Indian Journal of Science and Technology, Vol 8(35), DOI: 10.17485/ijst/2015/v8i35/87306, December 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Novel Unsupervised Classification Method for Customs Fraud Detection

Habibollah Arasteh Rad*, Saeed Arash, Farhad Rahbar, Ruhollah Rahmani, Zainabolhoda Heshmati and Maysam Mirzaee Fard

Department of Network and Information Technology, Faculty of Science and Modern Technology, Institute of Applied Intelligent Systems, University of Tehran, Tehran, Iran; Habib.Arasteh@ut.ac.ir, Saeed.arash@ut.ac.ir, frahbar@ut.ac.ir, R.rahmani@ut.ac.ir, zheshmati@ut.ac.ir, m.mirzaie@ut.ac.ir

Abstract

Objectives: In this paper, a very light and straightforward algorithm is proposed for customs fraud detection. **Methods/ Analysis:** in order to fraud detection we have proposed our algorithm based on unsupervised methods. Our approach is a combination of data clustering methods, Mahalanobis distance classifier, K Nearest Neighbor (KNN) method, and density-based methods**. Findings:** The results showed that the proposed method was able to accurately identify frauds, as more than 73 percent of high-risk goods that the proposed method is detected, has been violated. It is faster and more rapid than the other methods. The method requires less processing than other methods, and more than 30 percent CPU usage has been improved. The approach is independent of distribution and scattering of data samples. It also has the ability to work with samples by different clusters, densities, and no limitation on dimension of data. **Novelty of the Study:** For the first time, an unsupervised method is used for finding the frauds in customs. **Application/Improvements:** One of the most important applications of the results of this study is the Customs Risk Management System. Also, the proposed approach will enhance the ability of fraud detection in trade.

Keywords: E-Customs, Fraud Detection, Risk Management System, Unsupervised Method

1. Introduction

Numerous fraudulent acts related to customs including illegal cargo, hiding goods, declaring less or making false report[1]. On the one hand, because of the huge commodity volume and the time limit of trade activities, customs authorities only have the ability of inspecting 10% of all commodities. On the other hand, only 1% of commodities are detected as fraud in all inspected commodities.

In this study, a novel outlier detection algorithm is introduced for customs fraud detection. The algorithm is implemented and tested on Iranian customs inspection data. The Iranian Customs Organization uses Customs Intelligent System v.5 software to administer their interior customs procedures, commodities flow and goods audit.

The rest of this paper is organized as follows: Section 2 presents the related work. Section 3 describes our approach to the problems of identifying suspicious customs operations or fraudulent commodities. Section 4 presents the evaluation results. Finally, Section 5 presents a conclusion to this work and the identified for future works.

The rapid development of commerce and the increasing business connections between countries has complicated the customs enforcement due to limited resources of the customs officers especially when customs audit is based on the expertise. In the recent years, the advances of data mining and statistical approaches are becoming more popular within each day and create new aspects of services specifically for safe commerce. This approaches attempt to construct computational solutions to identify these fraudulent operations automatically. With using some solutions were devised in terms of the approaches, the amount of operations required in customs officer investigation for trade verification will

*Author for correspondence

be reduced. It is obvious that the method be caused of identification of dubious operations automatically or semi-automatically.

Fraud detection methods have widely applied in great importance domains, especially financial fields[2]. For instance, in[3] proposed a credit card fraud detection model. The model combined confidence value, neural network algorithm and receiver operating characteristic. Wen-Fang et al. introduced a fraud detection model for suspicious credit card by applying outlier mining. The detection method is distance-based for credit card transaction data according to the non-conformism and infrequency of fraud[3]. As another example, a classification model is developed based on Artificial Neural Networks (ANN) in Vallarino studies. The model applied for fraud detection on credit suspicious card[4].

