INTELLIGENT AUTOMATION & SOFT COMPUTING, 2017
https://doi.org/10.1080/10798587.2017.1321228

A Novel Strategy for Mining Highly Imbalanced Data in Credit Card Transactions

Masoumeh Zareapoor and Jie Yang

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT
The design of an efficient credit card fraud detection technique is, however, particularly challenging, due to the most striking characteristics which are; imbalancedness and non-stationary environment of the data. These issues in credit card datasets limit the machine learning algorithm to show a good performance in detecting the frauds. The research in the area of credit card fraud detection focused on detection the fraudulent transaction by analysis of normality and abnormality concepts. Balancing strategy which is designed in this paper can facilitate classification and retrieval problems in this domain. In this paper, we consider the classification problem in supervised learning scenario by creating a contrast vector for each customer based on its historical behaviors. The performance evaluation of proposed model is made possible by a real credit card data-set provided by FICO, and it is found that the proposed model has significant performance than other state-of-the-art classifiers.

KEYWORDS
Imbalanced data; frequent itemset mining; non-stationary data; credit card; fraud detection

1. Introduction

Imbalanced data is attempted in situations where the instances from one class outstrip the instances from the other class. An extreme case of this scenario occurs in the financial data-set, where the number of fraudulent transactions is extremely far from legitimate transactions. However, learning from highly imbalanced data is particularly difficult due to the absence or shortage of one of the class labels, which have an important role to find relevant information (Han, Lei, Zhao, & Yang, 2012; He & Garcia, 2009). The class imbalance problem has received more and more emphasis in recent years, and it occurs in many ranges of domain applications like; medical diagnosis, anomaly detection and credit card fraud detection (Ali, Shamsuddin, & Ralescu, 2015). In this paper, our work centralized on the imbalancedness problem in a credit card data-set, and we designed a framework for the credit card fraud detection technique based on the balancing strategy. In the context of the machine learning technique, classification is supervised learning while, clustering is unsupervised learning. As in this paper we are using supervised learning (where the class label is available), so, classification is the best solution. As billions of dollars of loss are caused every year due to fraudulent credit card transactions, credit card fraud becomes one of the dangerous frauds (Van-Vlasselaer et al., 2015). The cost of a fraud is often equal to the transaction amount; however, the small and big amounts must be treated with equal importance (Wong, Ray, Stephens, & Lewis, 2012). The existing fraud detection techniques address this issue by using sampling techniques, and these approaches are either too expensive or difficult to obtain (Dal-Pozzolo, Caelen, & Bontempi, 2015). Some other approaches neglect this issue and they just detect the frauds by analyzing the normal and abnormal behavior of genuine users, and thus cannot produce an optimal method (Yang & King, 2009). Moreover, various data mining and machine learning classifiers demonstrate poor performance evaluation when directly applied to credit card fraud detection. In such problems when using machine learning algorithms, practically all the instances are labeled as one class, which is the larger class (majority), while a few instances are labeled as the other class (minority). This is due to the fact that the classifiers will tend to predict based on the entire data-set, thus, categorizing all data as belonging to the larger class (Dong, Chung, & Wang, 2016; Japkowicz & Stephen, 2002; Weston, Hand, Adams, Whitrow, & Juszczak, 2008). Sampling techniques are the current outstanding techniques to address this issue. The typical sampling techniques include oversampling (Nekooeimehr & Lai-Yuen, 2016), under sampling (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Phua, Alahakoon, & Lee, 2004) and ensemble techniques (Sun et al., 2015). Resampling techniques can produce the noise by increasing the number of minority instances, while under sampling techniques, attempt to balance the data-set by removing some instances from the majority class, and sometimes the removed instances may contain important information and ignoring that information may be causes wrong in total results. Ironically, the minority class, which is the smaller class, is more interesting and shows more importance; therefore it must be recognized in the evaluation. Another issue in learning from credit card transactions is that the transactions change and evolve over time, for instance, costumer's behavior may change in different seasons or holidays. Even the fraudsters change their behavior or methods whenever planning to do new frauds. In fact, we call credit card data-sets as a non-stationary data, because of these certain sophisticated characteristics:

1. fraudulent behavior is very dynamic, may appear similar to genuine customer behavior and
2. fraudulent behavior is hidden in diversified customer behavior.

