AUTO CLAIM FRAUD DETECTION USING MULTI CLASSIFIER SYSTEM

Luis Alexandre Rodrigues and Nizam Omar

Department of Electrical Engineering,
Mackenzie Presbiterian University, Brazil, São Paulo
71251911@mackenzie.br,nizam.omar@mackenzie.br

ABSTRACT

Through a cost matrix and a combination of classifiers, this work identifies the most economical model to perform the detection of suspected cases of fraud in a dataset of automobile claims. The experiments performed by this work show that working more deeply in sampled data in the training phase and test phase of each classifier is possible obtain a more economic model than other model presented in the literature.

KEYWORDS

Fraud Detection, Multi Classifier, Data Mining.

1. INTRODUCTION

The detection of suspected cases of fraud aims to find anomaly patterns in a given population, could be performed in manually or automatically [1]. This task has been applied in various fields like insurance [2], finance [3] and telecommunications [4], etc.

The algorithms used in Data Mining to classification tasks are usually based on heuristics, and thus there is an optimal classifier to perform classification tasks in large datasets [5].

Using a set of 100 samples of training data, this work performs the training and testing of classifiers, whose are applied in an automobile claims dataset that has suspected cases of fraud. After this classifier are combined in a parallel topology that use a combination of results by vote techniques to perform a final classification of each objects.

The classifiers created by this work are evaluated economically. [6] presents a cost matrix that to identified the savings generated by models used in detection of suspected cases of fraud. This cost matrix will used by this work to create a set of classifiers containing the most saving models of detection fraud.

The section 2 from this work presents some researches created to detection suspected cases of fraud. The section 3 presents a methodology used to create the most saving model to detect suspected cases of fraud in an automobile claims dataset. The section 4 presents the results obtained when the classifiers are applied individually and when applied by set of classifiers in the testing dataset to detect suspected cases of fraud.

Natarajan Meghanathan et al. (Eds) : ACSIT, FCST, ITCA, SE, ICITE, SIPM, CMIT - 2014
pp. 37–44, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4604

2. RELATED WORKED

The most common technique to fraud detection by Data Mining is find patterns that shows a behavior uncommon inside of dataset [7]. The Data Mining works with different data exploration models and solutions to specifics fraud cases were proposed [7]:

- Insurance: [6] used individual classifiers and multi classifier system to detect fraud in an automobile claim dataset. The individual classifiers are Decision Tree by C4.5, Naïve Bayes and Artificial Neural Network. The multi classifier system is a combination of Decision Tree, Naïve Bayes and Artificial Neural Network by Stacking-bagging algorithm. The results showed that multi classifier system was the most saving model to detect suspected cases of fraud.
- Credit Card: [8] presents three techniques used in credit card fraud detection, Artificial Neural Network, Logistic Regression and Decision Tree. According [1] the most techniques used in credit card fraud detection are Outliers Detection and Artificial Neural Network.
- Telecommunications: The works to fraud detection in telecommunication field focus on trying to identify the use of services without authorization by Artificial Neural Network, Outliers Visualization and patterns recognition [1].
- Online Auction: [9] presents a model to fraud detection to online auctions. The model used decision tree created by C4.5 algorithm to classifiers suspicious transactions according to the time that they occur. The criteria used to create the decision tree's rules are the average of positives feedbacks that vendors have and the price average of their products.
- Health Insurance: [7] presents fraud cases performed in medical clinics, which impair financially the insurance companies. The cases are detected by a model based on outlier detection by Support Vector Machine.

3. METHODS TO FRAUD DETECTION

This section will present a methodology that this work is using to find the most saving model to detect suspected cases of fraud. Will be presented the classifiers used in fraud detection, the topology and a combination function used to perform a final prediction each objects.

