Machine Learning based Approach to Financial Fraud Detection Process in Mobile Payment System

Dahee Choi and Kyungho Lee*
CIST, Korea University, Seoul, Korea
{shoodol00, kevinlee}@korea.ac.kr

Abstract

Mobile payment fraud is the unauthorized use of mobile transaction through identity theft or credit card stealing to fraudulently obtain money. Mobile payment fraud is the fast growing issue through the emergence of smartphone and online transition services. In the real world, highly accurate process in mobile payment fraud detection is needed since financial fraud causes financial loss. Therefore, our approach proposed the overall process of detecting mobile payment fraud based on machine learning, supervised and unsupervised method to detect fraud and process large amounts of financial data. Moreover, our approach performed sampling process and feature selection process for fast processing with large volumes of transaction data and to achieve high accuracy in mobile payment detection. F-measure and ROC curve are used to validate our proposed model.

Keywords: Machine Learning, Mobile Payment Fraud, Financial Fraud Detection, Semi-Supervised Method, Feature Selection

1 Introduction

Mobile payment system is the fast growing issue since the mobile channel can facilitate nearly any type of payments. More than 87 percentage of merchants supports either mobile site, mobile application for online shopping or both[15]. Supporting for mobile wallets also helps to increase the overall use of mobile payment. As a result, mobile payments have reached $194.1 billion in 2017 and mobile proximity payments also increased to $30.2 billion in 2017, compared to $18.7 billion in 2016[1].

Due to the rapid increase of mobile commerce and the expansion of the mobile payment market, the fraud in mobile payment has arisen and becomes more common. Mobile payment fraud can be occurred in several ways, but the most frequent case is an unauthorized use of mobile payment via credit card number or certification number. The mobile payment does not require the presence of a physical payment tool. Instead, mobile payment needs some important information such as card number, expiration date, card verification code and pin number to make fraudulent payment in a mobile payment environment.

To address rapidly arising in mobile payment fraud problem, financial institutions employ various fraud prevention tools like real-time credit authorization, address verification systems (AVS), card verification codes, rule-based detection etc[21]. However, existing detection systems depend on defined criteria or learned records which makes it difficult to detect new attack patterns. Our proposed research is based on unsupervised learning methods to capture new mobile payment fraud and are also based on supervised learning for accurate classification of mobile payment fraud. Also, our research proposed the overall process of detecting mobile payment fraud by applying machine learning process to detect unknown and underlying fraud threats.

IT CoNvergence PRActice (INPRA), volume: 5, number: 4 (December 2017), pp. 12-24
*Corresponding author: Center for Information Security Technologies, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, Korea, Tel: +81-2180-3937

2 Related Work

As a traditional method of financial fraud detection, the Dempster–Shafer adder (DSA) based on Dempster–Shafer theory and use of Bayesian learning research has been proposed. A transaction conversed into the suspicion score, which can be referred as the probability of fraudulent, based on the index in the transaction history database. BLAST and SSAHA algorithm are sequence alignment algorithms and used as the alignment of sequences for an efficient technique to examine the spending behavior of customers[19]. To calculate and predict the probability from the user's existing financial information and to build a multilayer model of program behavior, Hidden Markov Model (HMM) has been proposed[30].

Machine learning based research has also been proposed, as a web service-based collaborative scheme for credit card fraud detection[23]. The detection of fraud based on the genetic algorithm calculation and customer‘s behavior [26], and an efficient financial fraud detection system which is adaptive to the behavior changes by combining classification and clustering techniques, scalable algorithm named BOAT, has also been proposed[29]. Decision trees and Support Vector Machine (SVM)[27], combined method of Decision tree, Neural Networks, Logistic regression[28], Self-Organizing Map (SOM) combined with Gaussian function[34], and Fuzzy logic combined with Self-organizing map have been introduced for financial fraud detection method[6]. A combined method of SVM, Random forests, Logistic regression[20], Self-Organizing Map Neural Network (SOMNN), Genetic Algorithm with behavior based technique and Hidden Markov Model (HMM) has been attempted[8].

The previous research papers are mainly related to specialized approach such as algorithms, which needs further step for implementation. Moreover, in applying the methods of machine learning, previous research only used one of the learning methods between supervised and unsupervised learning. However, our research has performed and described about the overall process of mobile payment fraud detection in practical perspective based on supervised and unsupervised learning method. Furthermore, our research expects to applicate for practical use since our experiment has validation score for each process and is based on real mobile payment data.

