Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230802

A Method for Anti-Leakage in Car Insurance Based on the Co-Occurrence Relationship Between Auto Parts

Jun JIANG, Fengyu YANG[1], Yu XIONG, Longhui LIU and Xin YANG
School of Software, Nanchang Hangkong University, Nanchang, China

Abstract. Car insurance fraud is a high-risk area in insurance and it accounts for as much as 80% of all insurance fraud, with a significant portion of it stemming from the risk of parts leakage. Current anti-leakage techniques in auto insurance mainly rely on the analysis of individual parts data such as vehicle accident records and parts damage lists, which neglects the consideration of the correlation among parts and leads to difficulty in identifying leaked parts effectively. This study proposes a method based on the co-occurrence relationship and density clustering of auto parts to detect parts leakage. In this research, an undisclosed dataset on car insurance fraud was utilized to conduct experiments, and the detection results of parts leakage were obtained. This method takes into account the correlation among auto parts, and possesses higher accuracy and practicality.

Keywords. Car insurance fraud, Parts Leakage Risk, Part Co-occurrence Relationship, Density clustering

1. Introduction

One common type of car insurance fraud involves part leakage, which refers to the fabrication of vehicle part damage by submitting fraudulent loss reports to insurance companies in order to obtain unlawful compensation[1].

The primary objective of this study is to provide efficient anti-part leakage techniques to car insurance firms, ultimately reducing fraud costs and improving the economic benefits of insurance companies. To achieve this goal, we proposed a methodology based on part co-occurrence relationships and density clustering to detect part leakage behavior. This method advocates for the consideration of inter-component correlations by utilizing a density clustering algorithm to cluster part data and evaluate part suspiciousness using the risk of potential part leakage as a determinant factor[2].

To validate the effectiveness of our proposed method, we used actual car insurance fraud case data. Notably, our methodology starts with the co-occurrence relationship of parts, and in the clustering process, considers the correlations among various parts, leading to outcomes that are more interpretable[3]. These findings imply that this approach has broad potential applications in the actual automotive insurance industry.

[1] Corresponding Author, Fengyu YANG, School of Software, NanChang Hangkong University, Nanchang, China; E-mail: 99770277@qq.com.

2. Methodology

This section provides a detailed explanation of the specific steps involved in the car insurance part co-occurrence based anti-leakage analysis method. The method consists of four main steps: data preprocessing[4], construction of fully connected relationships among parts[5], implementing adaptive DBSCAN clustering algorithm to cluster parts[6], and calculating the co-occurrence relationship and suspiciousness between various parts.

2.1. Data Preprocessing

This study performed an analysis on a non-public dataset of historical car insurance fraud cases. The dataset comprises several fields relevant to car insurance fraud cases. For the purpose of this study[7], we extracted and utilized only the relevant data fields, including a claim number, license plate number, and parts list.

This preprocessing step involved filtering and cleaning the original data, as shown in Figure 1, establishing a foundation for further data processing and analysis, and ensuring the accuracy and consistency of the data quality[8].

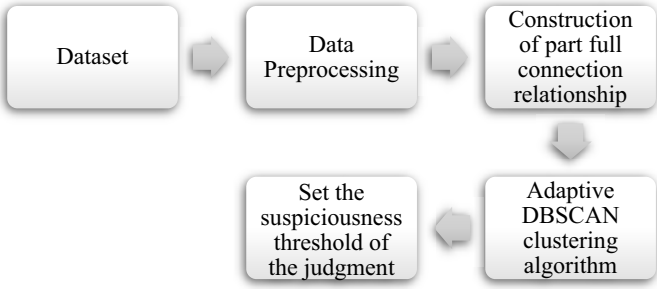

Figure 1. The specific flow chart of the analysis method of suspicious parts against leakage

2.2. Construction of Fully Connected Relationships Among Parts

Based on the dataset and list of damaged auto parts obtained through the preprocessing[9], we further filtered the auto part list. Subsequently, every auto part was connected with all the other parts, resulting in a fully connected dataset T containing all the auto parts with their corresponding relationships.