Fraud detection systems are also widely used in telecommunications. For example, a study done at Umm Al-Qura University. It explored the role of artificial neural networks in prevention of telecommunication fraud in detail. The study has shown that Artificial Neural Network, due to inherent ability to adapt, speed and efficiency, can be superior method for telecommunications fraud detection[5]. In[7] performed machine learning techniques to telecommunications fraud detection. In this study, fraudulent behaviors for subscriptions are detected by neural network. The Back-Propagation Neural Network (BPNN) is performed for telecommunication interpolation by[8]. It is observed that the performance of BPNN in predicting fraud was acceptable.

Several attempts have been made to detect terrorist networks. Brown proposed a method based on k-means and the nearest neighbor approach[6]. The spatial clustering methods are often employed in "hotspot analysis"[7,8].

Overall, there are various techniques and methods to take care of fraud detection. The use of rule based systems[9,10], neural networks[11], expert systems[12], the detection of statistical outliers[13,14] and Bayesian networks[15] are more emphasized. These methods can be separated in three groups; supervised, semi-supervised and unsupervised.

2. Supervised Methods

In supervised methods, records of both fraudulent and non-fraudulent are used to construct models which yield a suspicion score for new cases and allow one to assign new observations into one of the two classes[16]. This approach requires one to be confident about the true classes of the original data used to build the models. Furthermore, it can only be used to detect frauds of a type which has previously occurred[17]. Neural networks[18,19], Decision trees, rule induction, case-based reasoning are popular and Support Vector Machines (SVMs) are all popular methods which have been previously used. Rule based methods are also supervised learning algorithms that produce classifiers using rules such as Bayes[20] and RIPPER[21].

2.1 Semi-Supervised Methods

In the semi-supervised approach, the input contains both unlabeled and labeled data. Semi-supervised method deals with a small amount of labeled data with a large pool of unlabeled data. In many situations, assigning classes is expensive because it requires human insight. In these cases, it would be enormously attractive to be able to leverage a large pool of unlabeled data to obtain excellent performance from just a few labeled examples. The unlabeled data can help you learn the classes. For example, a simple idea to improve classification by unlabeled data is Using Naïve Bayes to learn classes from a small labeled dataset, and then extend it to a large unlabeled dataset using the EM (Expectation–Maximization) iterative clustering algorithm. In this approach, initially a classifier is trained using the labeled data. Then, it is applied to the unlabeled data for labeling them with class probabilities (the expectation step). Then, a new classifier is trained using the labels for all the data (the maximization step). Finally, these steps are iteratively repeated until convergence is achieved[22].

2.2 Unsupervised Methods

In unsupervised methods, there are no prior class labels of genuine or fraudulent observation. Techniques employed here are usually a combination of profiling and outlier detection methods. Initially a baseline distribution is modeled that represents normal behavior and then it attempts to detect observations that show greatest departure from this normal. Also, unsupervised approaches such as outlier detection, spike detection, and other forms of scoring have also been applied[23].

As in the unsupervised case, our observations (data) are unlabeled. We are therefore faced with an unsupervised fraud detection scenario. One of the main categories of unsupervised learning methods for fraud detection is outlier detection. Outlier detection approaches in a

general are divided into five categories. These approaches are explained as follows:

2.2.1 Statistical Test

These approaches established by assuming that all data follow a certain kind of statistical distribution (e.g., Gaussian), then the parameters will be calculated with considering to chosen statistical distribution (e.g., mean and standard deviation). The normal data occur in the high probability region of the model known such as normal data and follow the known distribution. The outlier data have a low probability to be generated by the preferred distribution, for instance, the data more than 3 times of standard deviation recognize such as outlier data.

2.2.2 Depth-based Approaches

The border of the data space, independent of the statistical distributions, will be searched for outliers by the approach. The data objects organized in convex hull layers. The normal objects supposed that the center of the data space and outliers are placed at the border of the data space. Convex hull computation have high complexity and only efficient in 2D/3D spaces.

2.2.3 Deviation-based Approaches

According to this approach, the points are not fit to the general characteristics of data set, recognized such outliers. For example when removing the outliers, the variance of the data set will be minimized. However the approach idea is similar to statistical approaches but they are independent from the chosen kind of distribution.