Moreover, the most striking characteristic of the credit card data-set is that it is highly imbalanced; (about 2–3% of

CONTACT Jie Yang ✉ jieyang@sjtu.edu.cn

the transactions are fraudulent), which represents the real world scenario (Japkowicz & Stephen, 2002). However, it is quite important to develop an intelligent model to overcome these important issues. Imbalancedness is not the only issue that decreases the classification performance, but another noteworthy problem with such data-sets is the amount of overlapping classes, which occurs due to limited information about transaction records. The overlapping can create misclassification problems and result in customer dissatisfaction as well as money loss. To this end, we introduce a "formalization of the learning problem" and present a novel way to create new instances in the data-set. As the focus of this work is on imbalancedness and the non-stationary environment of the data-set, we show the impact of the balancing the data-set before performing the classifiers on the final performance. To end this, we propose a "stepwise classification technique," which is based on frequent itemset mining, called (SWC-FIM). The proposed model, solve the imbalancedness problem by creating a contrast vector for each customer based on its historical behavior. Frequent itemset mining is the first step in association rule learning, to discover the repetitive information within a "multitude of data". With this merge technique, we can provide a suitable framework for credit card fraud detection that is able to handle class imbalance, overlapping and non-stationary environment of the data-set.

Taking the area of interest of this paper, credit card fraud as an example, the minority class would be considered "fraudulent transaction" while the majority class would be considered "non-fraudulent".

2. State of Art in Credit Card Fraud Detection

Credit card fraud detection is one of the most interesting domains of fraud detection (West & Bhattacharya, 2016; Yang & King, 2009). Learning from credit card transaction datasets is a challenging issue, because of highly class imbalance, non-stationary transactions (due to dynamic fraud behavior and diverse genuine behavior patterns) (Dal-Pozzolo, Caelen, Le Borgne, Waterschoot, & Bontempi, 2014). The credit transactions namely as non-stationary data since both legitimate and frauds transactions change over time due to instability behavior of customers and fraudsters. The most proposed techniques to detect the frauds rely on the "automatic analysis of transactions" (Dal-Pozzolo et al., 2014; Ali et al., 2015; West & Bhattacharya, 2016; Liu, Wu, & Zhou, 2009; Zhang, Krawczyk, Garcìa, Rosales-Pérez, & Herrera, 2016; Van-Vlasselaer et al., 2015). The existing techniques are based either on "supervised" or "unsupervised" techniques with regards to availability of the label class. Supervised methods make use of the labels of past transactions, which are available in training part, while unsupervised (Quah & Sriganesh, 2008; Weston et al., 2008) methods (where the label of transactions are not available) using clustering algorithm to group customers into different profiles and any outlier from customer profiles identify as fraudulent transactions (the recent survey by (Dong et al., 2016; Nian, Zhang, Tayal, Coleman, & Li, 2016). Researchers in the machine learning algorithms addressed the problem of class imbalance with various approaches including; different forms of re-sampling technique such as over-sampling the minority classes (OS), under-sampling the majority classes (US) ensemble techniques, and even lots of research has been done in comparing the various sampling techniques (Dal-Pozzolo et al., 2014; Sundarkumar & Ravi, 2015). Recently, some studies have generally used a combination of several machine learning based classification algorithms, including decision trees, and support vector machines, Bayesian networks to detect fraud detection by ignoring the most important characteristic of data, which is imbalancedness & non-stationary problems (Ali et al., 2015; Liu et al., 2009). Most often, the performance of aforementioned methods are less than ideal or may become completely intractable to achieve the adequate results. In consequence, how to deal with the imbalance data is still an emerging research filed due to the weak performance of standard classifiers whose algorithms are developed for the balanced data-set (Sun et al., 2015). In what follows we list a summary of existing methods with a specific focus on imbalancedness and non-stationary problems. Some studies in credit card fraud detection have been done on the combination of under-sampling of majority classes with over-sampling by increasing of minority class examples, while, they haven't been able to get significant improvement in their performance measures (Chawla et al., 2002; Dal-Pozzolo et al., 2015; Liu et al., 2009; Sundarkumar & Ravi, 2015). Zhang et al. in (2016) proposed a new ensemble method to handle the classification in the imbalanced data-set. This approach combines the bagging and boosting techniques together, in order to balance the data-set, where, the bagging technique reduce the variance for the classification model through resampling the original data-set, while the boosting technique can reduce the bias of the model. He and Garcia (2009) and Chen and He (2011) proposed REA model, where they suggest propagating examples of the minority class and then recommending a k-nearest neighbors algorithm as a classifier. Gama et al. developed a model to handle the non-stationary environment of the data-set, which is known as the concepts drift (Gama, Žliobaitė, Bifet, Pechenizkiy, & Bouchachia, 2014). Concept drift is designed to train the classifiers on the recently supervised samples, while the obsolete ones are discarded ("such as DWM (Kolter & Maloof, 2007)). Recently, some researchers have developed a 'new breed of classification technique' for handling the imbalancedness problems, which is called, 'hybridization techniques' (Ali et al., 2015). This technique is designed with more than one machine learning classification to mitigate the imbalanedness problem in real world data-sets such as the credit card transaction data-set. Besides that, most hybrid methods in class imbalance classifications focus more on neural networks, SVM and decision tree, only a few kinds of literature from other techniques are devoted to highly imbalanced data-sets. Dal-Pozzolo et al. (2014, 2015) designed a new fraud detection technique by focusing on two crucial issues; imbalancedness, and non-stationary. The analysis of the proposed technique is performed on a real credit card data-set. They proved that by solving those striking characteristics, we can obtain an optimal technique. They used three approaches (static, update and forgetting) to learn from high imbalance and non-stationary credit card data-set. In this paper, we focused on the supervised technique and we attempt to balance the data-set before any algorithm is applied. However, we didn't consider any sampling techniques to solve the challenge of imbalanced class, as might not be suitable in the highly imbalanced data-set, which is pointed by many researchers (Ali et al., 2015).