Summarized contribution of our proposed research are: 1) Performing and validating of the overall process for mobile payment fraud detection. 2) Based on either supervised and unsupervised learning method. 3) Proposing of practical method by applying sampling process and feature selection process for rapid detection in real world and verify our proposed model with real mobile transaction dataset.

3 Model and Methodology

The proposed model consists of data pre-processing, sampling, feature selection, application of classification and clustering algorithm based on machine learning. In this paper, verification step is performed for each step to verify the effectiveness of proposed mobile payment fraud detection model.

In pre-processing process, data correlation analysis and data cleaning process which cleans the noise data are performed. Also data transformation, integration and reduction are included in this process. Following process is the sampling process which evaluates dataset with various ratios for verification through random over-sampling and under-sampling method. Feature extraction and selection process have performed by the filter based method of feature selection algorithms. After the feature selection process, clustering process with the proposed algorithm performs and this result will be used as a training and validation set of classification process. By applying supervised algorithms to the previous result, which was derived in the clustering process previously, higher prediction could be achieved. The model validation process is performed with precision and recall rate through F1 measure. Figure 1 represents an overall flow of the processes described above in this paragraph.

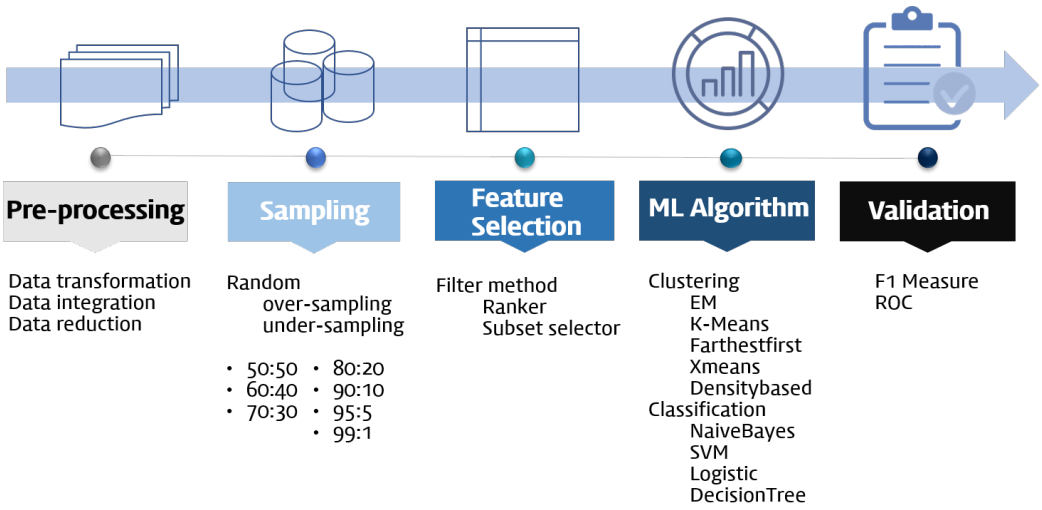

Figure 1: Overall framework of mobile payment fraud detection process

3.1 Data Description

Our research was conducted based on the actual mobile payment data occurred in Korea, 2016. With the agreement of a major financial institution, mobile payments provider collected mobile payment data for 6 months from June to November. A total of 260,000 pieces of data were extracted from the September data, which were used as learning data. The data received were labeled as normal, abnormal, or canceled. A normal transaction is a transaction that succeeds and is identified as a normal approval and the settlement is completed. The abnormal transaction is a transaction that has not been identified as a normal approval until six times after the settlement. The range of transaction prices, the place of settlement, the change in the transaction type and others are analyzed for classification. Among data. 21 characteristics are extracted as features.(transaction serial number, cash classification, transaction type, certification date, authentication time, transaction status, communication company, phone number, transaction amount, corporation ID, shop ID, service ID, Email usage, email hash, revenue breakdown, ip information, authenticated client version etc). For the protection of personal information, key information has been anonymized and data which can identify an individual has been converted.

3.2 Pre-processing

Table 1: Characteristic of the mobile payment dataset

Number of Normal Transactions	Number of Abnormal Transactions	Number of Attributes
274670	2402	21

3.3 Sampling

A real transaction dataset of mobile payment contains a data imbalanced problem. Imbalanced problem in the data could mislead mobile payment fraud detection process to misclassifying problem. Previous research has proposed a random minority over-sampling method and cluster based under-sampling approach for selecting the representative data as training data to improve the classification accuracy for minority class[32].