2.2.4 Distance-based Approaches

These approaches decisions are based on the distance(s) of a point to its neighbors. These approaches suppose that the outliers are far apart from their neighbors and normal data objects have a dense neighborhood. DB-Outliers and Outlier scoring based on KNN distances are two popular method of this kind[24]. Distance-based outlier detection models have problems with different densities.

2.2.5 Density-based Approaches

These approaches compare the density around a point with the density around its local neighbors. The relative density of a point compared to its neighbors is computed as an outlier score. Approaches also differ in how to estimate density. They suppose that the density around a normal data object is similar to the density around its neighbors and the density around an outlier is considerably different to the density around its neighbors. Local Outlier Factor (LOF)[25,26], Connectivity-based Outlier Factor (COF)[27], Influenced Out Liernes (INFLO)[28] and Local Outlier Correlation Integral (LOCI) are some methods of this kind of classification. LOF would not be appropriate while clusters with different densities are not clearly separated. Connectivity-based Outlier Factor (COF) is to treat "low density" and "isolation" differently. Influenced Out Liernes (INFLO) attempt to solve LOF problem, it is simply measured by the inverse of the KNN distance. Local Outlier Correlation Integral (LOCI) idea is similar to LOF, but taking the neighborhood instead of kNNs as reference set and testing multiple resolutions (here called "granularities") of the reference set to get rid of any input parameter.

3. Methodology

Our proposed approach for fraud detection is an unsupervised method. Our approach is a combination of data clustering methods, mahalanob is distance classifier, K Nearest Neighbor (KNN) method, and density-based methods. Since we may have samples of different kinds and ranges, we first divide the samples into different clusters. Then assign our input to one of clusters according to the mean and standard deviation of each of the clusters. Finally, compute the outlier score of the given input in the assigned cluster using a density-based outlier detection approach. Our proposed algorithm steps are as follow:

3.1 Data Clustering

Since our data samples are from different types, initially they should be separated into different groups via a clustering algorithm. The clustering algorithm that we have used in this problem is X-means clustering algorithm. This algorithm is an extended form of the classical K-means clustering algorithm with the difference that it does not need the number of clusters to be specified exactly. It takes a range for K. the algorithm starts with K equal to the lower bound of the given range and continues to add centroids where needed until the upper bound is reached. During this process, the centroid set that achieves the best score is recorded, and this is the one that is finally used.

3.2 Input Cluster Labeling

In this step we assign our given input to one of the clusters built in step 1 using mahalanobis distance classifier. According to this method the distance of a given point to a set of samples is computed as:

$$dist(p) = \sqrt{\frac{(p - mean(samples))}{std(samples)}}$$

The distance of the given input to all of the clusters is calculated and it is then assigned to the cluster with smallest distance.

3.3 Find K Nearest Neighbor

In this step, we should consider the given input's outlierness (unusualness) in the assigned cluster. For considering outlierness of an observation with respect to the data in the assigned cluster, we use a density-based approach. This approach compares the density of observations in neighborhood of a given observation with the densities of all observations and takes this as a good score for finding outlierness of the input. For this purpose, we find K-nearest observation to the given input according to its Euclidean distance from it as K-Nearest Neighbor. The parameter K is set as

$$k = \sqrt{n}$$

3.4 Outlier Score Calculation

We set the density of data in the neighborhood of the given input as the standard deviation of them. The density of all observations in the cluster is also the standard deviation of them. By dividing these two densities with each other, a number is obtained that is outlier score of a given input. There is a direct relation between this number and input abnormality or probability of fraud. The larger number show the greater probability that the given input is unusual or fraudulent.

$$unusuality = \frac{\sigma k}{\sigma}$$

3.5 Fraud Detection

By setting a threshold on the outlier score, we can label the input data with fraudulent or normal i.e., if the outlier score value of an input observation is bigger than the threshold, we assign it to the fraudulent category.

In comparison to the previous approaches, our proposed approach has some advantages, such as: Simple and straightforward algorithm, no assumption on type and independent of distribution and scattering of data samples. It also has the ability to work with samples by different clusters and densities, small run time and no limitation on dimension of data[33,34].

