Research on classification method of new energy vehicle information security risk assessment

Tonghong Chong\textsuperscript{1*}, Xianfeng Jia\textsuperscript{1}, Zhi Wu\textsuperscript{1}, Hao Zhao\textsuperscript{1}, Tianyu Liu\textsuperscript{1}

\textsuperscript{1} China Automotive Technology \& Research Center Automotive Data Co, Ltd, Tianjin, 300000, China

*Corresponding author’s e-mail: chongtonghong@catarc.ac.cn

Abstract. With the rapid development and popularization of intelligent Internet connected vehicle, the automobile and the physical world are combined together, which makes users have strong technology experience and life convenience. While enjoying the convenience and experience brought by the intelligent Internet connected vehicle, we are also faced with the risk of information security in the network world, even the security problems of the network world directly affect the security of the physical world. With the improvement of automobile ‘four modernizations’, the research on information security has become an important topic and research hotspot in the field of automobile. Risk assessment is one of the important research and work contents of automobile information security. Impact valuation is an important indicator of risk assessment. It is very important to determine the optimal cluster number for risk assessment model and risk assessment. In this paper, the fuzzy clustering impact estimates are divided into 2-12 categories during Evergrande’s risk assessment of an electric vehicle. The optimal number of clusters is determined by calculating the mixed F-statistics of each type of data. Conclusion shows that the best clustering number of impact estimates is 5.

1. Introduction

1.1. Research background

With the development of intelligent vehicle and Internet of vehicles technology, there are more and more automotive electronic control systems. The automobile will no longer be an isolated unit, but become a mobile intelligent network terminal\cite{1}. The Internet of vehicles is an Internet of things with vehicles, roads and other transportation infrastructure as its physical world. Information security includes not only the security problems in the Internet of things, but also the special characteristics of its physical objects. After active safety, passive safety and functional safety, automotive information security will become the fourth security issue in the automotive field\cite{2}.

With the increasing market share of 5g commercial and intelligent vehicles, the output value is expanding rapidly. Whether the potential safety hazards can be properly solved will affect the future development of intelligent vehicles. However, the huge "economic market" of Internet of vehicles industry with intelligent vehicle operating system will become the new “prey” of global hackers\cite{3}.

The intelligent Internet vehicle is based on the Internet of vehicles. Specifically, all the hardware are connected in the car and run in the cloud. For example, the car can be unlocked, honked, started and opened windows through the network. In addition, the information will also be connected to the Internet at the same time. The route, location and audio in the car will become valuable big data. These data are used to improve the accuracy of automatic driving and user experience. Once these data
fall into the hands of hackers, they will become transparent everywhere, whether for an individual or for a country. The biggest problem of smart cars is not about technology, but information security[4].

1.2. Definition of risk assessment
From the perspective of information security, risk assessment is the assessment of the threat, weakness and impact of information assets (i.e. the information set of an event or thing), as well as the possibility of risk caused by the comprehensive action of the three. As the basis of risk management, risk assessment is an important way for organizations to determine information security requirements and belongs to the planning process of organization information security management system[5].

1.3. Risk assessment methods
The common risk assessment methods are as follows:

(1) Risk factor analysis
Risk factor analysis method is a risk assessment method that can evaluate and analyze the factors that may lead to risk occurrence, so as to determine the probability of risk occurrence. The general idea is as follows: investigate the risk source → identify the risk transformation conditions → determine whether the transformation conditions are available → estimate the consequences of risk occurrence → risk evaluation.

(2) Internal control evaluation method
The internal control evaluation method is a method to determine the audit risk through the evaluation of the internal control structure of the auditee. Since the internal control structure is directly related to the control risk, this method is mainly used in the evaluation of control risk.

(3) Analytical review method
The analytical review method is to analyze the main ratios or trends of the audited entity by certified public accountants, including the investigation of abnormal changes and the differences between these important ratios or trends and the expected amount and relevant information, so as to speculate whether there is the possibility of material misstatement or omission in the accounting statements. The commonly used methods are comparative analysis, ratio analysis and trend analysis.

(4) Qualitative risk assessment method
Qualitative risk assessment method refers to those methods that can conduct qualitative assessment of audit risk through observation, investigation and analysis, and with the help of CPA's experience, professional standards and judgment. It has the advantages of convenience and effectiveness, and is suitable for assessing various audit risks. The main methods are: observation method, investigation and understanding method, logical analysis method, similar estimation method.