To solve this problem, our research generate various datasets using SMOTE and RUS for more accurate experiment. SMOTE is an over-sampling technique that uses a method of generating arbitrary examples, rather than simply oversamples through duplication or replacement[10]. Also, random under-sampling (RUS) method was applied for downsizing the normal transactions by extracting a sample data randomly. Since low ratio of anomalous data might lead to less precise results, our research applied both SMOTE and RUS for generating the different ratio of sampling dataset to increase the reliability and accuracy of our proposed research. Sampling ratios are divided into 50%, 60%, 70%, 80%, 90%, 95% and 99% of normal transaction ratio.

3.4 Feature Selection

Feature selection has been proven to be effective and efficient for data mining and machine learning problems. The objectives of feature selection include building simpler and more comprehensible models, improving data mining performance (predictive accuracy and comprehensibility), and preparing clean (redundancy and irrelevancy removal), understandable data[35][5].

Feature selection can be divided into three categories, which are wrapper method, filter method and embedded method. The wrapper method relies on the predictive performance of a predefined learning algorithm to evaluate the selected features. It repeats the searching step and evaluating criteria until desired learning performance is obtained. The drawback of wrapper method is that the search space is extremely large and it is relatively expensive than other methods. Filter method is independent of any learning algorithms and rely on certain characteristics of data to assess the importance of features. Features are scored based on the scores according to the evaluation criteria, and the lowest scored features are removed. Embedded method is a combined method between the filter and wrapper methods which embed the feature selection with the model learning[16][4].

3.5 Machine Learning

Machine learning can be divided into supervised learning and unsupervised learning based on the learning method. Supervised learning is a method of classifying with labeled data, mainly used for accurate classification and prediction. Supervised method could classify anomaly data by relatively high accuracy, but could not detect the underlying or unknown anomalies related to mobile payment fraud. The other common method used in machine learning is unsupervised learning which cluster unlabeled data with similar attributes. A clustering process in unsupervised method could manage the unlabeled data and be useful for detecting the hidden anomalies in mobile payment fraud detection. However, unsupervised method usually performs lower accuracy than supervised method.

In this research, we propose the combined method of supervised and unsupervised method called semi-supervised. Firstly, the unsupervised method to the clustering process divide the mobile transaction data in several groups. Since the dataset, which include the mobile payment fraud, in the real world are unlabeled, unsupervised learning performs the key role in our proposed detection process. Then the supervised method classifies the clustered groups into normal mobile transaction data and fraud data. In other words, our proposed approach applies unsupervised method for arrangement and summarization of discovering anomalies to detect mobile payment fraud. Afterwards, classification and regression process are performed for accurate prediction which is the supervised method. These overall combined processes, improve the detection efficiency by discovering hidden fraud data in unlabeled dataset and accurate classification.

3.5.1 Unsupervised Learning Algorithms

In this research, five kinds of unsupervised learning algorithms are applied for clustering mobile transaction data which is unlabeled. A brief description of applied algorithms used for experiments in this research is as follows.

EM algorithm is an iterative method for finding parameter of maximum likelihood or maximum a posteriori in statistical models, where the model depends on unobserved latent variables. EM algorithm generates an initial model randomly and then iteratively performs a refinement process to generate an optimized model. In addition, EM algorithm generates an optimal model by adjusting the probability that each object belongs to a mixture model through an iterative refinement process[9].

K-Means algorithm has K centroids that are defined for each cluster while costs are calculated by distances of centroids. The remaining objects are calculated from the distance to the selected k center points and assigned to the closest cluster and a new cluster center (average) is obtained for each cluster. This process is repeated until there is no movement of the center point, that is, until the center point converges[12].

FarthestFirst algorithm is a variant of Kmeans that places each cluster center in turn at the point furthest from the existing cluster centers. It proceeds through two scans of the dataset. The first scan constructs a number of hash table data structures equal to the number of characteristics based on the information about the characteristic value and frequency. In the second scan, the property values in the corresponding hash table are determined by the expected time and frequency. In a typical FarthestFirst algorithm, the starting point is randomly selected. The existing KMeans algorithm needs to find a new center point continuously as data is added. However, FarthestFirst is relatively fast since the number of center point changes is small[13].

XMeans algorithm is an extended type of KMeans, which is a top-down algorithm that starts from one group and gradually divides the group. A Bayesian Information Criterion (BIC) score is used as a criterion for partitioning. The BIC method is a statistic that approximates the posterior probability distribution calculated using the likelihood function and the prior probability distribution in Bayesian theory. BIC score increases until there is no further increase in the score. After a group is divided, the data assigned to each group are determined by KMeans[22].

