Frontiers in Signal Processing, Vol. 4, No. 1, January 2020
https://dx.doi.org/10.22606/fsp.2020.41007

Anomaly Detection of Vehicle Data Based on LOF Algorithm

Mengjia Yang[#], Daji Ergu[*]

College of Electrical & Information Engineering, Southwest Minzu University, Chengdu 610041, China
Email: [#] 945712922@qq.com, [*] ergudaji@163.com

Abstract. The official vehicle audit is an important issue in the government management, and it is very difficult to detect the potential doubts data in the collected data of official vehicles since most of them are unlabeled data. In this paper, a combination of DBSACN and Local Outlier Factor (LOF) algorithm is proposed for the official vehicle anomaly behavior detection by detecting the abnormal use data of the official vehicle under the same conditions. The detected data is regarded as a doubt data and submitted to the audit department for verification. Since the discrete features of the data set are too much and could not conform to the input type of the algorithm, the features are coded by One-Hot encoding, and a series of operations such as data cleaning and feature calculation are performed, and then compared with DBSCAN, LOF, and isolation forest anomaly detection algorithms. The experimental results show that the proposed algorithm outperforms the isolation forest, LOF and other machine learning algorithms in the anomaly detection of unmarked official vehicle data.

Keywords: DBSCAN, LOF, anomaly detection, audit of official vehicles

1 Introduction

In recent years, China has vigorously promoted official vehicles reform, strengthened big data audit of official vehicles, made clear the purchase, allocation, management and operation of official vehicles, strengthened financial fund management, reduced expenditure and improved financial fund efficiency[1]. Faced with the rapid growth of data volume in various industries, the experience of auditors often lags behind the development of audit data, making many new audit doubts impossible to be discovered in time. Data mining technology[2] solves this problem, it can quickly get valuable data and improve the efficiency of audit work.

The abnormal point detection algorithm can be applied to the auditing business. On the basis of selecting appropriate auditing indicators and preprocessing methods, a small amount of data with doubtful points can be filtered out from different forms of data, helping auditors to quickly and accurately locate auditing priorities[3].This paper needs the outlier detection algorithm to check the bus usage fee of each unit, and finds the abnormal point different from the normal data. It is an important and meaningful research work, and it is also a hot spot in the research field[4].

2 Related Work

In this section, we briefly introduce the literature on anomaly detection in recent years, as well as the application status of data mining technology in the field of auditing.

Related literature on abnormal point detection. In terms of outlier algorithm, the methods based on statistics, distance, deviation and density are also proposed successively. In the industrial situation, only a very small amount of marked abnormal data can be obtained, making it difficult to carry out abnormal detection. Nan Wang proposed a vertex weighted hypergraph learning method for abnormal detection[5]. J. Yang has designed an ICPSs anomaly detection method based on region division, which is used for attack detection of industrial network physical system[6]. Yingfu Huang proposed an algorithm combining K nearest neighbor (KNN) and local outlier factor (LOF) to detect abnormal behaviors of ships, aiming at the problems of low accuracy of detection methods based on global

[*] Corresponding author

variables and low computational complexity of detection based on local variables[7]. Boddy et al. proposed a density-based local outliers detection model, which aims to increase defense in depth of health care infrastructure. Patterns in EPR data are extracted to analyze user behavior and device interaction, so as to detect and visualize abnormal activities[8].Tsuyoshi Id 'e proposes a new method for anomaly detection using multivariable noise sensor data, which solves two major challenges: providing variable-wise diagnostic information and automatic processing of multiple operating modes[9]. Ya-Lin Zhang address the scenario when anomalies are partially observed, i.e., they are given a large amount of unlabeled instances as well as a handful labeled anomalies. We refer to this problem as anomaly detection with partially observed anomalies, and proposed a two-stage method ADOA to solve it[10].J Su, H Li et al. comprehensively used the decision tree, information gain and attribution-oriented induction model for 35 indicators in financial audit data, and proved the effectiveness of the model in obtaining audit evidence[11]. E Kirkos, C Spathis, Y Manolopoulos et al. discussed the practicability of using decision tree, neural network and bayesian belief network to identify fraudulent data in the audit of financial statements, and compared the recognition effects of the three methods[12]. Khattab M et al. used back propagation neural network to detect denial of service attacks (DoS) common in vehicle self-organizing network[13].