4. Evaluation Results

The dataset used here has been taken from real customs data, which is shown in Figure 1. This figure includes multiple scattering that helps us test different conditions. The horizontal axis is the number of the sample and the vertical axis is the value of the sample. In the following figures, the results of the proposed algorithm and computed outlier scores for different conditions are shown. The threshold value for the outlier score is set equal to 0.95, so inputs that have scores larger than 0.95 are labeled as fraudulent. This is a logical setting as it matches most statistical distribution thresholds as well.

Figure 1 shows the three clusters that were created using the X-means clustering algorithm. Following the clustering step, various input values were tested in the algorithm and the outlier scores were calculated. The results are shown in the Figure 2. As can be seen from the Figure 2, outlier scores above the threshold value, have a higher probability of being fraudulent and should

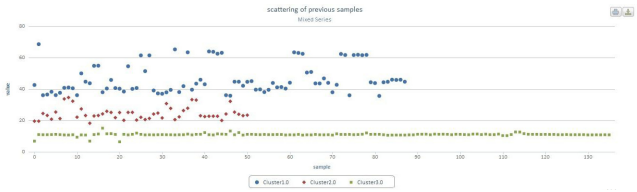

Figure 1. Scattering of previous samples.

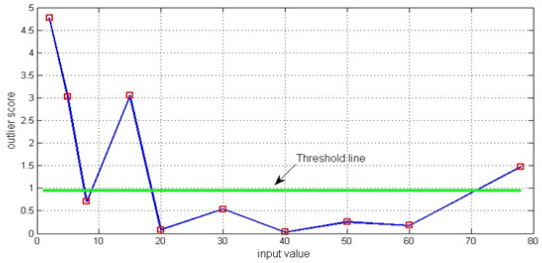

Figure 2. Outlier score VS input value.

hence be inspected further. These results were fed back to the customs operations and it was verified that the detected fraudulent observations were in fact illegitimate commodities. Hence the classification resulting from this approach is indeed very accurate.

In order to show superiority of our proposed method over previous outlier detection algorithms, we compare the results with other approaches under various input conditions; removing the first step of our algorithm and testing previous methods without the initial clustering clearly does not yield the same accurate outlier detection.

4.1 Statistical Approaches

If we assume that our data has a normal distribution, then input data towards the two tails of the distribution should show high probability of fraud. Hence, input values such as 10 and 60 should be detected as outliers, yet it is clear that these inputs are within the allowed regions with suitable densities.

4.2 Depth-based Approaches

This approach only detects inputs on the outer layers as fraudulent. For example, in the sample data used here, input values between 12 and 18, considering the distributions given, have a high chance of being fraudulent, yet depth-based approaches do not reflect this. On the other hand, input value 10 has a higher fraudulent probability with these approaches, yet it is a normal input.

4.3 Deviation-based Approaches

This approach, similar to the previous approaches mentioned above, only has the ability to detect inputs in the far tails of the distribution as fraudulent and fails to pick up instances that occur within the range of all samples.

4.4 Distance-based Approaches

This approach fails when the data does not have a uniform distribution and the densities are different in different parts of the dataset. As can be seen in Figure 1 the densities of the observations used here are variety in different parts of the dataset, which indicates that the distance-based approach will not be accurate.

4.5 Density-based Approaches

As mentioned, these approaches calculated the density around a point and compare with the density around its local. How define density have influences on the effectiveness of these approaches. For different definitions, the performance of these approaches has certain weaknesses. Our proposed method is categorized under this group and shows an improvement in comparison with the previous approaches.

Table 1. Output labels for different inputs and approaches

Method\input	78	60	50	40	30	20	15	8	5	2
Statistical test	yes	no	no	No	no	yes	yes	No	no	yes
Deviation based	yes	no	no	No	no	no	yes	yes	yes	yes
Depth based	yes	no	no	No	no	no	no	No	No	yes
Distance based(DB(ε, π)-Outliers)	yes	no	no	No	yes	no	yes	yes	yes	yes
Density based	yes	no	yes	No	yes	no	yes	Yes	yes	yes
Proposed approach	yes	no	no	No	no	no	yes	No	yes	yes

In Table 1 result of algorithm for different inputs in various conditions is shown. According to this table, we observe that other methods have some incorrect output labels, whereas our proposed approach has correct output for all inputs. The results showed that the proposed method was able to accurately identify frauds, as more than 73 percent of high-risk goods that the proposed method is known, has been violated. The results showed that the proposed method requires less processing than other methods, and more than 30 percent CPU usage has been improved.