Density-based clustering techniques have the advantage that prior information about the number or type of clusters is not needed as input variables. Density-based algorithm estimates the probability distribution of an unobservable underlying probability density function based on observed data. Clusters with a minimum number of instances within a given radius are available, even if the data has noise and anomalies[7].

3.5.2 Supervised Learning Algorithms

In this research, seven types of supervised learning algorithms were applied for accurate classification of fraud data in the mobile transaction dataset.

NaiveBayes is a kind of probability classifier applying Bayes theorem that assumes independence between properties. All Naive Bayes classifiers commonly assume that all property values are independent of each other. In some probabilistic models, the Naive Bayes classification can be trained very efficiently in a supervised learning environment. In many practical applications, parameter estimation of the Naive-Bayes model uses maximum likelihood estimation (MLE) and the training is possible without bayesian probabilistic or bayesian methods. In addition, the amount of training data to estimate the parameters required for classification is very small and works well in complex real-world situations[33].

Support Vector Machine (SVM) is a representative model of supervised learning method. SVM is a linear classifier that mainly determines which data category belongs to new data based on a given data

set. The basic principle of SVM is to measure the data distance of each group and to divide the boundary by the maximum margin size[31].

Logistic regression is a statistical technique that is used to predict the likelihood of an event using a linear combination of independent variables as a probability model. The objective of logistic regression is to provide a general regression in the relationship between the dependent variable and the independent variable expressed as a concrete function and used in future prediction models. Moreover, the result of the data is classified into a specific classification when the input data is given[14].

OneR is a classification algorithm that generates one rule for each predictor in the data, then selects the rule with the smallest total error. OneR uses OneR classifier to find out the attributes' weights. For each attribute, it creates a simple rule based only on that attribute and then calculates its error rate.

Decision tree is a predictive model that connects observation values and target values. In this tree structure, the leaf (leaf node) represents a logical product of features associated with a class label representing a class label. The learning of a decision tree is a process whereby a set of data used for learning is divided into subset . This process is recursively repeated on each separate subset of data until a new predictor is no longer added due to partitioning, or until the subset of nodes has the same value as the target variable. C4.5, RandomForest, RandomTree are the kinds of Decision tree.

C4.5 algorithm was proposed to overcome the limitations of ID3 algorithm. C4.5 algorithm improves the ID3 algorithm by dealing with both continuous and discrete attributes, missing values and pruning trees after construction. C4.5 algorithm handles continuous attributes by splitting the attribute's value range into two subsets and searches for the best threshold that creates the first subset and all other values constitute the second subset[25].

RandomForests or random decision forests are an ensemble learning method for classification and regression. Ensemble learning method is the method that generates many classifiers and aggregate their results. Two well-known methods are boosting and bagging, and RandomForests add an additional layer of randomness to bagging. In standard trees, each node is split using the best split among all variables, however, in a RandomForests, each node is split using the best among a subset of predictors randomly chosen at that node[17].

RandomTree constructs a tree that considers k randomly chosen attributes at each node. RandomTree operator works exactly like the Decision tree operator except that for each splitting, a random subset of attributes is available. Furthermore RandomTree selects a random subset of attributes before it is applied[2][3].

3.6 F-Measure

In machine learning method, which is based on statistics, F-measure is a well-known measurement for model performance between predicted class and actual class using Recall and Precision. In our research, the F-measure is used to measure the ratio between the actual value and the value that the algorithm detects and predicts[24].

Table 2: Confusion Matrix

	Predicted Positive	Predicted Negative
Positive	TP	FN
Negative	FP	TN

Recall is the proportion of real positive cases that are correctly predicted positive which can be defined as

$$Recall : \text{TP/(TP+FN)} \tag{1}$$

Precision denotes the proportion of predicted positive cases that are real positives. Precision is defined as

$$Precision : \text{TP/(TP+FP)} \tag{2}$$

As a result, Accuracy is defined as

$$Accuracy : \text{(TP+TN)/(TP+FN+FP+TN)} \tag{3}$$

and F-Measure is defined as

$$F-Measure : \text{2x(PrecisionxRecall)/(Precision+Recall)} \tag{4}$$

3.7 ROC Curve

ROC curve is widely used to determine the efficiency of diagnostic methods and suitable method to visualize classifier's performance in order to select a suitable operating point, or decision threshold. The ROC curve is a representation of sensitivity and specificity on the two-dimensional plane. The larger the area under the ROC curve (AUC), the better the diagnostic method[11].