In summary, outlier detection is often used in intrusion detection, fraud detection, medical and health anomaly detection, industrial damage detection, text data, sensor network, image data, voice recognition and other fields[14].

Application status of data mining technology in auditing field. Chichenlin USES data mining methods including Logistic regression, decision tree (CART) and artificial neural network (ANNs) to evaluate audit fraud[15]. Omid Pourheydari used four data mining classification techniques in Iran for the first time to establish a model that can identify auditors' opinions. The results demonstrate the ability of MLP neural networks to identify the opinions of different types of auditors[16]. Shao Jinwei proposed to improve Leaders operator to discover doubt points of corporate car audit data, and the algorithm realized the accurate clustering of big data of official vehicle audit, and identified the few abnormal clusters as potential doubts[3]. Zhong Jiaqi applies the association law to the hospital drug audit to find out whether there is an "abnormal" correlation between individual doctors and drug manufacturers, and applies the Logistic regression analysis to the audit of provident fund payers to find out the possibility of high-risk loans[17]. Zhang Cheng used the typical C4.5 algorithm in the decision tree algorithm to mine the loan data provided by the financial audit office, and found the information with possible irregularities as the audit doubt[18].

To sum up, at present, there are not many applications of outlier detection in computer audit in China, and most of the ones being studied are in the stage of theoretical research, which has not been implemented in computer audit[19]. Therefore, based on the current audit situation of official vehicles, this paper carries out outlier detection on official vehicles data, and uses the outlier detection algorithm to find potential doubts in official vehicles data, so as to provide audit basis for the personnel of the audit department and promote the reform of national official vehicles.

3 Related Knowledge and Methodology

An outlier is a data object that appears to be produced by a different mechanism and is significantly different from other data objects. Outlier detection (also known as anomaly detection) is a process to find out its behavior is different from the expected data points through a variety of detection methods. Outlier detection algorithm includes:

(1) statistical methods: firstly, a data model was established, and the anomalies were those objects that could not be perfectly fitted to the model; If the model is a collection of clusters, the exception is an object that does not belong to any cluster significantly. When using the regression model, exceptions are objects that are relatively far away from the predicted value. The advantage is that there is a solid theoretical basis for statistics, and these tests can be very effective when there is sufficient data and knowledge of the types of tests used. Disadvantages: there are fewer options available for multivariate data and poor detection possibilities for high-dimensional data.

(2) approach based on proximity: it is usually possible to define proximity measures between objects. Exception objects are those that are far away from other objects. These include: distance-based and

density-based methods. Usually cell methods, LOF algorithms. The advantages are: simple and easy to operate, the quantitative measurement of the object as outliers is given, and the data can be well processed even if there are different regions. Disadvantages: O(m2) time complexity, not suitable for large data sets; It is sensitive to parameter selection and difficult to choose parameters. Data sets with different density zones cannot be processed because it USES global thresholds and cannot account for variations in density. Although LOF algorithm deals with this problem by observing different k values and then obtaining the maximum outlier score, upper and lower bounds of these values still need to be selected.

(3) cluster-based method: outliers are detected by examining the relationship between objects and clusters. Advantages: clustering techniques based on linear and near-linear complexity (k-means) may be highly effective in finding outliers; The definition of a cluster is usually the complement of an outlier, so it is possible to find both a cluster and an outlier. Disadvantages: the resulting outliers and their scores may be very dependent on the number of clusters used and the existence of outliers in the data; The quality of clusters produced by the clustering algorithm has a great influence on the quality of outliers produced by the algorithm.

