the classification accuracy. Thus, Autoencoder has been seen as one of the best technique used for feature selection in many applications such as image analysis. For example, it has been utilized in medical imaging to reconstruct training samples in order to create compressed sensing images from MRI and CT [9].

Autoencoder can be regarded as a special form of neural network designed for unsupervised learning [10, 11]. The aim of an Autoencoder is to transform inputs into outputs with the least possible amount of deviation [12].

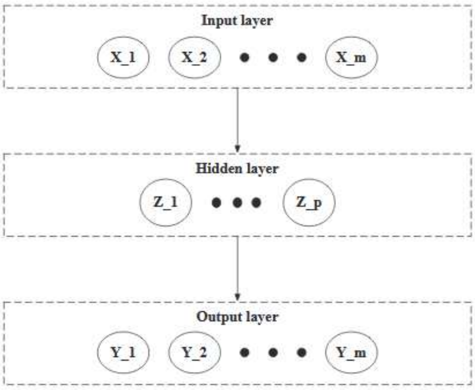

Fig. 2. The Autoencoder architecture

The basic version of an Autoencoder consists of three fully connected layers, i.e., one input layer, one hidden layer and one output layer. The reconstruction ability can be improved by introducing more hidden layers and hidden neurons. A shown in Fig. 2, the input (i.e., denoted as X) and output (i.e., denoted as Y) are set identical with a dimension of n and m, where n is the number of observations and m is the variable number. An Autoencoder consists of an encoder and a decoder. The encoder transforms the input data into high-level features (i.e., denoted as Z), while the decoder tries to reconstruct the input data using high-level features.

The Autoencoder is applied on the set of features that contains credit card transactions to generate the robust and discriminative features for fraudulent instances.

B. *K-means Clustering Method*

Clustering is a process of partitioning a set of samples into a number of groups without any supervised training. The goal of this technique is to categorize the records of a dataset in such a way that similar records are grouped together in a cluster and dissimilar records are placed in different clusters. The more the similarity among the data in clusters, more the chances of particular data-items to belong to particular group. Thus, K-Means is a heuristic clustering algorithm based on distance measure that partitions a data set into K clusters by minimizing the sum of squared distance in each cluster.

The algorithm consists of three main steps: i) initialization by setting center points (or initial centroids) with a given K, ii) Dividing all data points into K clusters based on K current centroids, and iii) updating K centroids based on newly formed clusters. K-Means algorithm converges after several iterations of repeating steps ii) and iii). These steps are shown as Fig. 3.

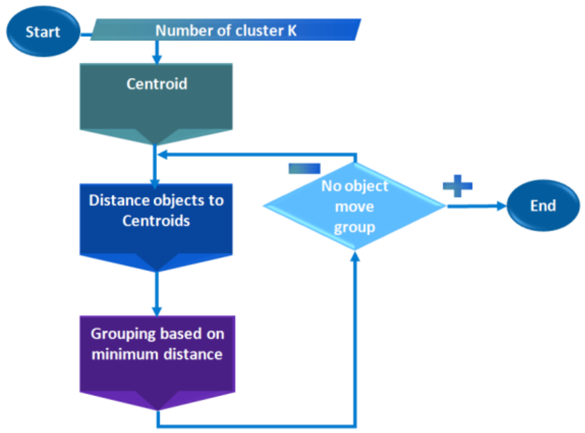
Start
Number of cluster K
Centroid
Distance objects to Centroids
No object move group
End
Grouping based on minimum distance

Fig. 3. K-Means Clustering Flow Chart

1) The K-Means Clustering algorithm is specified as follows:

- Let X = {x1, x2, x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.
- Step 1: Select ‘c’ cluster centers randomly.
- Step 2: Calculate the distance between each data point and cluster centers using the Euclidean distance metric as follows

$$Dist_{XY} = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2}$$

- Step 3: Data point is assigned to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
- Step 4: New cluster center is calculated using:

$$V_i = \left(\frac{1}{Ci}\right)\sum_{1}^{Ci} xi$$

Where, ‘ci’ denotes the number of data points in this cluster.

- Step 5: The distance between each data point and new obtained cluster centers is recalculated.
- Step 6: If no data point was reassigned then stop, otherwise repeat steps from 3 to 5.

In the proposed paper, the k-means algorithm is used to split minority class of fraud instances into clusters according to their similarities and generate new samples in these clusters. Thus, new samples are created from samples that are as much similar as possible. Thus, every cluster can obtain certain proportion of new samples, which can guarantee that new samples of whole minority class have better coverage and representation. Fig. 4 presents the schematic diagram of this step.