5. Conclusions and Future Work

In this study, a novel outlier detection algorithm is presented for customs fraud detection. The proposed algorithm is very simple and straightforward and unlike other previous outlier detection methods, places no assumptions and limitations on the scattering and distribution of the data. The results show that whereas other approaches have disadvantages in some given conditions, the proposed method yields suitable results in comparison with other methods.

As for future works, it is worth testing this algorithm on other fraud detection applications. Also, as there are no publicly available data sets for studying fraud detection, and obtaining real data from companies for research purposes is extremely hard due to legal and competitive

reasons, it is suggested to investigate this algorithm on synthetic data which matches closely to actual data. Work can also be done on improving the accuracy of this algorithm based on the application via enhanced clustering techniques.

6. References

1. Shao H, Zhao H, Chang G. Applying data mining to detect fraud behavior in customs declaration. Machine Learning and Cybernetic. 2002; 3:1241–44.
2. Prakash A, Chandrasekar C. An optimized multiple semi-hidden markov model for credit card fraud detection. Indian Journal of Science and Technology. 2015; 8(2):165–71.
3. Guo T, Li G. Neural data mining for credit card fraud detection. International Conference on Machine Learning and Cybernetics; Kunming. 2008. p. 3630–4.
4. Yu W, Wang N. Research on credit card fraud detection model based on distance sum. International Joint Conference on Artificial Intelligence (JCAI '09); Hainan Island. 2009. p. 353–6.
5. Sahin Y, Duman E. Detecting credit card fraud by ann and logistic regression. International Symposium on Innovations in Intelligent Systems and Applications (INISTA); Istanbul. 2011 Jun. p. 315–9.
6. Akhter MI, Ahamad MG. Detecting telecommunication fraud using neural networks through data mining. International Journal of Scientific and Engineering Research. 2012; 3(3):601–6.
7. Qayyum S, Mansoor S, Khalid A. Fraudulent call detection for mobile networks. International Conference on Information and Emerging Technologies (ICIET); Karachi. 2010 Jun. p. 1–5.
8. Mohamed A, Bandi AFM, Tamrin AR. Telecommunication fraud prediction using backpropagation neural network. International Conference of Soft Computing and Pattern Recognition; SOCPAR '09; Malacca. 2009 Dec. p. 259–65.
9. Liu H, Brown D. Criminal incident prediction using a point-pattern-based density model. International Journal of Forecasting. 2003; 19(4):603–22.
10. Oatley G, Brian E. Crimes analysis software: Pins in maps, clustering and bayes net prediction. Expert Systems with Applications. 2006; 25(4):569–88.
11. Bowers KJ, Johnson SD, Pease K. Prospective hot-spotting: The future of crimemapping? British Journal of Criminology. 2004; 44(5):641–58.
12. Benmakrouha F, Hespel C. An algorithm for rule selection on fuzzy rule-based systems applied to the treatment of diabetics and detection of fraud in electronic payment. IEEE International Conference on Fuzzy Systems (FUZZ); Barcelona. 2010 Jul. p. 1–5.
13. Costa PD, Mielke IT. A model-driven approach to situations: Situation modeling and rule-based situation detection. IEEE 16th International Enterprise Distributed Object Computing Conference (EDOC); Beijing. 2012 Sep. p. 154–63.
14. Xu W, Wang S, Zhang D. Random rough subspace based neural network ensemble for insurance fraud detection. 4th International Joint Conference on Computational Sciences and Optimization (CSO); Yunnan. 2011Apr. p. 1276–80.
15. Buoni A. Fraud detection: From basic techniques to a multi-agent approach. International Conference on Management and Service Science (MASS); Wuhan. 2010 Aug. p. 1–4.
16. Roman NT, Constantino ER, Ribeiro H, Filho JJ, Lanna A, Goldenstein SK, Wainer J. A decision support system for customs. Proceedings of ECML PKDD Workshop on Practical Data Mining: Applications, Experiences and Challenges; 2006 Sep. p. 100–3.
17. Martinez R, Cebrian M, Camacho D. Contextual information retrieval based on algorithmic information theory and statistical outlier detection. IEEE Information Theory Workshop. ITW '08; Porto. 2008 May. p. 292–7.
18. Moer V, Barbe W, Rolain KY. An automatic, statistical-based detection of outliers in an inter-laboratory comparison of nonlinear measurements. 38th European Microwave Conference, (EuMC); Amsterdam; 2008. p. 27–31.
19. Zhou L, Liu W, Chen H, Wang L, Yang X. Bayesian network-based detection and prediction of outliers in subspace. 7th World Congress on Intelligent Control and Automation. (WCICA 2008); Chongqing. 2008 Jun. p. 2479–85
20. Bolton RJ, Hand DJ. Statistical fraud detection: A review. Statistical Science. 2002; 17(3):235–49.
21. Bonchi F, Giannotti F, Mainetto G, Pedreschi D. A classification-based methodology for planning auditing strategies in fraud detection. ACM Knowledge Discovery and Data Mining (SIGKDD); 1999. p. 175–84.
22. Dorronsoro J, Ginel F, Sanchez C, Cruz C. Neural fraud detection in credit card operations. IEEE Transactions on Neural Networks. 1997; 8(4):827–34.
23. Lin J, Hwang M, Becker JA. Fuzzy neural network for assessing the risk of fraudulent financial reporting. Managerial Auditing Journal. 2003; 18(8):657–65.
24. Clark P, Niblett T. The cn2 induction algorithm. Machine learning. Machine Learning Journal. 1989; 3(4):261–85.
25. Cohen W. Fast effective rule induction. 12th Int'l Conf of Machine Learning; Lake Tahoe, California: Morgan Kaufmann. 1995. p. 115–23.
26. Kim J, Ong A, Overill R. Design of an artificial immune system as a novel anomaly detector for combating financial fraud in retail sector. Congress on Evolutionary Computation. 2003; 1:405–12.