3.1. Classifiers

The classifiers aims to identify the categories set that a object of given dataset belongs [10]. This work selected three algorithms used in related works to perform the classification task in automobile claims dataset:

- *Decision Tree C4.5:* rule induction is one the most used methods used in fraud detection, because is easy to analyze the decisions created by the algorithm [11]. The algorithm C4.5 is used to induction decision tree. The decisions created by this algorithm are performed by the evaluation of dataset's features [12].
- *Naive Bayes:* naive bayes is a static classifier based on Bayes Theorem that mix previous knowledge a class by evidence selected in dataset [13]. The algorithm has a good performance history compared to other algorithms applied in fraud detection of automobile claims [14]. According [6] the algorithm is very efficient in large datasets and very efficient to create classifiers.
- *Support Vector Machines (SVM):* SVM is a binary classifier has been successfully applied in tasks to pattern recognition [15]. The algorithm maximizes the decision limit between two classes using a kernel function [10]. According [16] SVM is used in fraud detection tasks because works very well in datasets with imbalanced class.

The algorithms presented by this work are instable, because can change their forms when the environment and conditions in which they applied change. This feature is important when the combination of classifiers is performed, because each change in training dataset, different classifications will be performed in each new classification model created in the training phase [17].

3.2. Combination of Classifiers

The combination of classifier aims to perform classification tasks by combination of results between different classifiers to predict the final classification of the each object in the dataset [18].

This work combined the result of each algorithm previously presented to detect suspected cases of fraud by parallel topology. According [5] the most systems that used combination of classifiers used a parallel topology, which executes parallel all classifiers and combining their results using a decision function. The Figure 1 presents a structure proposed by this work using a parallel topology.

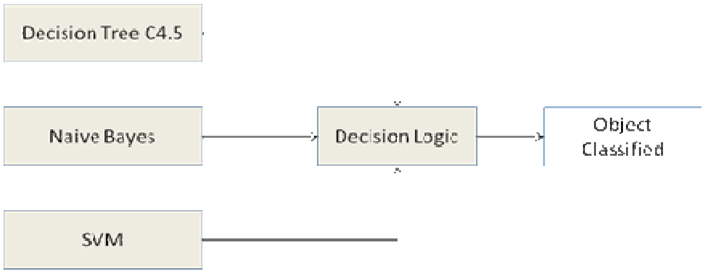

Figure 1. Structure of combination of classifiers using a parallel topology

Accord Figure 1 each object from dataset is applied trough all classifiers and each classifier will present a different classification to object, it can be fraud or legal. The decision function will be responsible for obtaining the classification object provide by each classifier and to perform the final classification to each object.

The decision function used to perform the final classification was the vote technique *AVGVote* [18]. As shown in equation 1, each classifier C_{ji} inside in set of classifiers R, the *AVGVote* function computes one vote in the object x classified as i.

$$AVGVote(x) = argmax_{i=1}^{c}\left(\frac{1}{R}\sum_{j=1}^{R} C_{ji}(x)\right) \qquad (1)$$

3.3. Automobile Claim Dataset

The experiments presents in this work used an automobile claims dataset with suspected cases of fraud. Each object from dataset is classified as fraud or legal. This dataset was used in [6] to identify the most saving model to detect suspected cases of fraud.

The dataset has suspected cases of frauds between 1994 and 1996 and has 15.421 objects, and each object has 6 numeric attributes and 25 categorical attributes. The preprocessing data was performed following the orientations proposed by [6].

The dataset was divided in two partitions to training and to testing the classifiers. The training partition has automobile claims between 1994 and 1995 years, and the testing partition has automobile claims of 1996 year.

There are imbalanced classes inside of dataset. This feature indicates that the classes are not distributed in the same quantity inside of dataset [10]. If dataset presents this feature, the generalization in each classifier can be adversely affected, thus classification tasks can be little precise in its test phase.

According [10], one way to solve the imbalanced classes' problem between the classes inside of dataset is the subsample generation from dataset. Thus this work created various subsamples and applied in the training and testing phase of each classifier.