(5) Risk rate risk assessment method
Risk rate risk assessment method is one of the quantitative risk assessment methods. The basic idea is: first calculate the risk rate, and then compare the risk rate with the risk safety index. If the risk rate is greater than the risk safety index, the system is in the risk state. The greater the difference between the two data, the greater the risk.

Currently, the common sharing evaluation method of automobile information security is based on iso-21434 regulations and industry rules and a set of mature risk assessment system is formed[6]. In the process of risk assessment, data or risk classification should be divided into several categories, usually using subjective judgment. For example, if it is divided into three levels, it corresponds to low risk, medium risk and high risk. If it's divided into four categories, it corresponds to low risk, medium and low risk, medium and high risk and extremely high risk. Scientific classification of the calculated data can better establish the risk assessment model, so as to better complete the risk assessment.
2. Theoretical basis of cluster analysis

2.1. Basic concepts of cluster analysis
Cluster analysis is widely used in many fields, including computer information science, mathematics, statistics and so on. Clustering analysis is one of machine learning and unsupervised learning methods. It is widely used in pattern recognition, market research, information retrieval and other research fields[7]. With the continuous expansion and extension of these fields, the theory and technology of cluster analysis have been constantly developed and improved. Especially in the era of big data, the application prospect of cluster analysis is unprecedented. Clustering analysis theory and technology are used to measure the similarity and mutual relationship between data sets. It can be divided into a specific number of classes. Clustering a large number of data to simplify the data structure, in many research areas can simplify the model parameters, so as to improve the reliability and efficiency of research. The data structure can be simplified by clustering a large number of data, model parameters are also simplified in many research areas to improve the reliability and efficiency of research.

There is a big difference between clustering data sets and classifying data sets. Generally speaking, the data structure in the data set corresponding to the clustering is not different and the number of clusters is not known in advance. However, classification is different. First of all, there are obvious structural levels within the data, and usually the number of classification is known before classification. However, there are some similarities between them: for the data sets in the same category or in the same cluster, some features or some features of the data have obvious similarity. For the data sets which do not belong to the same category or the same cluster, one or some characteristics of the data are obviously different. The above two points are usually used as indicators to evaluate the effect of clustering or classification.

2.2. Three key points of data clustering
According to the idea of "birds of a feather flock together", the sample sets of unknown categories are clustered according to the degree of similarity among samples. The similar ones are clustered into one category, and the dissimilar ones are clustered into other category. This classification is called clustering. There are three main points in data clustering:

(1) Similarity measurement: which indexes are used to represent the similarity between samples.

(2) Clustering criterion: what kind of clustering criterion function is used to make a certain clustering criterion reach the extreme value.

(3) Clustering algorithm: which algorithm is used to find the best clustering result that makes the criterion function take the extreme value.

2.3. Steps and application of cluster analysis
Clustering analysis is a complex process. In order to ensure the accuracy and effectiveness of clustering, there are generally many steps. Cluster analysis has many applications in practical applications.

For supervised learning, we assume that all patterns are represented by features and form dimensional eigenvectors. In order to complete a clustering task, the following steps must be followed:

(1) Feature selection: features must be selected appropriately to contain as much information as possible. In features, reducing and minimizing information redundancy is the most important goal. It is necessary to preprocess the features in supervised classification before using them.

(2) Proximity measure: used to measure quantitatively how two feature vectors are "similar" or "not similar". Naturally, it is important to ensure that all selected features have the same proximity and no dominant features.

(3) Clustering criteria: this depends on the expert's interpretation of "determinable", which is based on the types of classes implied in the data set. For example, dense class eigenvectors of dimensional space can be judged by one criterion, but elongated classes need another criterion. Clustering criteria can be expressed by cost function or other criteria.
(4) Clustering algorithm: the nearest neighbor measure and clustering criteria have been adopted. This step involves selecting a specific algorithm to reveal the clustering structure of the data set.

(5) Result verification: once the result is obtained by clustering algorithm, the correctness must be verified. The approximation test is usually used.

(6) Result judgment: in many cases, experts in the application field must use other experimental evidence and analysis to determine the clustering results, and finally make the correct conclusion.

In many cases, the "clustering trend" step should be included. This step includes various tests, mainly testing whether the valid data has the clustering structure[8]. For example, data sets may be completely random, and trying to cluster is meaningless.