3. Learning Strategy

We formulate the credit card fraud detection technique as a binary classification where the detection methods classify each transaction as legitimate or fraudulent. Each transaction is represented by a feature vector f and a label Z. Features in f, are the information for each transaction (such as credit card number, transaction amount, transaction date, time of the transaction, etc.). Label in Z could be, fraudulent "1" and legitimate "0". The proposed model "K" typically update after every transaction "t", due to nature of the model (as our model train on pattern database). If the proposed model "K":

K_t: $\mathbb{R}^n$ $\rightarrow \{0, 1\}$, so, each feature vector $f \in$ $\mathbb{R}^n$, and label $K_t(f) \in 0.1$

"1" denotes a fraud transaction (minority class) and "0" a legal transaction (majority class). If (X) denoted as a particular customer and X_{ij} be the transaction number j of a card number i, so the transactions are ordered in time such that:

$$(If\ X_{im}\ occurs\ before\ X_{in} | then\ "m < n")$$

Let $\{T_t\}$ be new transactions, so, each transaction T_j is assigned a binary status Z, the goal of a detection system is to learn P (Z|X) and predict the class of a new transaction Z∊(0, 1). In this paper, we consider the credit card transaction data as a non-stationary data, because the customers and fraudsters are having dynamic behaviors due to different reasons; for instance, the customer's profiles are found to be changing gradually over an unreasonable period of time (different seasons, different places & times), and also the fraudsters try to mimic & follow the genuine customer behaviors to easily catch the frauds, therefore these instability behaviors make the data-sets as a non-stationary data. "Consequently, the instability behaviors result in the changes in "distributions P (X)". Any incoming transaction which arriving at the proposed model, namely T_t, is trained by the classifier K_{t-1}. Therefore, the riskiest transactions can be those are having less pattern in the database, where K = 0 (since there is no sufficient pattern in our model). In order to mitigate the imbalance effects, we used frequent itemset mining to create a new data-set in training phase.

FIM algorithm is an association rule mining, which its task is to find "frequently occurring attributes" in order to identify a pattern. A set of multiple itemsets (attributes) considers as the input for FIM algorithm to identify patterns. The support of an itemset, is the number of those transactions that contain all the items of that itemset. With this observation that support indicates its regularity or frequency within the database, it plays an important role in frequent item mining, thus in this work, we set the minimum support as 0.9 and selected the large itemset as the pattern.

$$Frequency\ (\delta) = sup\ (\delta)$$
$$= \frac{No.oftransactionswhichcontaintheitemsetX}{Totalno.oftransactions}$$

We clarified the above equation with an example:

For example, if the database contains 700 records and the itemsetδappears in 500 records then the support (δ)=500/700=0.7 = 70%. So, 70% of transactions support the itemsetδ.

The pseudocode of FIM is given in Algorithm 1.

```
Input
Transaction database consists of set of transactions T and set of items I Support, δ
Algorithm
Step 1:   Find all subsets of T of size "S". Let it be TS. TS={$TS_1$,$TS_2$,...$TS_k$} where
            each $TS_i$ is a subset of T of size "δ".
Step 2:   For each $TS_i$ in TS find the set of items that are common in all the
            transactions in $TS_i$. Let $FIS_i$ is the set of items common in all the
            transactions in $TS_i$. Then $FIS_i$ is a frequent item set and insert it
            into the set of closed frequent item sets CFIS,
          Else, already present in CFIS.
Step 3:   Return closed frequent itemset, CFIS.
```

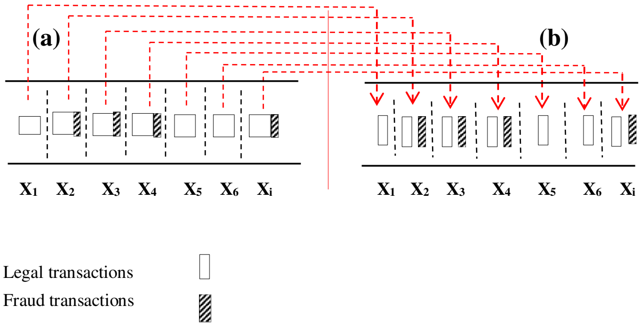

Figure 1. (a) Here, we assume the No. of customers at the rate of "i". We created separate chunks for each customer, which contains the transaction history of a particular customer. (b) The proposed algorithm used to create a new database for each customer to mitigate the imbalancedness problem, (which contains a legal and/or fraud pattern for each customer separately).