3.4. Creating subsample data to training phase

According [10] the performance of a classifier depends of training data used in training phase. Thus with the goal of finding the best subsample to train algorithms, 100 subsamples were created. Between the first subsample and subsample number 71, there was a variation in the quantity of objects, and the balanced class was between 50% fraud objects and 50% legal objects. The variation of size of first subsample until the subsample 71 was on the order of 20 to randomly selected objects and no repetition of objects. The subsample number 1 was composed of 20 objects and the size of each new subsample was the sum of size of the previous subsample plus 20 objects. Thus the subsample number 71 was composed by 1420 objects.

Because the training dataset has imbalanced classes, the variation of quantity of objects between subsample number 72 until subsample number 100 was performed on the order 10 to 10 objects only to objects that belongs to majority class, and there was not repetition of objects. The subsample number 72 was composed by the sum of size of subsample plus the random selection of 10 objects of majority class. This variation happened until subsample number 100.

3.5. Cost Model

[6] used a cost matrix to identify the most saving model to perform detection suspected cases of fraud in the dataset. Based on the year 1996, the average cost per claim was about USD$ 2,540.00 and the average cost per investigation of suspected cases of fraud was about USD$ 203.00.

Using a confusion matrix [6] defined variables to identify the costs in each experiment performed in his work. According Table 1 the quantity of True Positives (Hits) and the quantity of False Negatives (False Alarm) were used to calculate the cost of suspected fraud claim. The quantity of items classifieds like True Negatives (Normal) and the quantity of items classified like False Negatives (Misses) were used to calculate the cost of each claim.

Table 1. Model Cost to Fraud Detection.

Variable	**Cost**
Hits	Quantity of Hits * Average cost per Investigation.
False Alarms	Quantity of False Alarms * (Average cost per Investigation + Average cost per claim)
Misses	Number of Misses * Average cost per claim
Normal	Quantity of Normal * Average cost per claim

The False Alarm items are the most expensive model, because this variable is defined by the cost per investigation and by the cost per claim. The saving total of each model created was defined in the Model Cost Savings variable by [6], as shown in equation 2.

$$Model\ Cost\ Savings = No\ Action - [Misses\ Cost + False\ Alarms\ Cost + Normals\ Cost + Hits\ Cost] \tag{2}$$

The variable No Action is considering that all claims are Normal. Thus this variable is defined by quantity claims in dataset multiplied by cost per claim. This work used the variable Model Cost Savings to identify the best model created by each algorithm in each testing phase. This variable was used too to compare the cost of combination of classifiers related the classifiers applied manually and related the results presented in [6].

4. EXPERIMENTS

This work performed four experiments to detect suspected cases of fraud in the automobile claim dataset. Three experiments that are divided by algorithm used the set of subsample to find the most saving model, and the last one is related the combination of classifiers created by each algorithm.

The Figure 2 show the process used to create the classification model proposed by this work. In the first moment is performed a preprocessing data according [6]. The preprocessing data was necessary to eliminate missing values and create new attributes that can grow the performance of each classifier created in the training and testing phase.

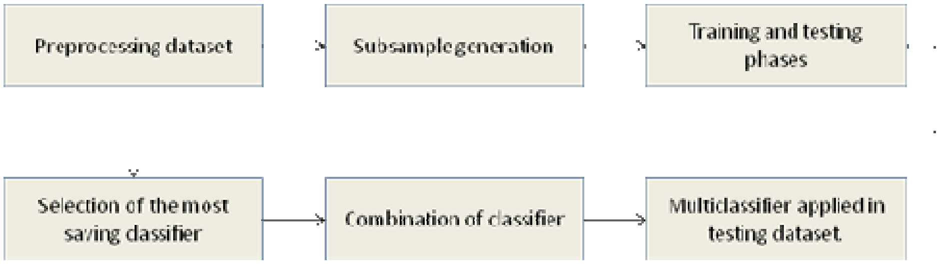
Preprocessing dataset
Subsample generation
Training and testing phases
Selection of the most saving classifier
Combination of classifier
Multiclassifier applied in testing dataset