4 Experiment

4.1 Feature Selection

In this research, we aim to analyze the characteristics of inherent mobile transaction data and propose a process specific to mobile fraud detection by applying various algorithms. In this reason, we applied filter based feature selection algorithms for feature selection method among filter, wrapper and embedded methods. Feature selection based on filter method is categorized into ranker and the subset selector[4]. We applied 8 subset selector feature selection algorithms and 6 ranker feature selection algorithms to select features among 21 existing features. We assigned the scoring of subset approach to evaluate features based on frequency, and ranker algorithms to calculate weight with higher rank. The results of two feature selection algorithms are combined to prioritize the features by selecting features which exceed the setting number or threshold in frequency and ranking. We also measured the accuracy and f-measure by applying the various clustering algorithms before and after selecting the features in order to verify the effectiveness of our proposed research in feature selection method for mobile payment fraud detection.

After the feature selection process, there was a difference according to the sampling rate, but average accuracy was increased in all algorithms. For more accurate comparison the f-measure was measured and the results are as follows.

The f-measure also differed according to the sampling rate, but the average f-measure value increased in all kinds of algorithms. The validity of the derived features was verified by comparing the measured values of the f-measure with the accuracy before and after the feature selection process.

There was a difference in accuracy and F1 score value according to the sampling rate, but the average value increased in all algorithm types including EM, KMeans, FarthestFirst, XMeans and MakeDensity. This proves the utility of the process we propose.

4.2 F-Measure validation

After feature selection, data were clustered by applying unsupervised learning algorithm. Each clustering is composed of different characteristic values by the clustering algorithm, and the ratio of mobile financial

Table 3: Accuracy and F-Measure before Feature Selection

Ratio	Measurement	EM	KMeans	FarthestFirst	XMeans	MakeDensity
50:50	Accuracy	50.4746	52.3897	51.9317	52.3897	52.2315
	F-Measure	0.5297	0.4815	0.2692	0.4815	0.4806
60:40	Accuracy	51.2906	50.7369	54.9542	50.7369	50.6037
	F-Measure	0.5787	0.5186	0.6869	0.5186	0.5179
70:30	Accuracy	52.4646	51.1657	62.0275	51.1657	51.2531
	F-Measure	0.6222	0.6159	0.7527	0.6159	0.6163
80:20	Accuracy	53.3722	52.6436	68.7552	52.6436	52.6644
	F-Measure	0.6575	0.6540	0.8085	0.6540	0.6541
90:10	Accuracy	53.98	54.4671	75.6495	54.4671	55.0541
	F-Measure	0.6841	0.6892	0.8590	0.6892	0.6932
95:5	Accuracy	53.7094	55.3705	82.448	80.7868	55.7202
	F-Measure	0.6896	0.7053	0.9028	0.8926	0.7076
99:1	Accuracy	61.0958	56.5792	71.4167	56.5792	57.6125
	F-Measure	0.7574	0.7212	0.8328	0.7212	0.7296
AVG	Accuracy	53.7696	53.3361	66.7404	56.967	53.5913
	F-Measure	0.6456	0.6265	0.7307	0.6532	0.6284

Table 4: Accuracy and F-Measure after Feature Selection

Ratio	Measurement	EM	KMeans	FarthestFirst	XMeans	MakeDensity
50:50	Accuracy	68.3847	66.0117	50.1374	66.0117	66.4363
	F-Measure	0.6134	0.6443	0.6647	0.6443	0.6492
60:40	Accuracy	64.6794	65.4996	59.8834	65.4996	65.7077
	F-Measure	0.6319	0.6820	0.7473	0.6820	0.6835
70:30	Accuracy	61.0033	64.7627	69.6545	51.5987	64.8626
	F-Measure	0.6453	0.7112	0.8202	0.54	0.7121
80:20	Accuracy	57.1274	53.9092	79.4255	62.2689	50.3331
	F-Measure	0.6540	0.6268	0.8849	0.7178	0.5859
90:10	Accuracy	53.2598	51.0866	89.1799	69.6503	52.8268
	F-Measure	0.6610	0.6402	0.9426	0.8126	0.6813
95:5	Accuracy	52.627	52.627	84.1299	62.7352	52.6395
	F-Measure	0.6732	0.6732	0.9130	0.7608	0.6733
99:1	Accuracy	52.0083	58.6875	55.7583	60.7542	51.5917
	F-Measure	0.6829	0.7373	0.7143	0.7538	0.6775
AVG	Accuracy	58.4414	58.9406	69.7384	62.6455	57.7711
	F-Measure	0.6516	0.6735	0.8124	0.7016	0.6661

fraud data is different through random sampling. However, since various sampling ratios with randomly extracted in stratified structure are set, it is general compared to the other previous sampling method.