(4) classification based method: if the training data has class labels, a classification model that can distinguish normal data from outliers can be trained. A class of models is usually used.

3.1 DBSCAN Algorithm

Among density-based outlier detection algorithms, DBSCAN algorithm is the representative one, which can find outliers while clustering and is insensitive to outliers in the data set. It can divide data into several clusters according to density and distance, and filter data that does not belong to any clusters as abnormal points.

The calculation steps of this algorithm are as follows:

Take each data point xi as the center of the circle and draw a circle with eps value as the radius. This circle is called the eps neighborhood of xi and count the points contained in this circle.

(1) if the number of points in a circle exceeds the density threshold MinPts, then the center of the circle is denoted as the core point, also known as the core object.

(2) if the number of eps neighborhood points of a certain point is less than the density threshold but falls within the neighborhood of the core point, it is called the boundary point.

(3) the point that is neither the core point nor the boundary point is the noise point.

The algorithm is sensitive to user-defined parameters, and even subtle differences may lead to very different results. However, the selection of parameters is irregular and can only be determined by experience. Compared with k-means, BIRCH, which are generally only applicable to clustering of convex sample sets, DBSCAN can be applied to both convex and non-convex sample sets. Compared with the traditional k-means algorithm, the biggest difference of DBSCAN is that it does not need to input the number of categories K. Moreover, its clustering results are not biased, and the initial value of clustering algorithms such as k-means has a great impact on the clustering results.

3.2 LOF Algorithm

In many cases, outliers do not have binary properties, meaning that an object is not black and white, and it makes sense to explore how separate it is from its neighbors. Assign an exception level value, the local exception factor (LOF), to each object. This is an unsupervised outlier detection method, which is a representative algorithm of density-based outlier detection methods. Density-based LOF algorithm can effectively detect local outliers and global outliers in the data set, with high detection accuracy.

This algorithm will calculate an outlier factor LOF for each point in the data set, and determine whether it is an outlier by judging whether the LOF is close to 1.If LOF is much larger than 1, it is considered to be an outlier factor, and if it is close to 1, it is a normal point. It mainly determines whether the point is an abnormal point by comparing the density of each point p and its neighbors. If the density of point p is lower, it is more likely to be considered as an abnormal point. The density is calculated by the distance between the points, the farther the points are, the lower the density, and the closer the points are, the higher the density. Moreover, because the density of LOF is calculated through

the kth neighborhood of the point, rather than the global calculation, it is named as "local" abnormal factor. The disadvantage is that the setting of parameters has a great impact on the results.

The calculation steps of this algorithm are as follows:

(1) calculate k distance of p: the kth distance of point p, that is, the distance of the point k away from p, excluding p.

(2) calculate k-distance neighborhood of p: the kth neighborhood of point p is all the points within the kth distance of p, including the kth distance.

(3) calculate reach-distance: accessible distance; if less than the kth distance, the accessible distance is the kth distance; if greater than the kth distance, the accessible distance is the true distance.

(4) calculate local reachability density: local reachable density.

(5) calculate local outlier factor: local outlier factor.

3.3 Anomaly Detection Based on LOF Algorithm

For most of the official vehicles audit data are unlabeled data, namely abnormal data are unlabeled, so this paper choose use unsupervised DBSCAN clustering algorithm first to abnormal data of the initial screening, because unsupervised LOF algorithm needs to set contamination parameters, but for unlabeled data, it is difficult to determine the contamination parameters, this parameter setting of numerical size is different, the results of anomaly detection is different, so need to use DBSCAN to roughly determine the abnormal point proportion in the sample data, In this way, the range of contamination parameter of LOF algorithm is determined.