Figure 2. Process to perform fraud detection

When the classifiers were applied in testing dataset a confusion matrix was extracted to calculate all cost variables. It was possible calculate the value of Model Cost Saving variable, as shown in equation 2, and get the most saving model created by each classifier. Each model selected as the most saving was compare with the most saving model shown in [6]. This model is composed by combination of classifiers created by C4.5, Naïve Bayes and Artificial Neural Network, and the max saving cost by this combination was about USD$ 167,000.00.

According Figure 3 the C4.5 algorithm showed the most saving model to detect suspected cases of fraud when it was applied in the subsample number 12. The subsample number 12 consists of 240 objects, with 50% of objects classified as fraud and 50% of objects classified as legal. Reaching a saving of USD$ 177,592.00, the model is the most saving when compared with other classifiers and the most saving when compared the model proposed in [6].

The SVM algorithm created the most saving fraud detection model when it was applied in the subsample number 80. The subsample consists of 1510 objects, with 40% of objects classified as fraud and 53% of objects classified as legal. The most saving model created by SVM has a saving about 158,732.00, but according Figure 4, the model is not more economic than model proposed in [6].

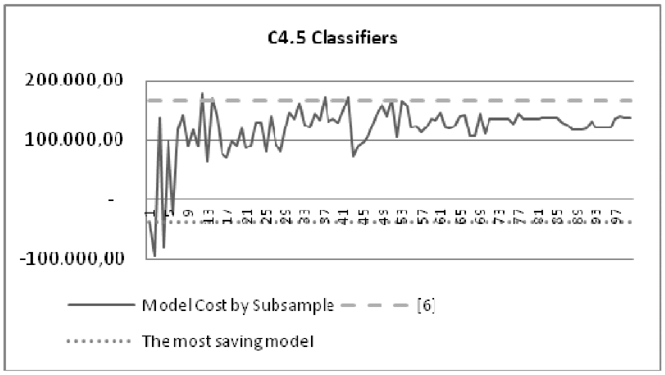

Figure 3. Performance of C4.5 classifier in each subsample

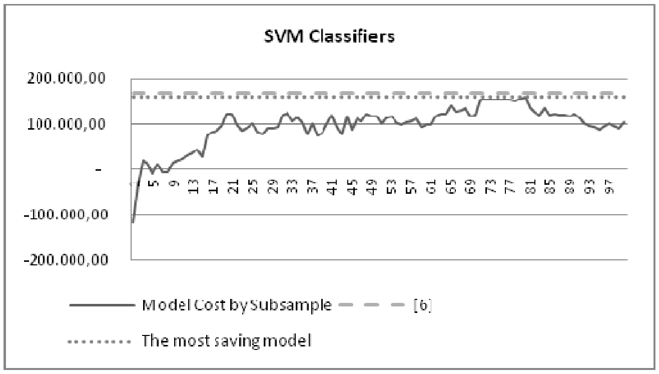

Figure 4. Performance of SVM classifier in each subsample

The most saving model using Naïve Bayes algorithm was created by subsample number 20. This subsample consists of 400 objects, with 50% of objects classifieds as fraud and 50% of objects classifieds as legal. This model has the worst saving compared with other classifiers proposed by this work and the worst performance compared with the most saving model proposed in [6]. According Figure 5, the model presented a saving of USD$ 117,486.00.

