Data Mining Strategy of Information Fingerprint Self-learning Mechanism

Liu Xin[1], Wang Wenting[1], Chen Jianfei[2], Wang Rui[1], Zhang Hao[1]

[1]State Grid Shandong Electric Power Company Electric Power Research Institute, Jinan City, Shandong Province, 250003

[2]State Grid Shandong Electric Power Company, Shandong Jinan, 250003

Keywords: data mining; data preprocessing; data reduction; dimension reduction; feature extraction; fingerprint recognition

Abstract: Data mining needs various analysis of data, and data preprocessing is needed before all analysis. However, after data cleaning, data integration and data transformation, the data set will still be very large! It will take a long time to analyze and mine complex data directly on massive data, which makes this analysis unrealistic or impractical. Data reduction technology can be used to get the reduction representation of data sets, which is much smaller, but still close to maintaining the integrity of the original data. In this way, mining on reduced data sets will be more effective and produce the same or almost the same analysis results. Through this method, the most representative features are extracted from a large number of features and useful information is analyzed according to needs. With the development of society, traditional security systems based on tokens or passwords are becoming more and more fragile, which can not meet the needs of modern security systems. The fingerprint recognition based on feature extraction comes into being. Two distinct features, endpoint and bifurcation point, are extracted from numerous fingerprint attributes for data mining and analysis.

1. Introduction

In data mining, various kinds of data analysis are needed, and data preprocessing is needed before all analysis. However, after data cleaning, data integration and data transformation, the data set will still be very large! It will take a long time to analyze and mine complex data directly on massive data, which makes this analysis unrealistic or impractical. At this time, data reduction technology is particularly important. Data cube aggregation, dimension reduction, data compression, numerical compression, discretization and conceptual hierarchical generation strategies of data reduction technology are used to reduce data sets and maintain the integrity of the original data. In this way, mining on reduced data sets will be more effective and produce the same or almost the same analysis results. Through this method, the most representative features are extracted from a large number of features and useful information is analyzed according to needs. Feature extraction in data mining is widely used, and fingerprint recognition is the most typical application.

2. Data Mining Strategy

2.1 Dimensional reduction

Data used for data analysis may contain hundreds of attributes, most of which are not related to mining tasks and are redundant. Although domain experts can pick out useful attributes, this can be a difficult and time-consuming task, especially when the behavior of data is unclear. It is harmful to omit relevant attributes or leave irrelevant attributes, which may slow down the mining process. Dimension reduction reduces the amount of data by deleting irrelevant attributes or dimensions. Generally, feature extraction is used to select a subset of attributes.

2.2 Concept of feature extraction

Feature extraction is to compress high-dimensional attribute space into low-dimensional attribute

  DOI: 10.25236/isrme.2019.106

space by mapping, and get the smallest set of attributes, so that the conceptual distribution of data classes is as close as possible to the original distribution using all attributes. The results of data mining are similar to or completely consistent with those of data mining in which all features participate. For D attributes, there are 2 possible subsets.

2.3 Feature Extraction Analysis

It may be unrealistic to find the best subset of attributes by exhaustive search, especially when the number of D and data classes increases, so the heuristic algorithm of compressed search space is usually used for feature extraction. The basic heuristic methods of feature extraction include the following four kinds:

(1) Step by step forward selection: the process starts with empty attribute set as reduction set, determines the best attribute in the original attribute set, and adds it to the reduction set. In each subsequent iteration, the best attributes in the remaining set of original attributes are added to the set.

(2) Gradually delete backwards: the process starts with the entire set of attributes. In each step, delete the worst attributes in the attribute set.

(3) The combination of progressive forward selection and progressive backward deletion: we can combine progressive forward selection with progressive backward deletion, select one of the best attributes at each step, and delete one of the worst attributes in the remaining attributes.

(4) Decision tree induction: Decision tree induction constructs a flow chart-like structure in which each internal (non-leaf) node represents an attribute test, each branch corresponds to a result of the test, and each external (leaf) node represents a class prediction. At each node, the algorithm chooses the "best" attribute and classifies the data.

3. Fingerprint feature analysis

In real life, people's identities are often identified, such as identifying the boarder when boarding an airplane, verifying whether the money withdrawer is the legal owner of a specified account, checking the operator's authority when using a computer, and so on. With the rapid development of national economy and social informationization, financial institutions, government agencies, enterprises and individuals are increasingly closely linked through the Internet. On the one hand, it provides conditions for information sharing, on the other hand, it also provides opportunities for people with bad intentions to illegally access other people's information. Therefore, how to automatically, accurately and efficiently identify people's identity is an important issue in the field of information security. Fingerprints have uniqueness and stability, so they are used as the main basis for identifying a person. However, how to accurately find fingerprints representing a person in thousands of fingerprint databases for data mining, which requires the use of fingerprint identification technology. Compared with other biometric identification technologies, fingerprint identification is an ideal identification technology.

At present, there are many kinds of fingerprint feature points extraction algorithms, such as fingerprint feature extraction algorithm based on direct gray image, image feature extraction algorithm based on thinning, fingerprint feature extraction algorithm based on ridge direction filtering, fingerprint feature extraction algorithm based on binarization, and so on. Generally speaking, there are two kinds of most commonly used feature points extraction algorithms: one is to extract feature points directly from gray images; the other is to extract feature points from pre-processed and thinned images.

Based on the application of feature extraction in fingerprint identification and analysis technology, it can be seen that after a series of feature extraction, the original multiple fingerprint attributes are reduced to the minimum, but still representative. It not only completes the recognition more accurately, but also greatly reduces the workload. Obviously, the importance of feature extraction in data mining and its importance can not be ignored.

4. Conclusion

It will take a long time to analyze and mine complex data directly on massive data, which makes this analysis unrealistic or impractical. Therefore, data reduction is needed, and special name extraction is particularly important at this time. After data mining, the results of the extracted data sets are not only similar to or completely consistent with the data mining results of all features, but also greatly reduce the workload, so it is widely used in real life. With the development of society, traditional security systems based on tokens or passwords are becoming more and more fragile, which can not meet the needs of modern security systems. The fingerprint recognition based on feature extraction comes into being. Two distinct features, endpoint and bifurcation point, are extracted from numerous fingerprint attributes for data mining and analysis.

References

[1] Medeiros J P S, Brito A M, Pires P S M. A Data Mining Based Analysis of Nmap Operating System Fingerprint Database[C]// International Workshop on Computational Intelligence in Security for Information Systems. 2009.

[2] Shih C C, Chiang D A, Lai S W, et al. Applying hybrid data mining techniques to web-based self-assessment system of Study and Learning Strategies Inventory[J]. Expert Systems with Applications An International Journal, 2009, 36(3):5523-5532.

[3] Tong Y. Research on Data Mining of Complex Multi -dimensional Fingerprint Data of TCM[J]. New Technology of Library & Information Service, 2011.

[4] Othman M L, Aris I, Abdullah S M, et al. Knowledge Discovery in Distance Relay Event Report: A Comparative Data-Mining Strategy of Rough Set Theory With Decision Tree[J]. IEEE Transactions on Power Delivery, 2010, 25(4):2264-2287.

[5] Nagaty K A. On learning to estimate the block directional image of a fingerprint using a hierarchical neural network.[J]. Neural Netw, 2003, 16(1):133-144.