Generally, for a clustering task, we hope to get clusters that are as close as possible in the cluster and as far away from each other as possible. The Silhouette Coefficient is an evaluation index of the intensity degree of clusters. The formula (1) is expressed as follows:

$$s = \frac{b - a}{\max(a, b)} \tag{1}$$

where a represents the mean distance between samples in the same cluster and each other, which is called intra-cluster dissimilarity, b represents the average distance from the sample to all the samples of the nearest cluster except the cluster in which it is located, which is called inter-cluster dissimilarity. The closer s is to 1, the better the clustering effect of samples will be; The approximation of s to 0 means that the sample is on the boundary of the two clusters.

In this paper, we only care about abnormal data, so as long as its silhouette coefficient is greater than 0, the closer it is to 1, the better. Adjust the two parameters of DBSCAN algorithm: eps value and MinPts value, and when the silhouette coefficient of the clustering results is relatively the best, calculate the proportion of abnormal points according to the clustering results. Then set the contamination parameter of LOF algorithm, and use LOF anomaly detection, and finally calculate the abnormal point. The abnormal detection model of official vehicles audit in this paper is showinfigure1:

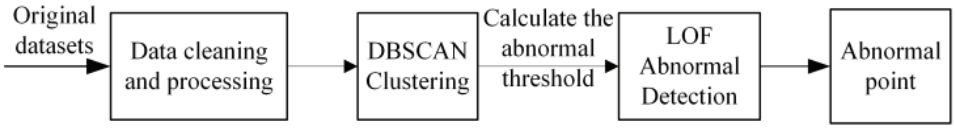

Figure 1. Anomaly detection model for official vehicles

4 Experiment

4.1 Data Extraction

The experimental data in this paper are derived from the bus data of a district audit bureau in Chengdu, covering the data of maintenance, maintenance, car rental and fuel, etc., most of which are tabular data. This paper extracts all the bus maintenance data in recent years from the bus audit data table as the research object, and extracts the marked data as the sample data. The main fields of the maintenance

data table are: license plate number, repair order number, repair unit, repair date, repair unit, repair amount, and repair item. However, maintenance data sheet is not enough, and bus audit also needs comprehensive vehicle information. Therefore, combined with the basic data collection table of bus, the fields are as follows: vehicle model, starting date of use, displacement, capacity, usage purpose, owner, vehicle number, brand, and phone number of vehicle owner.

4.2 Data Cleaning and Selection

Data cleaning. Firstly, unreasonable sample data should be screened and deleted according to the actual situation and business characteristics. Data with maintenance item =0 should be excluded. Delete the data of maintenance unit =0; Delete the data with maintenance date =0; Cull data using start date =0.

Feature selection. Two tables some fields, such as license plate number, BaoXiuChan, repair service units, such as telephone all motor vehicles, are has nothing to do with mining model, combined with the audit regulations regarding the use of the expert experience and the reform of official management, analysis of maintenance audit key attributes: cars, maintenance amount, use the start date, date of maintenance.

4.3 Data Preprocessing

The features are divided into two categories: discrete data and continuous data, which are classified as follows: (1) discrete features, also known as typed variables, present discrete state. Including vehicle models and maintenance projects, such as vehicle model 'FV7146FBDGG',it's a combination of letters and Numbers;(2)continuous characteristics: it can take any value within a certain range, and the value is continuous, including the use time, maintenance money, etc.

(1) calculate attributes.

Since the audit maintenance data needs to take into account the service time of official vehicles and judge the reasonableness of maintenance costs in combination with vehicle types, but this feature is not found in the original data table, the difference between the maintenance date and the starting date of use should be calculated to obtain the service time of each data from the starting date to the maintenance day. Since we're directly subtracting it out in days, we need to convert it to months and round it up. According to the features of the maintenance amount with decimals and different values, the features of the amount are rounded to the decimal. Finally, the processed data is saved as a new file so that it can be imported directly into the calculation later.

(2) data type conversion.

Discrete data converted to numeric data. One-hot coding is used to extend the value of discrete features to the Euclidean space. A value of discrete features corresponds to a point in the Euclidean space, which makes the calculation of distance between features more reasonable. After one-hot coding of discrete features, the features of each dimension can be regarded as continuous features. We can normalize each one-dimensional feature just as we can normalize continuous features.