Also, it is important to maintain accuracy and reflect the original data's characteristics which is based on previous model. For these reasons, we applied stratified sampling after the entire model building process has finished. Unlike simple random sampling, which draws a sample from the population in entirety, stratified sampling picks separate samples from separate groups, called as strata or sub-populations [18]. Stratified sampling is a practice of selecting individual records with probability proportional to the variance of estimating statistics on their strata. Stratification randomly divides the dataset so that each class

is correctly distributed in the training and test sets.

Conclusively, in our research, EM, KMeans, FarthestFirst, XMeans and MakeDensity algorithms were applied. For accurate classification, we applied supervised learning algorithms to clustered data. The supervisied learning algorithms used for classification are Naive Bayes(NB), SVM, Logistic Regression (LR), OneR, C4.5, RandomForest (RF). Since both supervised and unsupervised learning methods are used, our proposed research can be called semi-supervised learning approach. The F-score value after applying the classification algorithm are as Table 5.

Table 5: F-measure after classification

	Ratio	NaiveBayes	SVM	Logistic Regression	OneR	C4.5	RandomForest
EM	50:50	1	0.988	0.999	1	1	1
	60:40	1	0.989	0.999	1	1	1
	70:30	0.999	0.999	1	0.983	0.999	1
	80:20	1	0.994	1	1	1	1
	90:10	1	1	1	1	1	1
	95:5	1	1	1	1	1	1
	99:1	0.993	1	1	1	1	1
KMeans	50:50	0.949	0.995	1	0.912	0.997	0.998
	60:40	0.952	0.996	1	0.910	0.998	0.999
	70:30	0.952	0.996	1	0.912	0.999	0.999
	80:20	0.908	0.999	1	0.906	0.999	0.999
	90:10	0.908	1	1	0.898	0.999	1
	95:5	1	1	1	1	1	1
	99:1	0.959	1	1	0.929	1	1
FarthestFirst	50:50	0.998	0.999	1	0.983	1	1
	60:40	0.998	0.999	1	0.983	0.999	1
	70:30	0.999	1	1	0.983	0.999	1
	80:20	0.999	1	1	0.983	1	1
	90:10	0.999	1	1	0.984	1	1
	95:5	0.995	0.999	1	0.828	1	1
	99:1	0.979	1	1	0.736	1	1
XMeans	50:50	0.949	0.995	1	0.912	0.997	0.998
	60:40	0.952	0.996	1	0.910	0.998	0.999
	70:30	0.996	1	1	0.988	1	1
	80:20	0.949	0.997	1	0.915	1	1
	90:10	0.959	1	1	0.880	1	1
	95:5	0.947	0.999	1	0.913	1	1
	99:1	0.999	1	1	0.908	1	1
MakeDensity	50:50	0.956	0.989	0.997	0.927	0.998	0.999
	60:40	0.956	0.992	0.998	0.918	0.999	0.999
	70:30	0.955	0.993	0.998	0.919	0.999	0.999
	80:20	0.975	1	1	0.977	1	1
	90:10	0.974	1	0.999	0.975	1	1
	95:5	1	1	1	1	1	1
	99:1	1	1	1	1	1	1

4.3 AUC validation

In addition to the f-measure value, the AUC value was specified for more accurate verification. The x-axis represents the false positive value and the y-axis represents the true positive value. By calculating the scope of the verification of auc, we used the k-fold cross validation method for verification. The k-fold cross validation method divides the data into k groups, sets the remaining data except for one group as training data, and one group is divided into test data for later verification. The evaluation in k-fold cross validation was carried out by obtaining the mean and standard deviation of the precision obtained from the test data. Training data then can be divided into training set and validation set for more accurate verification. The k value for k-fold cross validation was set to 10 in our research.