Based on dataset T, we utilized a fully connected topology builder to manipulate and transform the information among the nodes in the list. The specific method for constructing the fully connected topology is detailed below:

1. Firstly, Input the list of nodes that needs to be connected and the pre-set adjacent matrix; compute the number of elements within the list, and verify if the adjacent matrix is already properly initialized.
2. Initialize the list of node names, and add every element from the input list of nodes to it. Apply necessary sorting or de-duplication to the list of node names.
3. Initialize the adjacent matrix based on the set of node names, and initialize all matrix cells with the value 0.
4. Iterate through all elements in the list of nodes, and calculate their edge connections. For each node and its neighboring node, increment the value of the

corresponding element in the adjacent matrix by 1.

5. After iterating through the list of nodes, modify the diagonal elements as follows: add the sum of the out degree and in degree of all nodes to the diagonal elements, and reset the non-zero elements located symmetrically.

6. Finally, a relationship network object is obtained, where each node is connected to every other node.

This step was taken to quantify the relationships between auto parts, enabling us to conduct cluster analysis and calculate their respective degree of suspicion.

2.3. Adaptive DBSCAN Clustering Algorithm

To achieve adaptive clustering of the dataset, we adopted the DBSCAN algorithm for cluster division and proposed a strategy for adapting threshold parameters. Based on the clustering results, we counted the number of samples and non-core points connected to each cluster. We treated these counts as new features and clustered the data to obtain new clustering results, as shown in Figure 2, which illustrates the pseudocode for adaptive clustering.

Algorithm 1: Adaptive Clustering Based on DBSCAN

Input: Data matrix X, initial radius r, initial minimum number of samples *min samples*, initial threshold t
Output: Cluster labels
Initialize *r new* = *r*;
Initialize *min samples new* = *min samples*;
Initialize *iteration* = 0;
Initialize *AR new* = 0;
Initialize *AR old* = 0;
while *iteration* < *max iterations* **do**
    Compute distance matrix $D(X)$;
    Train DBSCAN with $D(X)$ to get labels;
    Extract features for each cluster;
    Train DBSCAN with the new features to get new labels; Calculate *AR_new*;
    if *AR new* > *t* **then**
        r new = *r new* ∗ (1 + *alpha*);
        min samples new = *min samples new* ∗ (1 + *beta*);
    end
    else if *AR_new* < *t* **then**
        r new = *r new* ∗ (1 − *alpha*);
        min samples new = *min samples new* ∗ (1 − *beta*);
        if *r_ new* < *r_min or min_samples new* < *min_samples_min*
        then
            beak;
        end
    end
    else
        break;
    end
    AR_old= AR_new;
    iteration = *iteration* + 1;
end
return *New cluster labels*;

Figure 2. Pseudocode for adaptive clustering

To improve the accuracy and efficiency of clustering, we calculated the adjusted Rand index (AR) by comparing the new clustering results with the old ones. The formula for calculating AR is TP + TN / (TP + TN + FP + FN), where TP represents the correct identification of core points in the cluster, TN represents the correct identification of noise points, FP represents the number of noise points that were wrongly identified as core points, and FN represents the number of core points wrongly identified as noise points.

2.4. Calculating the Degree of Suspicion for Co-occurrence of Auto parts

Based on a clustering approach, we can represent each auto part as a vector. By calculating the similarity between the least similar pair of auto parts in an individual case, we can represent the degree of suspicion for the co-occurrence of all auto parts in that case. The smaller the similarity value, the more suspicious the co-occurrence of the auto parts.

Next, we used an iterator to traverse the adjacency matrix to find the least similar pair of auto parts and calculated their similarity value. If this similarity value was smaller than the current minimum similarity value, we updated the minimum similarity value and corresponding element names with this new value.

We used the similarity value between the least similar pair of auto parts and their names as the basis to determine the degree of suspicion[10]. Specifically, we used the similarity value of the least similar auto part pair as the degree of suspicion, and set a threshold value for the degree of suspicion. If the similarity value of the least similar pair of auto parts was lower than the set degree of suspicion threshold value, we output a sorted list of auto part names.