(3) normalization.

The numerical features are normalized. To numerical numeric data with average filling empty, adjust the distribution of the feature data into standard fall, also known as the Gaussian distribution, which makes the data average d 0, variance 1, standardization of the reason is that if some features of the variance is too big, will dominate the objective function, so that the parameter estimator is unable to correctly to learn other characteristics.

Finally, the ColumnTransformer package of sklearn library in Python can be used to preprocess columns of different types in parallel with the above different methods.

4.4 Evaluation Criteria

Evaluation indexes for evaluating the effectiveness of abnormal detection are: detection rate is denoted as DR, false alarm rate is denoted as FAR, and accuracy is denoted as ACC. Three indexes are used to evaluate the effectiveness of the algorithm. The calculation method is shown in the following formula (2):

$$DR = \frac{TN}{N}$$
$$FAR = \frac{FN}{P} \tag{2}$$
$$ACC = \frac{TP + TN}{P + N}$$

In the above formula(2), P represents the number of normal samples in the test data set, N represents the number of abnormal samples in the test data set, and the meanings of TP, FN, FP and TN are shown in the confusion matrix in table 1:

Table 1. Confusion matrix

	labeled as positive	**labeled as negative**
predicted as positive	True Positive(TP)	False Positive(FP)
predicted as negative	False Negative(FN)	True Negative(TN)

The confusion matrix above clearly shows that FN represents the case where the actual sample is normal, but the output of the algorithm is abnormal.TN represents the situation where the output of the algorithm is an abnormal sample; TP means that the actual sample is normal, and the output of the algorithm is also abnormal. The detection rate reflects the ability of the algorithm to recognize anomalies. The false alarm rate indicates the false alarm probability of normal samples. And the classification accuracy shows the algorithm's ability to distinguish normal samples from abnormal samples on the whole[20].

4.5 Results and Analysis

In order to verify the validity of this model, a comparison experiment is made between the model and other anomaly detection algorithms. Three anomaly detection algorithms have been widely used in the industry, namely isolated forest, LOF and DBSCAN. The detection rate, false alarm rate and classification accuracy were compared, and the comparison results of evaluation indexes were shown in table 2.

Table 2. Evaluation metrics comparison of the four techniques

Model	ACC	DR	FAR
DBSCAN+LOF	99.12%	26%	0.38%
DBSCAN	98.27%	24.91%	0.39%
LOF	97.76%	21.43%	0.85%
Isolation Forest	98.60%	7.14%	0.98%

The comparison results show that the classification accuracy of DBSCAN is similar to that of isolated forests, but the performance of LOF based on DBSCAN is still the best, with ACC of 99.12%, higher than that of isolated forest algorithm with ACC of 98.60%, its detection rate is 26%, and the false alarm rate is 0.38%, which is better than other algorithms, indicating that this model has strong abnormal detection ability in official vehicles audit data.

5 Conclusion

In order to discover potential doubts from unlabeled official vehicles audit data and assist auditors to complete official vehicles audit efficiently and accurately, This paper proposes a potential doubts discovery model for official vehicles audit data combining DBSCAN and LOF, and compares it with isolated forest algorithm, LOF and DBSCAN. Experiments show that the anomaly detection model

based on LOF is superior to the other three algorithms in detecting unmarked official vehicles audit data and has strong generalization ability. In the further research, we will try to experiment on a larger data set or try to tune the model to achieve the optimal performance of the model.

Acknowledgments. This work was supported by grants from the National Natural Science Foundation of China #71373216, in part by the Innovation Scientific Research Program for Graduates in Southwest Minzu University (No. CX2019SP33).