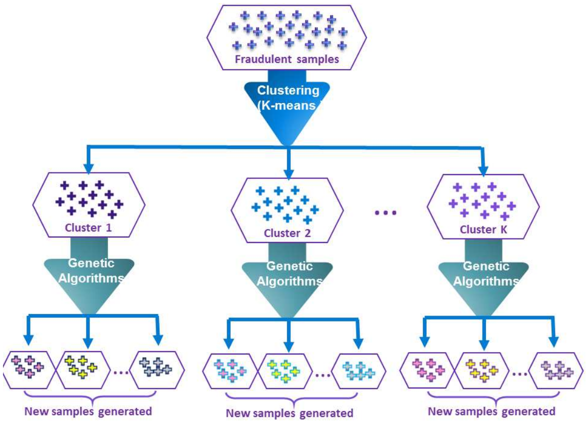

Fig. 4. GA generation process

C. Genetic Algorithms

Genetic algorithms were inspired from Darwin’s theory of evolution and were pioneered by John Holland [13]. A genetic algorithm can be defined as a search algorithm based on the mechanics of natural selection and natural genetics [14]. Genetic algorithms have at least the following elements in common: Populations of chromosomes, selection according to fitness, crossover to produce offspring, and random mutation of a new offspring [15]. In a broader usage of the term, a genetic algorithm is any population-based model that uses selection and recombination operators to generate new sample points in a search space. Fig 5 depicts a schematic view of different steps involved.

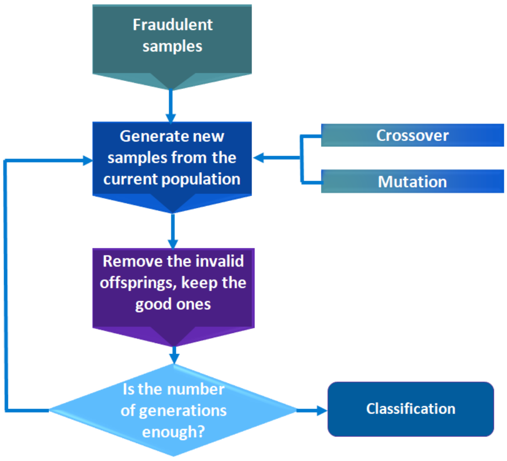
Fraudulent samples
Generate new samples from the current population
Crossover
Mutation
Remove the invalid offsprings, keep the good ones
Is the number of generations enough?
Classification

Fig. 5. GA Flow

In further details, the algorithm starts with a population of fraudulent transaction samples. Each sample within the population is coded in a so-called chromosome into a specific type of representation (i.e. binary, decimal, float, etc).

Each chromosome (sequences of genes) is assigned a “fitness” according to how valid new individuals’ property features are, based on a given fitness function. Fitness is calculated in the evaluation step.

While the termination condition of number of generations is not met, the processes of selection, recombination, mutations and fitness calculations are done. Selection process chooses individuals from population for the process of crossover. Recombination (or crossover) is done by exchanging a part (or some parts) between the chosen individuals, which is dependent on the type of crossover (Single point, Two points, Uniform, etc). Mutation is done after that by replacing few points among randomly chosen individuals. Then fitness has to be recalculated to be the basis for the next cycle.

1) Pseudo code of genetic algorithm

```
• Initialize the population
• Evaluate initial population
• Repeat
      Perform competitive selection
      Apply genetic operators to generate new solutions
      Evaluate solutions in the population
• Until some convergence criteria is satisfied
```

2) Selection process

The selection process is used for choosing the best individuals from the population which lead to the further breeding in the next generation. The selection operation takes the current population and produces a ‘mating pool’ which contains the individuals which are going to reproduce. The higher of the fitness value, the higher the probability to be selected and copied many times. There are three common types of selection operator, Fitness proportional / Roulette Wheel Selection [16], Stochastic Universal Sampling [17] and Binary Tournament Selection Tournament [16].

3) Crossover process

Crossover is also a genetic operator which is succeeded by the selection. It takes two individuals and cuts their chromosome strings at some chosen position to produce two “head” segments and two “tail” segments. The tail segments are then swapped over to produce two new full length chromosomes as depicted in Fig 6. Each of the two offspring inherits some genes from each parent. Crossover is made with the hope that new chromosomes will contain good parts of old chromosomes. As a result, the new chromosomes are expected to be better. If crossover is performed, the genes between the parents are swapped, and the offspring is made from parts of both parents' chromosomes. If no crossover is performed, the offspring is an exact copy of its parents.

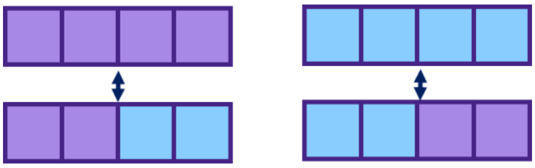

Fig. 6. Crossover

4) Mutation process

The last process in genetic algorithm is the mutation process. It is mainly used to maintain heterogeneity in the population. It alters some of the genes to create a unique and more fit offsprings in the population. Mutation is applied to each child individually after the crossover that alters each gene with a low probability, typically in the range 0.001 and 0.01, and modifies elements in the chromosomes [16]. Mutation is often seen as providing a guarantee that the probability of searching any given string will never be zero, it acts as a safety net to recover the good genetic material that may be lost through the action of selection and crossover. Mutation prevents the GA from falling into local extremes and provides a small amount of random search that helps ensure that no point in the search space has a zero probability of being examined. If mutation is performed, one or more parts of a chromosome are changed, and if there is no mutation, the offspring is generated immediately after the crossover (or directly copied) without any change [16].