The first mining task in this work is to separate each particular customer from the data-set (Figure 1(a)).

$$\left(If\,(X_i) denotedasaparticularcostumer | i \in \{1, 2, 3, 4, \ldots ..i\}\right)$$

We built an individual chunk for each customer, which contains customers' transactions history. These chunks can contain either both fraud and legitimate transactions *(X_2, X_3, X_4, X_i)*, or one of them *(X_1, X_5)*. The blank chunks represent the legitimate transactions, and the hatching chunks represent the fraudulent transactions. From Figure 1(a) it is apparent that most of the transactions belong to the legitimate class *(Z∈{0})*, and only a few transactions belong to the fraudulent class *(Z∈{1})*. Next step, we applied the proposed FIM algorithm to each Chunks (customers' profile), to find the legitimate pattern as well as the fraudulent pattern of each customer. We set the minimum support as S = 0.9 and selected the largest itemset as the pattern for each particular customer. At the end, for each customer, we obtained a balanced profile, which contains "one" legitimate transaction, "one" fraudulent transaction. Each time any new transaction enters the database, a model is learned from the new data-set, which we obtained in this step (Figure 1)(b). Our claim is illustrated in next Section in detail.

As shown in Figure 1, after applying the proposed algorithm to each customer's profile (chunk), we got an obvious different size of customer databases before & after the experiment. Therefore we obtained a fully balanced data-set for each customer since any customers are having only one legitimate and/or one fraudulent transaction in their databases. The main goal of our work is to give some guidelines to researchers on how to tackle the imbalancedness & dynamic behaviors (non-stationary data-set) problems. Although this trick helps us to treat any new behavior, since we created the new pattern database from the original data-set. The proposed strategy not only helps to balance the data-set, but our training data-set will be updated after any incoming transactions and provides recent up-to-date patterns and information for classifiers without removing or discarding any transaction vector, because the classifiers consider patterns for their training.

4. Experimental Assessment

This paper formalizes a new framework for the credit card transactions data-set to reduce the imbalancedness & dynamic environment (non-stationary) problems in order to improve the working conditions of fraud detection techniques. We rigorously evaluated our model by conducting a series significant data to assess the effectiveness of our proposed model. If we assume any incoming transactions, which arriving at our detection model namely T_t, (t corresponds to number of transactions, t∈ { 1,2,….,t}), they are processed by the classifier (model) K_{t-1}. But the riskiest transactions of T_t are the former transactions where t=1, since we do not have that much sufficient supervised samples to create a strong pattern. During processing, legal pattern (LP) and fraud pattern (FP) will be exploited for training and updating the classifier K_t. Our proposed scenario is based on the pattern database, and consequently, the classifiers must be updated after any incoming transactions, since each transaction having different behavior. In this work we plan to develop an FDS to return accurate alert; refer to frauds (correct alerts), which are the true positives (TP). Thus, what in fact matters is to attain the highest precision in the Fraud catching rate. The precision can be measured as:

Let μ be the number of fraudulent transactions in the original data-set. Out of the t% top-ranked candidates, suppose *f (t)* is truly fraud *(f (t) <= t)*.

We can then define precision as *P (t) = (f (t))/t.*

Positive predictive value = Precision

$$\left(\frac{TruePositives}{TruePositives * FalsePositives}\right)$$

4.1. Data-set

In order to evaluate the proposed model, UCSD-FICO data-set is used (FICO is the “leading provider of analytics and decision management technology”). The data-set is real data and consists of a supervised data-set and a label mark added for each transaction by the company. Based on whether the label class is available, credit card transaction data can be classified into supervised & unsupervised methods. Since, in practice, there is a shortage of labeled data-sets, but it is a great significance for us to evaluate our method with a real world credit card dataset. Since the fraud transactions are a negligible portion of total data-set (transactions), thus, are called as minority class, and the legitimate transactions one as majority class. If we define label Z where Z ϵ {0, 1}, 1 denotes a fraudulent (minority class) and 0 a legitimate transaction (majority class). The data-set contains the 100,000 transactions of 76,729 customers in spanning over a period of 91 days (Table 1). This data-set is highly imbalanced (the percentage of fraudulent transactions is 2.1%).

Table 1. Structure of the Data-set.