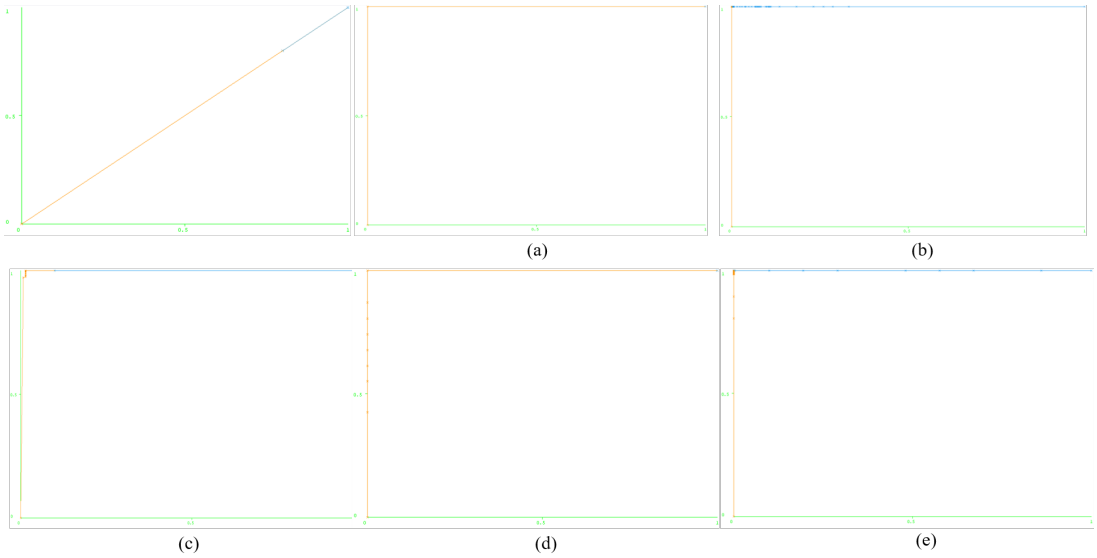

Figure 2: Variance of AUC graph in mobile financial fraud detection

Figure 2 shows the result of classifying the raw data of mobile payment transaction data with C4.5 algorithm that has undergone only the pre-processing process. The raw dataset of mobile payment transaction data results 0.4997, which means if the AUC value is less than 0.5, it indicates that the classification is not performed correctly. Other graphs, (a) through (e) which (a) means the clustering result of EM, (b) indicates KMeans, (c) indicates FarthestFirst, (d) indicates XMeans, and (e) indicates MakeDensity, are the result of sampling, feature selection, clustering and classification with C4.5 algorithm performed by our proposed process. The data from (a) to (e) is a ratio of 90:10. As a result, (a) records 1, (b) records 0.9997, (c) records 0.996, (d) records 0.9998, and (e) records 0.9997 from the perspective of AUC value.

5 Conclusion

In this paper, we proposed a process for detecting mobile financial fraud transaction. Our research examined the performance of data mining method in mobile payment fraud detection with generative process in specifically, including feature selection method. Considering that previous research in mobile payment fraud detection commonly focused on a single perspective of data mining methods, our proposed research combined both supervised and unsupervised methods. Moreover, our proposed research on semi-supervised approach expected to detect underlying mobile payment fraud and new fraudulent behaviors with unlabeled data in an effective way. Effective way indicates the direct classification of mobile payment fraud detection with a good precision via regression. We also demonstrated the practicality of

our research by using actual mobile payment transaction data. The effectiveness of our proposed research is assessed by f1 score and AUC.

5.1 Future Work

As the number of financial transactions increases, financial fraud is increasing by using a different method than before, which means that more and more frauds cannot be detected with existing rule-based models. Therefore, in addition to rule-based detection, a self-detectable model should be designed using neural networks with proper process.

5.2 Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-2015-0-00403) supervised by the IITP(Institute for Information and communications Technology Promotion).

References

[1] Javelin strategy and research. Technical report, Javelin, 2016.

[2] D. Aldous. The continuum random tree. i. *The Annals of Probability*, 19(1):1–28, 1991.

[3] D. Aldous. The continuum random tree iii. *The Annals of Probability*, 21(1):248–289, 1993.

[4] F. Bagherzadeh-Khiabani, A. Ramezankhani, F. Azizi, F. Hadaegh, E. W. Steyerberg, and D. Khalili. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. *Journal of clinical epidemiology*, 71:76–85, March 2016.

[5] K. E. P. Baksai. *Feature Selection to Detect Patterns in Supervised and Semi Supervised Scenarios*. PhD thesis, Pontificia Universidad Católica de Chile, 2010.