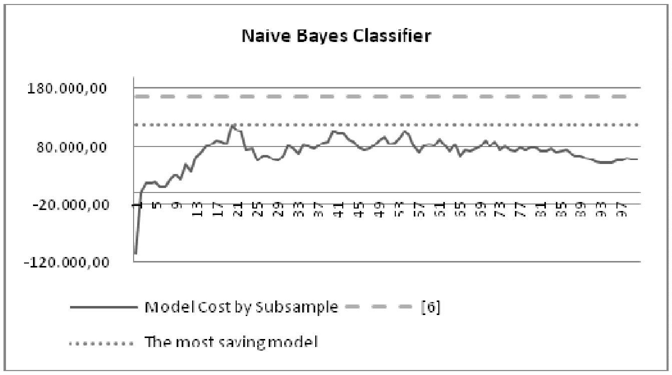

Figure 5. Performance of Naïve Bayes classifier in each subsample

The three models created when applied in testing dataset showed different classification of objects. As shown in Figure 6 there is a diversity of quantity positive class classified by each the most saving model. According [5] the diversity of results is important to make the combination of classifiers and created a final prediction of each object in dataset.

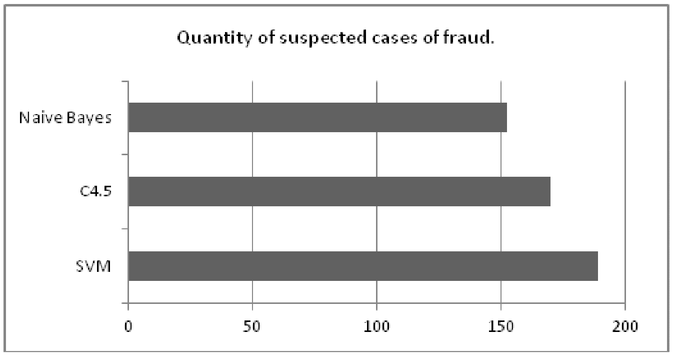

Figure 6. Quantity of suspected cases of fraud identified by each the most saving model

Using this diversity presented by the most saving models, they were combining and applied in testing dataset. The combination of classifiers proposed by this work, which uses the parallel topology and *AVGVote* decision function, presented the most saving model compared to all models applied individually and compared to model proposed in [6]. According Figure 7 the combination of classifiers reaching a saving about USD$ 183,089.00.

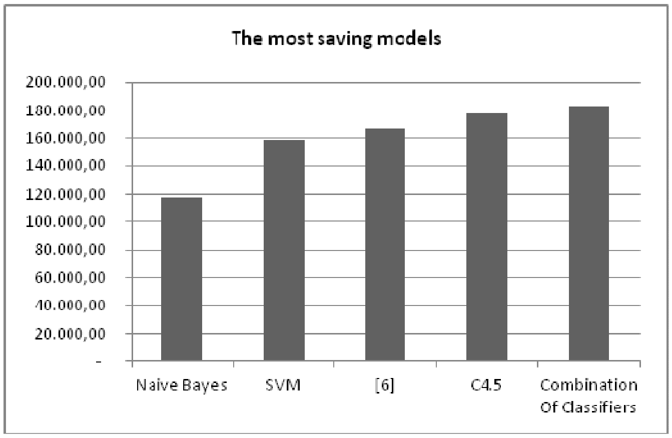

Figure 7. Ranking of the most saving model

5. CONCLUSION

This work combined the classifiers create by C4.5, SVM and Naïve Bayes algorithm to find the most saving model to detect suspected cases of fraud. Working more deep with subsample of training automobile claim dataset, the most saving models were selected, combined and applied in testing dataset. The combination of classifiers was performed by parallel topology and each object was classified by *AVGVote* function decision.

These experiments showed that a good subsample can be efficient to build classifiers and to build a cheaper model to identify suspected cases of fraud. The combination these classifiers presented better performance than the most saving model proposed in [6], which used combination of classifiers.

REFERENCES

[1] Y. Kou, C.-t. Lu, S. Sinvongwattana & Y.-P. Huang, (2004) "Survey of Fraud Detection Techniques," *International Conference on Networking, Sensing & Control*.

[2] K. D. Aral, H. A. Güvenir, I. Sabuncuoğlu & A. R. Akar, (2012) "A prescription fraud detection model," *Computer methods and programs in biomedicine*, Vol. 106, pp37-46.