3. Result and Discussion

In this section, we utilized the adaptive algorithm DBSCAN to cluster the components in a non-public dataset of vehicle insurance historical cases and determined the suspiciousness of all components in each case through vectorized calculation of their co-occurrence relationship. We further evaluated the effectiveness of our approach by analyzing and interpreting the clustering results. The non-public dataset of vehicle insurance historical cases consists of a total of 100,000 cases, of which about 3,000 cases were detected with vehicle component leakage issues.

3.1. Analysis and Presentation of Component Clustering Results

Based on the evaluation of the co-occurrence degree of all components using the DBSCAN clustering algorithm and vectorization method, we obtained the clustering results of the components and further analyzed and interpreted them.

Different from traditional single feature classification[11], this method clusters data points based on multiple features they possess, which can more fully reflect the actual relationships between components.

Presenting and analyzing the clustering results can help us better understand the dependency relationships between components in each case and support us in making informed decisions on preventing component leakage[12].

We conducted an analysis on the clustering results as shown in Table 1 and found that a majority of the components have been successfully clustered into the corresponding vehicle systems with no directional errors detected.

Furthermore, the tabular visualization also highlights the co-occurrence relationships among the components and validates the efficiency and precision of the component clustering process.

Table 1. Analysis of Adaptive Clustering Results

Clustered parts list	System	Collision Direction
[Turning Horizontal Pull Rod Ball Joint (Right), Front Half Shaft (Right), Turning Horizontal Pull Rod (Right), Front Shock Absorber (Right), Front Lower Arm Ball Joint (Right), Front Stabilizer Link (Right), Front Wheel Hub Bearing (Right), Front Lower Arm (Right), Front Steering Knuckle (Right)]	Front suspension system	Right front wheel direction
[Front Door Frame Seal (Right), Front Door Glass Felt Channel (Right), Front Door Glass (Right), Front Door Glass Outer Pressing Strip (Right), Front Door Shell (Right), Front Door Upper Hinge (Right), Front Door Lower Hinge (Right), Front Door Outer Handle (Right)]	The door and window system	Right front door direction
[Front Headlight (Left), Front Bumper Support (Left), Front Fender Liner (Left), Front Fog Light Cover (Left), Front Fog Light (Left), Front Fender (Left)]	Exterior body safety device system	The left front direction

3.2. Presentation of Component Co-occurrence Suspiciousness

We used a vectorization method to calculate the similarity measure between the least similar pair of components as a suspiciousness indicator and exported the results in table form as shown in Table 2.

We introduce the following equations to compute the suspicion matrix of auto parts:

$$data_{mat(i,i)} = \max_{j \in rows} \left(data_{mat(i,j)} \right) \tag{1}$$

In the equation (1), the symbol i denotes the ith auto parts in the data matrix. The symbol N refers to the number of rows in the data matrix (i.e., number of autoparts) while $rows$ refers to the set of all rows of the data matrix. Moreover, the $\max$ function in equation (1) is used to locate the maximum value in the ith row of the data matrix, which is then utilized to set the element $\mathrm{data}_{\mathrm{mat(I,I)}}$ on the diagonal.

$$sim_{mat(:,i)} = \sqrt{\sum_{j=1}^{N} \left(data_{mat(:,j)} - data_{mat(:,i)} \right)^2} \tag{2}$$

In the equation (2), this formula calculates the similarity matrix between each pair of samples in the dataset using the Euclidean distance.

Table 2. Suspicious Degree of Component Co-occurrence (Partial Results)

Suspicion degree	Part1	Part2
0.0001198	Front Door Outer Handle (Left)	Front Bumper
0.0001269	Rear Bumper Parking Sensor	Front Door Glass Outer Pressing Strip (Right)
0.0001441	Engine Dress Up Cover	Inner Tail Light (Left)
0.0001619	Engine Dress Up Cover	Rear Bumper Inner Liner
0.0001681	Front Wheel Mudguard	Rear Bumper Inner Liner
0.0001716	Front Windshield Wiper Washer Pump	Rear Fog Light (Left)
0.0001734	Engine Splash Shield (Front)	Rear Panel

Each row in Table 2 represents a list of parts with low co-occurrence in the damaged components of the vehicles involved in a case. The higher the value of similarity, the lower the level of suspicion.