2.4. Common clustering methods

Since the concept of clustering was first proposed to the present research, a large number of clustering methods[9] have been developed. In practical application, the clustering algorithm will be determined according to different application fields.

At present, K-means algorithm, hierarchical clustering algorithm, SOM algorithm and FCM clustering algorithm[10] are relatively mature and widely used.

2.4.1. k-means algorithm

The algorithm is the representative of the objective function clustering method based on the prototype. This algorithm is a hard clustering algorithm. It is a function with certain distance from the data point to the prototype as the optimization objective. The adjustment rule of iterative operation[11] is obtained by calculating the extreme value of the objective function. Euclidean distance is the similarity measure index of clustering algorithm, which is to find the optimal classification corresponding to a certain initial clustering center vector V, so that the evaluation index J is the minimum. The algorithm uses the sum of square error as the clustering criterion function. It is one of the classic clustering algorithms in partition method. Because of its high efficiency, this algorithm is widely used in clustering large-scale data.

2.4.2. hierarchical clustering algorithm

According to the decomposition order of the algorithm, the order is from bottom to top or from top to bottom. There are two kinds of hierarchical clustering algorithms: bottom-up hierarchical clustering algorithm and up-bottom hierarchical clustering algorithm. Among them, the leaf node of the tree represents the network node, and the non-leaf node is usually obtained by merging the similar or close distance sub nodes. Figure 1 is a schematic diagram of hierarchical clustering method.

Fig.1 Schematic diagram of hierarchical clustering method
2.4.3. **SOM clustering algorithm**

Self-organizing maps (SOM) algorithm is a typical unsupervised learning method. SOM network algorithm has been widely used in machine learning and other related research and application[12], because it has many advantages, such as good visualization and self-organization.

SOM neural network consists of two parts, input layer and output layer. Generally, a high-dimensional input vector serves as the input of the input layer. The output layer of SOM network is composed of some ordered nodes on the two-dimensional grid. The relationship between output nodes and input nodes depends on their weight vectors. When training the network, the weight coefficient between nodes is trained to obtain the optimal weight vector.

2.4.4. **FCM clustering algorithm**

On the basis of general classification, J. V. Bezdek proposed a fuzzy clustering method based on membership degree and fuzzy set. FCM Clustering algorithm is to determine the relationship of data by calculating the membership value of data points to each cluster center, and the membership value is calculated by functions[13]. Compared with the traditional "hard clustering" algorithm, FCM algorithm has made a significant improvement.

The general principle of FCM clustering algorithm is as follows:

The set \( X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^s \) is Euclidean s Dimension data set, i.e. \( x_j = (x_{i1}, x_{i2}, \ldots, x_{is}) \). If \( X \) is subdivided into \( c \) kinds (2 \( \leq c \leq n \) ), there are \( c \) Cluster centers, \( V = \{v_1, v_2, \ldots, v_c\} \). The final result of clustering should make \( U \) is extremely small in the objective optimization function \( J(U, V) = \sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij}^m d_{ij}^2 \), in which \( u_{ij} \) represents the membership value of a sample data point \( x_j \) to cluster center vector \( v_i \), \( U = (u_{ij})_{c \times n} \) is a fuzzy partition matrix, which consists of the membership of sample \( x_j \) in the \( i \)-th membership, \( m \) is a fuzzy factor that determines the weight index of ambiguity, \( d_{ij} \) is the distance between a sample \( x_j \) and cluster center \( v_i \), \( d_{ij} = \|x_j - v_i\|^2 \) represents the Euclidean distance between a sample data point \( x_j \) and the first \( i \) Cluster-like center vector \( v_i \). From the above descriptions, it can be concluded that the FCM algorithm is transformed into The minimum solving process of \( J(U, V) \). The degree of membership is calculated:

\[
  u_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}^2(x_j, v_i)}{d_{ij}^2(x_j, v_k)} \right)^{1/(m-1)}} \quad (1)
\]

The calculation formula of cluster center is:

\[
  v_i = \frac{\sum_{j=1}^{n} (u_{ij})^m x_j}{\sum_{j=1}^{n} (u_{ij})^m} \quad (1 \leq i \leq c) \quad (2)
\]

Error \( e \) is:

\[
  e = \sum_{k=1}^{c} \|v_k - v_{k+1}\|^2 \quad (3)
\]

Specific algorithm steps are as follows:

1. Initialization parameters \( c \), parameters \( m \) and the value of error zeta;
2. Cluster center initialization: make the number of cycles \( k = 1 \), initialize the cluster center vector \( V = \{v_1, v_2, \ldots, v_c\} \);
3. Calculate the membership value according to the membership calculation formula (1);
4. According to formula (2), the clustering center value is obtained;
(5) Calculate the error $e$ according to formula (3), if the error $e$ does not fulfill the requirements, then go to step (3) until the requirements are met.