[3] E. Ngai, Y. Hu, Y. Wong, Y. Chen & X. Sun, (2011) "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, Vol. 50, pp559-569.

[4] C. S. Hilas & P. A. Mastorocostas, (2008) "An application of supervised and unsupervised learning approaches to telecommunications fraud detection," *Knowledge-Based Systems*, Vol. 21, pp721-726.

[5] M. Woźniak, M. Grañab & E. Corchadoc, (2014) "A survey of multiple classifier systems as hybrid systems," *Information Fusion*, Vol. 16, pp3-17.

[6] C. Phua, D. Alahakoon & V. Lee, (2004) "Minority Report in Fraud Detection : Classification of Skewed Data," *ACM SIGKDD EXPLORATIONS*, pp50-59.

[7] M. Kirlidog & C. Asuk, (2012) "A Fraud Detection Approach with Data Mining in Health Insurance," *Procedia - Social and Behavioral Sciences*, pp989-994.

[8] A. Shen, R. Tong & Y. Deng, (2007) "Application of Classification Models on Credit Card Fraud Detection," *Service Systems and Service Management*, pp2-5.

[9] W.-H. Chang & J.-S. Chang, (2012) "An effective early fraud detection method for online auctions," *Electronic Commerce Research and Applications*, pp346-360.

[10] P.-N. Tan, M. Steinbach & V. Kumar, (2005) *Introduction to Data Mining*, Addison-Wesley.

[11] Y. Sahin, S. Bulkan & E. Duman, (2013) "A cost-sensitive decision tree approach for fraud detection," *Expert Systems with Applications*, Vol. 40, n. 15, pp5916–5923.

[12] J. R. Quinlan, (1993) *C4.5: programs for machine learning*, San Francisco: Morgan Kaufmann Publishers Inc.

[13] G. H. John & P. Langley, (1995) "Estimating continuous distributions in Bayesian classifiers," *UAI'95 Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp338-345.

[14] S. Viaene, R. A. Derrig, B. Baesens & G. Dedene, (2002) "A comparison of State-of-the-Art Classification Techiniques for Expert Automobile Insurance Claim Fraud Detection," *The Journal of Risk and Insurance*, vol. 69, n. 3, pp373-421.

[15] H.-C. Kim, S. Pang, H.-M. Je, D. Kim & S. Y. Bang, (2003) "Constructing support vector machine ensemble," *Pattern Recognition*, Vol. 36, n. 12, pp2757–2767.

[16] S. Bhattacharyyaa, S. Jhab, K. Tharakunnelc & J. C. Westlandd, (2011) "Data mining for credit card fraud: A comparative study," *Decision Support Systems*, Vol. 50, n. 3, pp602–613.

[17] T. G. Dietterich, (2000) "Ensemble Methods in Machine Learning," *Multiple Classifier Systems*, Springer Berlin Heidelberg, pp1-15.

[18] R. Ranawana & V. Palade, (2006) "Multi-Classifier Systems: Review and a roadmap for developers," *International Journal of Hybrid Intelligent Systems*, Vol. 3, n. 1, pp35-61.

[19] S. Thiruvadi & S. C. Patel, (2011) "Survey of data-mining used in Fraud detection and Prevention," *Information Technology Journal*, pp710-716.

AUTHORS

Luis Alexandre Rodrigues

Degree in Information Systems and MSc student in Electrical Engineering at Mackenzie University. Currently he is working like software architect at Insurance Company and is interested in Data Mining techniques to detect suspected cases of fraud in large datasets.

Nizam Omar

Degree in Mechanical Engineering from the Technological Institute of Aeronautics (ITA), MSc in Applied Mathematics from the ITA and Ph.D. in Applied Informatics by Pontifical Catholic University (PUC). He is currently professor at Mackenzie Presbiterian University. Has experience in Computer Science, Artificial Intelligence, Automata and Formal Languages.