Legitimate instances	Fraudulent instances	Distinct Costumers
97858	2142	76729

4.2. Data Pre-processing

Data-set pre-processing is one of the most important and vital theories in data mining (García, Luengo, & Herrera, 2016). The task of data pre-processing is to “organize the original data-set”, clear those attributes, which are “irrelevant to our work”, and “simplified data”. In order to address this issue we include a pre-processing phase into our work. The data-set contains “20 fields” including; class label, custAttr1, custAttr2, hour1, and many others. From the data-set we removed those attributes, which are unique for each customer, for example, “custAttr1” is the account/card number and “custAttr2 is e-mail id” of the customer. Both these fields are unique to a particular customer and thus we keep only “custAttr1”. For the best evaluation of our model (SWC-FIM) we removed the transactions corresponds to those customers who have only one transaction in the data-set since for a single transaction it is infeasible to create a pattern.

Table 2. Data-set.

No.of Customers	Number of transactions in Training set			Number of transactions in Testing set		
	Legal	Fraud	Total	Legal	Fraud	Total
200	652	25	489	660	17	677
600	1716	64	1780	1244	48	1292
1000	2604	131	2735	2002	102	2104
1400	3440	158	3598	3083	147	3230

4.3. Formalization of the Learning Problem

Our intuition is that “fraud and legal transactions have to be trained separately,” because, they refer to different classification problems:

1. “fraud misclassification leads to money lose”;
2. “legal misclassification leads to customer dissatisfactions”.

Therefore we create legal and fraud patterns from their legal and fraud transactions respectively. We then train a designed classifier on the created data-set (Table 2).(Figure 2(a)).(Figure 2(c)).(Figure 2(c)).

- Separate *“i”* customers with their transactions’ history from the data-set .
- From each customer split his/her legal and fraud transactions (Figure 2(b)).
- Apply frequent itemset mining to the set of legal transactions of each customer. Store these legal patterns in legal pattern databases
- Apply frequent itemset mining to the set of fraud transactions of each customer. Store these fraud patterns in fraud pattern database
- At the end, for each customer, we have a fully balanced data-set (one legitimate transaction, one fraudulent transaction).
- Each time a new transaction is available, a model is learned on the legal & fraud patterns, which are stored in the database.
- Since this approach leads to training sets and decrease in size the training data-set, so able to avoid overloading.

On the basis of the state-of-art of this work, we can address the imbalanced nature of datasets by conceiving the following strategies. Figure 2 assumes that a particular customer having both fraud & legal transactions in their database, we have three steps (A, B, C) to create a balance profile for each customer.

In step A, we select those customers who are having both fraudulent and legitimate transactions in their database.

Subsequently, in step B we separate, legitimate and fraudulent transaction of each customer.

In step C, after applying FIT algorithm, we create a balanced profile (database) for each customer (since one legitimate and one fraudulent transaction are available for each customer).

The pseudocode of training algorithm is given in Algorithm 2.

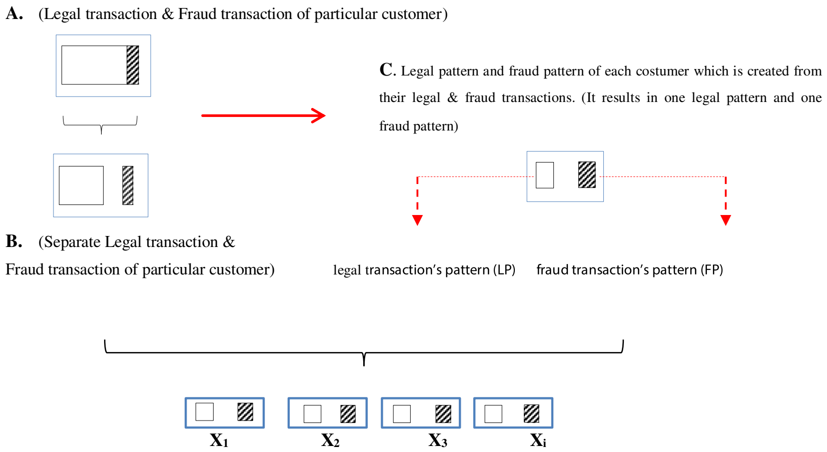
A. (Legal transaction & Fraud transaction of particular customer)
B. (Separate Legal transaction & Fraud transaction of particular customer)
C. Legal pattern and fraud pattern of each costumer which is created from their legal & fraud transactions. (It results in one legal pattern and one fraud pattern)
legal transaction's pattern (LP)
fraud transaction's pattern (FP)
X1
X2
X3
Xi