The FCM clustering algorithm is a classical clustering method with high accuracy and efficiency[14]. In this paper, FCM clustering method is used to determine the number of clusters that affect the valuation.

The spatial distribution of impact estimates to be clustered in the evaluation process is shown in Figure 2 below.

![Impact Assessment](image)

Fig. 2 Impact evaluation data to be classified

3. Cluster analysis of impact estimation based on FCM

3.1. Definition of Impact Valuation
Safety impact assessment value is an important index to measure automobile information security. Safety impact assessment includes six factors: safety, economy, operation, privacy, public safety and regulations, and hazard time. A detailed description of each factor is shown in Figure 3 below.
Fig.3 Influencing factors of impact assessment

Meanings of $S$, $F$, $O$, $Pri$, $Ps$ and $D$ are listed in the figure above. Calculation of the safety impact assessment value is as follows:

$$IL = a \cdot S + b \cdot F + c \cdot O + d \cdot Pri + e \cdot Ps + f \cdot D,$$

in which $a$, $b$, $c$, $d$, $e$, $f$ are fixed coefficient values. A Reference example:

$$0.952 \cdot Pri + 0.953 \cdot Ps + 2.381 + D \cdot 0.952 (10 \leq IL \leq 40).$$

In this paper, the impact evaluation formula is used to evaluate the information security risk of a certain vehicle.

3.2. Definition and evaluation criteria of clustering effectiveness

Generally speaking, as a good clustering result, the more obvious the difference between different classifications, the better the similarity within a class. The process of evaluating clustering effect based on this idea is the process of evaluating Cluster Validity[15].

A clustering structure $C$ is obtained by a certain clustering method from data set $X$. Generally speaking, clustering results can evaluate the evaluation effectiveness from two aspects[16]. First, the clustering structure $C$ can be evaluated by a structure-independent method. To analyze the data set $X$, a priori value is applied. In this kind of clustering evaluation, the criteria used are external criteria. Second, the quantitative evaluation and measurement of the clustering structure $C$ is made through its internal data set $X$. The criteria used in this kind of evaluation are called internal criteria.

3.3. Determine the optimal cluster number of influencing valuation data

Many scholars have done a lot of researches on determining the best cluster number[17]. Scholars consider determining the best cluster number through F statistics[18], which is a difficult problem in cluster analysis. Benefiting from the inspiration of variance analysis theory in mathematical statistics, F statistics can be applied to the theory of determining the best cluster number. The F statistics has a obvious deficiency: it is only applicable to one-dimensional samples. For the clustering analysis of multidimensional data, according to the theoretical knowledge of one-way ANOVA and based on F statistics, some scholars proposed to use mixed F statistics to solve the problem of optimal cluster number.

 Definition:
In which $n_i$ is the i-th number of samples in the category; $v_{ik}$ is the k-th cluster center of variables in the i-th Class sample set; $v_k$ is the the average value of each cluster center of the k-th variable; $x_{ijk}$ is the value of the k-th variable of the j-th sample. Theoretically, it can be proved that the above statistics $F$ obeys the $F$ Distribution with degree of freedom $(c-1,n-c)$:

$$F(k) = \frac{\sum_{i=1}^{c} n_i (v_{ik} - \bar{v}_k)^2 (n-c)}{\sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ijk} - \bar{v}_k)^2 (c-1)}$$ \quad (4)$$

$\bar{v}_k$ is the average value of each cluster center of the k-th variable.

Mixed $- F$ is the expression of mixed $F$ statistics, which can be proved by theory that the mixed $F$ statistics obeys the $F$ Distribution with degree of freedom $(c-1,n-c)$. According to formula 5, $F(k)$ contains the information of each variable and each category. The larger the value, the greater the similarity and correlation of variables in each category, while the smaller the similarity and correlation between each category. Mixed $- F$ It also contains the information of each variable and each category, and can also reflect the similarity and correlation of all variables in each category within and between classes. In order to show the impact of smaller $F(k)$ on clustering results, and in order to ensure higher classification accuracy, the reciprocal weighting method is adopted in Mixed $- F$.