27. Phua C, Gayler R, Lee,V, Smith-Miles K. A comprehensive survey of data mining based fraud detection research. Artificial Intelligence Review. 2010 Sep; 336–48.
28. Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. ACM on Management of Data (SIG-MOD). 2012; 29(2):427–38.
29. Breunig MM, Kriegel HP, Ng RT, Sander J. Lof: Identifying density-based local outliers. ACM SIGMOD Int Conference on Management of Data (SIGMOD), ACM; New York, NY, USA. 2000. p. 93–104.
30. Breunig MM, Kriegel HP, Ng RT, Sander J. Optics-of: Identifying local outliers. Springer-Principles of Data Mining and Knowledge Discovery (PKDD). 1999; 1704(1):262–70.
31. Tang J, Chen Z, Fu AWC, Cheung DW. Enhancing effectiveness of outlier detections for low density patterns. Pacific-Asia Conf on Knowledge Discovery and Data Mining (PAKDD). 2002; 23(36):535–48.
32. Jin W, Tung A, Han J, Wang W. Ranking outliers using symmetric neighborhood relationship. Knowledge Discovery and Data Mining (PAKDD). 2006; 3918(3):577–93.
33. Kasaeipoor A, Ghasemi B, SM Aminossadati. Convection of Cu-water nanofluid in a vented T-shaped cavity in the presence of magnetic field. International Journal of Thermal Sciences. 2015; 94:50–60.
34. Rad MZ, Kasaeipoor A. A numerical study of similarity solution for mixed-convection copper-water Nanofluid boundary layer flow over a horizontal plate. Modares Mechanical Energineering. 2015:14(14):190–8. (In Persian).