Figure 2. Example of a particular customer- Section A is an example of the particular customer, which having both fraud & legal transactions in his profile. In Section B, we separated the legal & fraud transactions from the corresponded costumer. Section C is showing the final result, where the FIM algorithm applied on transactions to create legal & fraud patterns.

```
Require: sample L; F
D: original data-set; containing legitimate set L and fraudulent set F;
m: number of costumers;
Z: class label;
Output: D'
Begin
S1 = 0.9
    Select Z // class label; from D; // original data-set
    ZЄ{0}
for i=1 to m
    PlD = max (FIS); //Large Frequent Itemset for legitimate set in D
    Pl (i) = PlD;
    Select Z // class label; from D; // original data-set;
    Z∊ {1}
    PfD=max (FIS); //Large Frequent Itemset for fraudulent set in D
    Pf (i) = PfD;
End for
D' = new pattern database;
Return Pl & Pf & D';
End
```

If we assume Table 2 as training sample, which is called the “original data-set {D}”, we created fraud and legal patterns for any individual customer by using the proposed algorithm (Algorithm 1). The new data-set, which is based on legal and fraud patterns of corresponding customers is called {D'}. We set the “minimum support or frequency rate, as 0.9” and selected the “large itemset as the pattern”. For example, let the largest itemset be the following example:

Hour	*0*	*Flag1*	*0*	*Indicatore1*	*0*
Zip	*50*	***Flag2***	*0*		
Field1	*3*	***Flag3***	*0*		
Field2	*0*	***Flag4***	*0*		
Field3	*7654*	***Flag5***	*1*		
Filed4	*18*	***Indicatore1***	*0*		

Then the corresponding pattern will be as Table 3.

In order to evaluate the benefit of FIM algorithm on imbalanced & non-stationary environments of credit transaction datasets, we develop a classifier, which is based on “score matching”, and then compare the performance of proposed classifier with other outstanding classifiers in credit card fraud detection. Classification is the result of supervised learning, which means that there is a known label that we want the system to generate, while clustering comes under unsupervised learning where the class labels are not available, as in this paper we are using supervised learning (where the class label is available), so, classification is the best solution. After creating a new database for each customer and make a data-set as balanced data, the proposed classifiers must cross these fraud and legal pattern databases in order to detect frauds. These new databases (pattern databases) are much smaller than original credit card transaction databases, as they contain “only one record corresponding to a legal pattern and one record corresponds to a fraud pattern for each particular customer”. In this work we propose a new classifier to show the impact of our new strategy in the credit card fraud detection technique. The novelty of this research centralized on pattern database, we have converted our original imbalanced data-set (D) to smaller pattern database (D'). Here, we assume the size of pattern databases is m × s, where m is the number of customers and s is the number of features. First we select the new database, which is D', it contains legal & fraud patterns of each costumers.

Table 3. “9999” represents an “invalid value,” which means that the field has dis-similar values in each transaction, hence it is not contributing to the pattern.

0	9999	50	3	0	7654	18	0	0	0	0	0	1

Step 1. “*CPl* is the number of features in the incoming transaction, which is matching with that of the legal pattern of the corresponding customer”.

Step 2. “*CPf* is the number of features in the incoming transaction, which is matching with that the fraud pattern of the corresponding customer”.

Step 3. “If $CPf = 0$ & *CPl* is more than the defined threshold, the incoming transaction is legal”.

Step 4. “If $CPl = 0$ & *CPf* is more than the defined threshold, the incoming transaction is fraud”.

Step 5. “If both *CPl* & $CPf > 0$ and $CPf \geq CPl$, the incoming transaction is fraud or else it is legal”.

The “pseudocode of (SWC-FIM) is given in Algorithm 3”

```
Require: Pl (legal pattern); Pf (fraud pattern);
D: original data-set; containing legitimate set L and fraudulent set F;
D' = the new pattern database;
m: number of costumers;
Z: class label;
S: number of features
$T_t$: incoming transaction
φ: predefined threshold
Output: legal (0); fraud (1)
Begin
Select D'
CPl=0;//legal feature match count,
CPf=0;//fraud feature match count
    for j=1 tom do
    if (PI(j,1) = $T_t$(1)) //First attribute
    for i=2 to s do
    if (PI(j,i) is valid and PI(j,i)=$T_t$(i))
    CPI = CPI + 1;
    endif; endfor; endif; endfor;
{\rm K}_{t}: $\mathbb{R}^n$ ⟶ {0,1}, so, each feature vector $f$∈$\mathbb{R}^n$, and label {\rm K}_{t} \left(
  { f} \right) \in \{ 0,1\}
    for j=1to m do
    if (Pf(j,1) = $T_t$(1))
    endif; endfor;
    φ = K;
    if (CPf = 0) & (CPI >=K);
    return (1);
    else return (0);
    endif
elseif (CPI=0) & (CPf > K)
return (1);
elseif (CPI > 0 && CPf > 0)
    if(CPf >=CPI) then return(1);
    else return(0);
    endif; endif;
End
```