[6] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland. Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3):602–613, February 2011.

[7] F. Cao, M. Estert, W. Qian, and A. Zhou. Density-based clustering over an evolving data stream with noise. In *Proc. of the 2006 SIAM International Conference on Data Mining (),*, pages 328–339. Society for Industrial and Applied Mathematics, April 2006.

[8] M. Carminati, R. Caron, F. Maggi, I. Epifani, and S. Zanero. Banksealer: A decision support system for online banking fraud analysis and investigation. *Computers and Security*, 53:175–186, September 2015.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 39(1):1–38, 1977.

[10] H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. volume 3644, pages 878–887, August 2005.

[11] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, April 1982.

[12] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[13] Z. He. Farthest-point heuristic based initialization methods for k-modes clustering. *arXiv*, cs/0610043, 2006.

[14] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley and Sons, 2013.

[15] Kount. Mobile payments fraud survey report. Technical report, Kount, 2016.

[16] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu. Feature selection: A data perspective. *arXiv*, 1601.07996, September 2016.

[17] A. Liaw, M. Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[18] T. Liu, F. Wang, and G. Agrawal. Stratified sampling for data mining on the deep web. *Frontiers of Computer Science*, 6(2):179–196, 2012.

[19] P. Matheswaran, E. Siva Sankari, and R. Rajesh. Fraud detection in credit card using datamining techniques. *International Journal for Research in Science Engineering and Technology*, 2(1):11–18, February 2015.

[20] F. N. Ogwueleka. Data mining application in credit card fraud detection system. *Journal of Engineering Science and Technology*, 6(3):311–322, 2011.

[21] S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar. Credit card fraud detection: A fusion approach using dempster–shafer theory and bayesian learning. *Information Fusion*, 10(4):354–363, October 2009.

[22] D. Pelleg, A. W. Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proc. of the 17th International Conference on Machine Learning (ICML'00), Stanford, California, USA*, volume 1, pages 727–734, June 2000.

[23] C. Phua, V. Lee, K. Smith, and R. Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv*, 1009.6119, 2010.

[24] D. M. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[25] J. R. Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[26] K. RamaKalyani and D. UmaDevi. Fraud detection of credit card payment system by genetic algorithm. *International Journal of Scientific & Engineering Research*, 3(7):1–6, January 2012.

[27] Y. G. S.ahin and E. Duman. Detecting credit card fraud by decision trees and support vector machines. In *Proc. of the 2011 International MultiConference of Engineering and Computer Scientists (IMEC'11), Hong Kong, China*, volume 1, March 2011.

[28] A. Shen, R. Tong, and Y. Deng. Application of classification models on credit card fraud detection. In *Proc. of the 2007 International Conference on Service Systems and Service Management, Chengdu, China*, pages 1–4. IEEE, June 2007.

[29] K. Sherly and R. Nedunchezhian. Boat adaptive credit card fraud detection system. In *Proc. of the 2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC'10), Coimbatore, India*, pages 1–7. IEEE, December 2010.

[30] A. Srivastava, A. Kundu, S. Sural, and A. Majumdar. Credit card fraud detection using hidden markov model. *IEEE Transactions on dependable and secure computing*, 5(1):37–48, January 2008.

[31] J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, June 1999.

[32] S.-J. Yen and Y.-S. Lee. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):5718–5727, April 2009.

[33] H. Zhang. The optimality of naive bayes. In *Proc. of the 17th International Florida Artificial Intelligence Research Society Conference, Miami, Florida, USA*, January 2004.

[34] Y. Zhang, F. You, and H. Liu. Behavior-based credit card fraud detecting model. In *Proc. of the 5th International Joint Conference on INC, IMS and IDC (NCM'09), Seoul, South Korea*, pages 855–858. IEEE, August 2009.

[35] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *Proc. of the 24th international conference on Machine learning (ICML'07), Corvalis, Oregon, USA*, pages 1151–1157. ACM, June 2007.

Author Biography

Dahee Choi received the B.S. in Information System from Hanyang University in 2016. Currently she is in Ph.D. degrees in Korea University. Also she is a member of Risk Management Laboratory. Her research interests include Machine Learning, Deep Learning and Data Mining. She is a member of CISSP associate.

Kyungho Lee received his Ph.D degree from Korea University. He is now a professor in the Graduate School of Information Security, Korea University, and has been leading the Risk Management Laboratory in Korea University since 2012. He was a former CISO at NHN Corporation, and CEO of SecuBase Corporation.