Impact valuation is analyzed and calculated for each feature. The maximum evaluation objects contained in each Feature ID are shown in table 1 below. A total number of 84 Feature IDs were evaluated, and 338 impact valuations were obtained, see attached Table 1.

Table 1. Evaluation object of feature

| Feature ID | Evaluation object |
|------------|-------------------|
| TBA TSP   | TBOX               |
| BCP       | application       |
| OMP       | ion                |
| XT        | hive communication |
|           | CAN/Message sending |

Using FCM clustering method, the impact valuation data are classified into 2-12 categories in turn. The classified data are calculated using mixed $F$ statistics, and the mixed $F$ statistics are obtained. The mixed $F$ statistics under different cluster numbers is in fig. 4.
Fig. 4 Mixed F-statistics with different cluster numbers

It can be seen from fig. 4 that when the cluster number of observation indexes is divided into five categories, the mixed F-statistics reaches the maximum, and it is considered that the best cluster number is 5. That is, the impact valuation should be divided into five risk levels, namely, extremely low, low, medium, high and extremely high, which can best reflect the overall distribution of data.

4. Conclusions

In the application of FCM clustering algorithm, the impact estimates are divided into 2-12 categories. The mixed F-statistics can well reflect the space allocation. By calculating the mixed F-statistics of each classification, the conclusion shows that the mixed F-statistics reaches the maximum when the impact estimates are divided into five categories. This is of great significance to the establishment of risk assessment model and the preparation of risk assessment analysis.

References

[1] Wang Chen, Yao Wenye, et al. A preliminary study on the security of 5G smart car networking [J]. China New Communications, 2021,23(06).

[2] Zhong Yongchao, Yang Bo, Yang Haonan, et al. Overview of Intelligent Networked Vehicle Safety [J]. Information Security Research, 2021,7(06).

[3] Wang Jianqiang, Wang Xin. Intelligent Networked Automobile Architecture and Key Technologies [J]. Journal of Chang 'an University (Social Science Edition) .2017 (06).

[4] Yansong Li. Intelligent Networked Vehicle Security Threat Analysis and Remote Intrusion Research under Complex Network Environment [D]. xidian university 2019.

[5] definition of risk assessment: https://baike.baidu.com/item/%e9%a3%8e%e9%84%ae/%e8%af%84%e4%bc%b0/2785997?fr=aladdin

[6] GB/T 22239-2019. Basic requirements for network security level protection of information security technology [S]. 2019

[7] Cluster analysis [EB/ol]. https://baike.baidu.com/item/ cluster analysis /3450227.

[8] Data Structure [EB/OL]. https://baike.baidu.com/item/ Data Structure /1450? fr=aladdin.

[9] Park Shangzhe, Chao Mu Rilige, Yu Jian. Evaluation of clustering effectiveness of fuzzy C-means algorithm [J]. Pattern Recognition and Artificial Intelligence .2015 (05).
[10] summary of clustering algorithms [J]. Zhang yonglai, Zhou yaojian. computer applications .2019 (07).
[11] Cluster center initialization algorithm for K-means clustering[J]. Shehroz S. Khan, Amir Ahmad. Pattern Recognition Letters . 2004 (11).
[12] Tian Bing et al. Rock burst prediction based on SOM neural network clustering and gray TOPSIS evaluation method. China Mining Industry.2021.20 (21).
[13] Collaborative filtering recommendation algorithm based on improved fuzzy partition and clustering [J]. su Qing, Zhang jingfang, Lin zhengxin, Li Xiaomei, Cai zhaquan, Zeng yongan. computer engineering and application.2019 (05).
[14] Bezdek J C. Pattern recognition with fuzzy objective function algorithms /[M]. Plenum Press, 1981.
[15] Study on the Effectiveness of Fuzzy Clustering [D]. Geng Jiayi. Jiangnan University, 2019
[16] Cluster index [EB/OL]. https://blog.csdn.net/Haiyang_Duan/article/details/77980387.
[17] Fuzzy Correlation and Clustering Effectiveness [J]. Fan Jiulun, Pei Jihong, Xie Weixin. Journal of xidian university. 1998(01)
[18] Sun Caizhi, Wang Jingdong, Pan Jun. Study on the method of determining the best cluster number in fuzzy clustering analysis [J]. Fuzzy Systems and Mathematics, 2001(01):89-92.