5. Discussion

The effect of the imbalancedness problem is ignored in previous credit card fraud detection (Ali et al., 2015; Dal-Pozzolo et al., 2015; Liu et al., 2009; Nekooeimehr & Lai-Yuen, 2016; Yang & King, 2009; Zhang et al., 2016). In this paper, we proposed a new model to handle the imbalanced and non-stationary problems in credit card data, which are the most striking characters in fraud detection classification process. Due to dynamic environment of data-set, the customer & fraudulent behaviors are found to be changing gradually over an unreasonable period of time. This makes it difficult to develop efficient fraud detection methods or may degrade the performance of methods. Therefore the effective fraud detection methods should be adaptive with these two strike characters (highly class imbalance & non-stationary problems). The specification of the proposed model is, i) provide a different framework for credit card transaction data-set to overcome the imbalancedness problem by creating legal and fraud patterns of each individual customer, in this way we can create a new database that is much smaller than the original database. ii) We document a new insight in non-stationary environment by creating a new profile for each customer and the behavioral changes (customer & fraudulent behaviors) can be incorporated into our model by updating the fraud and legal pattern databases after any transactions. The proposed model shows that learning from those customers that having both legal & fraud pattern in their database is a different problem than learning from others, (which having either legal or fraud patterns in their databases). One of the differences is, applying two times FIM algorithm in order to create a legal and fraud pattern. Because as we discussed earlier, our intuition is that "fraud and legal transactions have to be trained separately" since they refer to different classification problems. The other difference, which is evident, provides recent up to date information for classifiers, since two patterns (legal and fraud) are available for classifiers.

Also the performance of the proposed. "The fraud detection model (SWC-FIM) is compared with tree states of the art classifiers used for credit card fraud detection; support vector machine, random forest and naïve bayes. These are the base classifiers used in the state-of -art financial fraud detection models described in the literature review (Figure 3 & Figure 4). We applied SMOTE (Chawla et al., 2002; Dal-Pozzolo et al., 2014), an oversampling technique that used (common-to-handle class imbalance) before giving the data to the classifiers and compared the result with our proposed model. But the performance of the classifiers in using SMOTE is found degrading, because of the highly imbalanced' nature of the data-set. From the performance evaluation (Figure 3) it is found that the FIM algorithm is having a significant impact for balancing the data-set. SWC-FIM model is showing the highest performance (Figure 3 & Figure 4) than other classifiers. From the results we can obtain that, our proposed model are capable to handling class imbalance and SWM-FIM showed very good performance according to these measures (AUC, Precision) compared with other classifiers (Figure 2 & Figure 3).

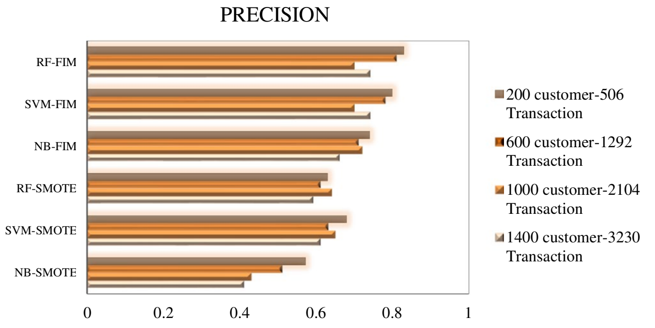

Figure 3. Comparison of sampling technique (SMOTE) and our suggested scenario in terms of precision. The results show that our scenario has considerable results in comparing SMOTE, (which the popular resampling technique in handling imbalanced datasets).

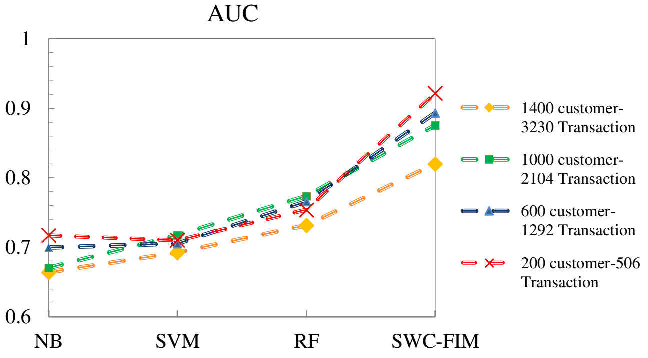

Figure 4. The performance of the proposed model (SWC-FIM) is compared with tree states of the art classifiers used for credit card fraud detection; support vector machine, random forest and naïve bayes (these are the base classifiers used in the state-of-art financial fraud detection models described in the literature review) in terms of AUC. The results shows that our proposed model has better performance even with increasing the number of transactions.

6. Conclusion

This paper formalizes a new framework for the credit card transactions data-set to reduce the imbalancedness & dynamic environment (non-stationary) problems in order to improve the working conditions of fraud detection techniques. To this end, the proposed model is created by two set patterns; fraud & legal patterns. Our intuition is that "fraud and legal transactions have to be trained separately," because they refer to different classification problems, since;

1. -"fraud misclassification leads to money lose";
2. -"legal misclassification leads to customer dissatisfactions".