4. Conclusions

This study used a non-public vehicle insurance dataset to test the feasibility of our method, and only extracted essential fields from the dataset, such as case numbers, license plates of insured vehicles, and corresponding lists of damaged components.

In the experiments, by analyzing the clustering results and the suspiciousness of component co-occurrence, we can identify the collision dependency relationships between various components in each case and assist in making effective decisions on component leakage. However, in cases involving a large number of components, some of the co-occurrence results may be challenging to interpret and explain reasonably.

Although this study faces many challenges in solving vehicle insurance fraud problems, we believe that this data-driven analysis approach can provide new insights and methods for future research.

References

[1] Anti-Fraud Joint Task Force of Vehicle Insurance, Study on Vehicle Insurance Fraud and Anti-Fraud Measures, Insurance Studies, vol. 06, pp. 3-10, 2021.

[2] Kim, J.H., Choi, J.H., Yoo, K.H., Lee, Y. and Kim, Y, AA-DBSCAN: An Approximate Adaptive DBSCAN for Finding Clusters with Varying Densities, The Journal of Supercomputing, vol. 75, no. 1, pp. 142-169, Jan. 2019.

[3] Tran, T.N., Drab, K., Daszykowski, M., Revised DBSCAN algorithm to cluster data with dense adjacent clusters, Chemometrics and Intelligent Laboratory Systems, vol. 120, pp. 92-96, 2013. DOI: 10.1016/j.chemolab.2012.10.003.

[4] Yu, W., Feng, G., and Zhang, W., Research on motor vehicle insurance fraud detection system and gang identification, Insurance Research, vol. 2017, no. 02, pp. 63-73, 2017, doi: 10.13497/j.cnki.is.2017.02.007.

[5] Macedo, A.M., Cardoso, C.V., Neto, J.S.M. and Cunha, C.A.C.B., Car insurance fraud: The role of vehicle repair workshops, International Journal of Law, Crime and Justice, vol. 65, 100456, 2021. DOI: 10.1016/j.ijlcj.2021.100456.

[6] Subudhi, S. and Panigrahi, S., Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection,Journal of King Saud University-Computer and Information Sciences, vol. 32, no. 5, pp. 568-575, 2020. doi: 10.1016/j.jksuci.2017.09.010.

[7] Ye, M. Y., Research on identifying insurance fraud based on BP neural network – taking claims in China's motor vehicle insurance as an example, Insurance Studies, vol. 2011, no. 03, pp. 79-86, 2011, doi: 10.13497/j.cnki.is.2011.03.012.
[8] Aslam, F., Hunjra, A.I., Ftiti, Z., et al., Insurance fraud detection: Evidence from artificial intelligence and machine learning, Research in International Business and Finance, vol. 62, 101744, 2022. DOI: 10.1016/j.ribaf.2021.101744.
[9] Benedek, B., Ciumas, C., Nagy, B.Z., Automobile insurance fraud detection in the age of big data – a systematic and comprehensive literature review, Journal of Financial Regulation and Compliance, 2022. DOI: 10.1108/JFRC-07-2020-0099.
[10] Li, P., Shen, B., Dong, W., An anti-fraud system for car insurance claim based on visual evidence, arXiv preprint arXiv:1804.11207, 2018.
[11] Debener, J., Heinke, V., Kriebel, J., Detecting insurance fraud using supervised and unsupervised machine learning, Journal of Risk and Insurance, 2023. DOI: 10.1111/jori.12427.
[12] Hanafy, M., Ming, R., Using machine learning models to compare various resampling methods in predicting insurance fraud, J. Theor. Appl. Inf. Technol, vol. 99, no. 12, pp. 2819-2833, 2021.