References

1. Jiang Ma. Analysis of the current situation and countermeasures of the reform of public institutions' official vehicles in China [J]. Times finance,2019,(9): 85-86, 90
2. Deron Lianga; Fengyi Lin and Soushan Wuc. Electronically auditing EDP systems With the support of emerging information technologies[J]. International Journal of Accounting Information Systems, 2001,Vol.2: 130-147
3. Jinwei Shao; Jun Lin; Yating Liu; Jia-li Xiao. Discovery of potential audit doubts based on improved Leaders operators [J]. Computer and modernization, 2018,(4): 79-83
4. Yanna Tan. Study on cluster outlier detection in the field of auditing [D]. Harbin engineering university,2011
5. Nan Wang; Zizhao Zhang; Xibin Zhao; Quan Miao; Rongrong Ji; Yue Gao. Exploring High-Order Correlations for Industry Anomaly Detection[J].IEEE Transactions on Industrial Electronics,2019,Vol.66(12): 9682-9691
6. Yang, Jun; Zhou, Chunjie; Yang, Shuanghua; Xu, Haizhou; Hu, Bowen. Anomaly Detection Based on Zone Partition for Security Protection of Industrial Cyber-Physical Systems.[J].IEEE Transactions on Industrial Electronics, 2018,Vol.65(5): 4257-4267
7. Yingfu Huang; Qiurong Zhang. Identification of Anomaly Behavior of Ships Based on KNN and LOF Combination Algorithm.[J]. AIP Conference Proceedings, 2019,Vol.2073(1): 020090(1-8)
8. Boddy, Aaron J.; Hurst, William; Mackay, Michael; El Rhalibi, Abdennour. Density-Based Outlier Detection for Safeguarding Electronic patient Record Systems[J]. IEEE ACCESS, 2019,Vol.7: 40285-40294
9. Ide, T (Ide, Tsuyoshi); Khandelwal, A (Khandelwal, Ankush); Kalagnanam, J (Kalagnanam, Jayant). Sparse Gaussian Markov Random Field Mixtures for Anomaly Detection[J]. 2016 IEEE 16TH INTERNATIONAL CONFERENCE ON DATA MINING (ICDM), 2016,: 955-960
10. Ya-Lin Zhang; Longfei Li; Jun Zhou; Xiaolong Li; Zhi-Hua Zhou. Anomaly Detection with Partially Observed Anomalies[A]. WWW '18: Companion Proceedings of the Web Conference 2018[C], 2018
11. Sun, J.; Li, H.. Data mining method for listed companies' financial distress prediction (Article)[J]. Knowledge-Based Systems, 2008,Vol.21(1): 1-5
12. Kirkos, E.; Spathis, C.; Manolopoulos, Y.. Data Mining techniques for the detection of fraudulent financial statements (Article)[J]. Expert Systems with Applications, 2007,Vol.32(4): 995-1003
13. Khattab M. Ali Alheeti; Klaus McDonald-Maier. Intelligent intrusion detection in external communication systems for autonomous vehicles[J]. Systems Science & Control Engineering, 2018,Vol.6(1): 48-56
14. Chandola, V.; Banerjee, A.; Kumar, V.. Anomaly detection: A survey (Review)[J]. ACM Computing Surveys, 2009,Vol.41(3)
15. Lin, Chi-Chen; Chiu, An-An; Huang, Shaio Yan; Yen, David C.. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments[J]. KNOWLEDGE-BASED SYSTEMS, 2015,Vol.89: 459-470
16. Omid Pourheydari; Hossein Nezamabadi-pour and Zeinab Aazami. Identifying qualified audit opinions by artificial neural networks[J]. African Journal of Business Management,2012,Vol.6(44): 11077-11087
17. Jiaqi Zhong. Application research of data mining technology in computer audit [D]. Nanjing university of posts and telecommunications, 2017
18. Cheng Zhang. Research and application of data mining technology in financial audit [D]. Anhui university, 2014
19. Xiaowei Zhang. Audit evidence acquisition system based on data matching and outlier detection technology [D]. Nanjing university of aeronautics and astronautics, 2009
20. Renyu Liu. Network anomaly detection algorithm based on outlier detection [D]. Chongqing university, 2018