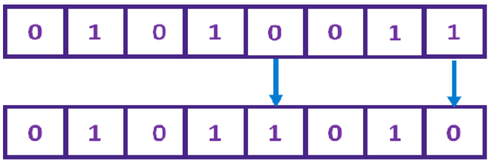

Fig. 7. Mutation

In the proposed paper, a number of genetic operations are applied on group to code the samples of each cluster into chromosome, then crossover operator is used to get a new generation group. New individuals will inherit the individual’s characteristic of father’s generation, but not simple duplication. It aims to get a new balanced training set that allow classifiers to be more accurate in detecting fraudulent transactions.

IV. Conclusion

In this study, a new method for data generation of imbalanced data set's minority class was proposed to enhance fraud detection in e-banking by using K-Means clustering and genetic algorithm as an oversampling strategy.

Although genetic algorithms have been applied in many areas, our application domain aims to handle imbalanced data set issue by generating new minority class instances to gain new training sets. Applying this algorithm into bank credit card fraud detection system aims to reduce fraudulent transaction and decrease the number of false alert. A further work is to implement this approach using python programming language, this will allow us to validate our work and produce pertinent experimental results.

References

[1] Using HDDT to avoid instances propagation in unbalanced and evolving data streams. A. Dal Pozzolo, R. A. Johnson, O. Caelen, S.Waterschoot, N. V. Chawla, and G. Bontempi. s.l. : Proc. Int. Joint Conf. Neural Netw., 2014, pp. 588–594.

[2] When is undersampling effective in unbalanced classification tasks?. A. Dal Pozzolo, O. Caelen, and G. Bontempi. s.l. : Machine Learning and Knowledge Discovery in Databases. Cambridge, U.K.: Springer, 2015.

[3] SMOTE: Synthetic minority over-sampling technique. N. Chawla, K. Bowyer, L. O. Hall, and W. P. Kegelmeyer. s.l. : . Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[4] Bagging Predictors. Breiman, Leo. (1996). s.l. : Machine Learning. 24 (2), 123–140.

[5] Experiments with a New Boosting Algorithm. Machine Learning. Freund, Yoav, and Robert E Schapire. (1996). s.l. : Machine Learning: Proceedings of the Thirteenth International Conference, 148–156.

[6] Cost-sensitive boosting for classification of imbalanced data. Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y. (2007). s.l. : Pattern Recognition, 40 (12), 3358-3378.

[7] Classification with class imbalance problem: A Review. Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). s.l. : Int. J. Advance Soft Compu. Appl, 7 (3).

[8] Learning from imbalanced data. Knowledge and Data Engineering. He, H., and Garcia, E. A. (2009). s.l. : IEEE Transactions on, 21 (9), 1263-1284.

[9] RODEO: Robust de-Aliasing autoencoder for Real-Time Medical Image Reconstruction. Mehta, Janki, and Majumdar, Angshu. 2017. s.l. : Pattern Recognition . 63. pp. 449- 510.

[10] Representation Learning via Semi-supervised Autoencoder for Multi-task Learning. al., Fuzhen Zhuang at. s.l. : EEE International Conference on Data Mining, 2015.

[11] Unsupervised Feature Extraction with Autoencoder Trees. Ozan úIrsoy, Ethem Alpaydõn. s.l. : Neurocomputing (2017), doi:10.1016/j.neucom.2017.02.075.

[12] *Advanced Phishing Filter Using Autoencoder and Denoising Autoencoder.* Samira Douzi, Meryem Amar, Bouabid El ouahidi. s.l. : Proceedings of the International Conference on Big Data and Internet of Thing Pages 125-129.

[13] Adaptation In Natural And Artificial Systems. Ozan úIrsoy, J. Holland. s.l. : University of Michigan Press, 1975.

[14] Computer-Aided Gas Pipeline Operation Using Genetic Algorithms And Rule Learning, Ph.D. thesis. D. Goldberg. s.l. : University of Michigan, Ann Arbor, 1983.

[15] Intelligent exploration for genetic algorithms: Using self-organizing maps in evolutionary computation. H. Ben Amor, and A. Rettinger. s.l. : In proceedings of the 2005 conference on Genetic and evolutionary computation, 2005, pp. 1531–1538.

[16] Genetic Algorithms in Search, Optimization and Machine Learning. Goldberg, D. E. (1989). s.l. : Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition.

[17] Reducing bias and inefficiency in the selection algorithm. Baker, J. E. (1987). s.l. : In Proc of the Second International Conference of Genetic Algorithms and their application, pages 14-21.

978-1-5386-7045-3/18/$31.00 ©2018 IEEE