Due to dynamic environment of the data-set, the customer & fraudulent behaviors are found to be changing gradually over an unreasonable period of time. This makes difficulties to develop efficient fraud detection methods or may degrade the performance of the methods. Therefore the effective fraud detection methods should be adaptive with these two strike characters (highly class imbalance & non-stationary problems). The specification of the proposed model is; to consider and handle these two issues, and the behavioral changes that refer to non-stationary problem can be incorporated into our model by updating the fraud and legal pattern databases after any transactions. Moreover "our proposed model takes very less time, which is an important parameter of real-time applications," because our model is done by crossing the smaller pattern databases rather than the large transaction database. In future work we can focus on infrequent itemset mining to further improve the precision in non-stationary environment.

Acknowledgment

This research is partly supported by NSFC, China (No: 61572315, 6151101179) and 973 Plan，China (No. 2015CB856004)

Disclosure statement

No potential conflict of interest was reported by the authors

Notes on contributors

Masoumeh Zareapoor is working as a postdoctoral researcher at the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University. Prior to that, she was an Associate Researcher in Tokyo University of Science, Japan. Her research interest includes data mining & machine learning.

Jie Yang is the director of Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University. He has served as the principal investigator of over 30 nation and ministry scientific research projects in image processing, pattern recognition, data mining. He is the holder of 32 patents and has been awarded six research achievement awards from ministry of Education, China.

References

Ali, A., Shamsuddin, S.M., & Ralescu, A.L. (2015). Classification with class imbalance problem: A Review. *International Journal of Advances in Soft Computing & its Applications, 7*, 176–204.

Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligent, 16*, 341–378.

Chen, S., & He, H. (2011). Towards incremental learning of nonstationary imbalanced data stream: A multiple selectively recursive approach. *Evolving Systems, 2*, 35–50.

Dal-Pozzolo, A., Caelen, O., Le Borgne, Y.L., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications, 41*, 4915–4928.

Dal-Pozzolo, A., Caelen, O., & Bontempi, G. (2015). When is undersampling effective in unbalanced classification tasks? *Machine Learning and Knowledge Discovery in Databases. ECML PKD. Lecture Notes in Computer Science, 9284*, 200–215.

Dong, A., Chung, F.L., & Wang, S. (2016). Semi-supervised classification method through oversampling and common hidden space. *Information Sciences, 349*, 216–228.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drifts adaptation. *ACM Computing Surveys, 46*, 1–37.

García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems, 98*, 1–29.

Han, F., Lei, M., Zhao, W., & Yang, J. (2012). New support vector machine for imbalance data classification. *Intelligent Automation & Soft Computing, 18*, 679–686.

He, H. & Garcia, E.A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*, 1263–1284.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis, 6*, 429–449.

Kolter, J.Z. & Maloof, M.A. (2007). Dynamic weighted majority: An ensemble method for drifting concepts. *Journal of machine learning research, 8*, 2755–2790.

Liu, X.Y., Wu, J., & Zhou, Z.H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transaction on Cybernetics, 39*, 539–550.

Nekooeimehr, I., & Lai-Yuen, S.K. (2016). Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets. *Expert Systems with Applications, 46*, 405–416.

Nian, K., Zhang, H., Tayal, A., Coleman, T., & Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. *The Journal of Finance and Data Science, 2*, 58–75.

Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: classification of skewed data. *ACM SIGKDD Explorations Newsletter, 6*, 50–59.

Quah, J.T.S., & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. *Expert Systems with Applications, 35*, 1721–1732.

Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., & Zhou, Y. (2015). A novel ensemble method for classifying imbalanced data. *Pattern Recognition, 48*, 1623–1637.

Sundarkumar, G.G., & Ravi, V. (2015). A novel hybrid undersampling method for mining unbalanced datasetsin banking and insurance. *Engineering Applications of Artificial Intelligence, 37*, 368–377.

Van-Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems, 75*, 38–48.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers and Security, 57*, 47–66.

Weston, D.J., Hand, D.J., Adams, N.M., Whitrow, C., & Juszczak, P. (2008). Plastic card fraud detection using peer group analysis. *Advances in Data Analysis and Classification, 2*, 45–62.

Wong, N., Ray, P., Stephens, G., & Lewis, L. (2012). Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results. *Information Systems Journal, 22*, 53–76.

Yang, H., & King, I. (2009). Ensemble learning for imbalanced e-commerce transaction anomaly classification. *ICONIP 2009, Lecture note in computer science, 5863*, 866–874.

Zhang, Z., Krawczyk, B., Garcìa, S., Rosales-Pérez, A., & Herrera, F. (2016). Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data. *Knowledge-Based Systems, 106*, 251–263.