International Core Journal of Engineering Volume 5 Issue 9, 2019
ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0020

A Survey of Data Preprocessing in Data Mining

Chenggang Zhen[1, a], Yingxin Zhang [2, b]

[1]School of North China Electric Power University, Baoding 071000, China;

[2]School of North China Electric Power University, Baoding 071000, China.

[a]zhencg@163.com, [b]609741160@qq.com

Abstract

With the increasing amount of data, data preprocessing has become an indispensable part of data mining. This paper introduces the data preprocessing process and enumerates the methods of data preprocessing related processes in the existing literature.

Keywords

Data mining, Data preprocessing, Data quality.

1. Introduction

With the arrival of the era of big data, the amount of data is increasing day by day, and the order of GB and TB cannot meet the scale of data processing. However, these data contain a lot of redundant information, making it difficult for people to extract useful information from the massive data. Therefore, data mining comes into being. By integrating and analyzing the data, data mining transforms the original massive data sets according to certain algorithms, which can achieve the extraction of internal effective information of the massive data, so as to provide support for future decisions. Data preprocessing is the most important link in the mining work. This paper mainly introduces the data preprocessing process.

2. Relevant concepts

2.1 The concept of data mining

Data mining is capable of extracting information of real value that is difficult for people to obtain in a conventional manner from a large data set of various forms according to a certain algorithm [1]. The tasks of data mining mainly come down to feature rules, distinguishing rules, classification, relevance, clustering, prediction, and analysis of changes and deviations. Data mining generally needs to go through the analysis of problems and clear the ultimate purpose of mining. Through the preprocessing of the data in the mass data set, the preprocessed data will be stored, and finally presented to users in a visual form, which can better support the user's decision.

2.2 The concept of Data preprocessing

Data preprocessing is a series of data processing work for obtaining high quality data sets.The most important step of data mining is to preprocess the data. Data mining uses a wide range of data sources, including databases of different forms and many heterogeneous data sources. At present, the data in the database are easily interfered by noise, data collection equipment failure, human or computer internal errors, so it is very necessary to preprocess the data.The relevant literature on preprocessing indicates that the data preprocessing work needs to occupy 60% of the entire excavation work, and the quality of data preprocessing plays a direct and decisive role in data mining quality and efficiency

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0020

[2]. In this paper, data preprocessing is introduced in four methods: data cleaning, data integration, data transformation and data protocol.

2.3 The concept of data quality

Data quality is to guarantee the correctness of logic of the data and remove the problem and repeated data, data quality is the key to complete the final support data analysis and decision factors, in the relevant references in the definition of data quality has a different, one is for the final analysis of the data available, one is the enterprise to the degree of trust and meet their expectations. High quality data sets resulting from a series of processes are often beneficial for maximum utilization. Generally, data quality includes four elements, namely, accuracy, integrity, consistency, and timeliness, which are used to describe data quality.

3. Data preprocessing method

3.1 Data cleaning

The process of data cleansing is very important for transforming complex and heterogeneous big data into a complete, high-quality and reliable data set that can support decision making. Dirty data is common. Due to human or machine reasons, there will be missing data and inconsistent data, etc. Data cleaning is to process these dirty data to ensure the data quality of mining work.

3.1.1 Missing data processing

Missing values are a common problem when data is processed. Usually, there are many reasons for data missing values, such as device failure or artificial avoidance of privacy. Usually ignoring missing data is the easiest way to handle missing values, but this method is only suitable for cases where there are fewer missing values. If you use this method with too many missing values, Losing a lot of important data may bias the final result prediction. At this stage, there is a method of surrogate value filling. In the method of substituting value filling, the method of mean and median filling is included. Which method is used for filling needs to be determined by looking at the actual situation of the data. From the literature in recent years, there are more and more researches on the processing of missing values. Zheng Yashi [3] proposed a suitable radar data based on data cleaning based on regression interpolation and ordinary least squares method. Missing value cleaning method, which can select known information from the sliding record window according to the correlation of radar data correlation, and then predict the missing value by linear regression relationship between attributes. Yin Zhentao[4] used the idea of iteration and combined with the characteristics of panel data, based on this, proposed a method based on factor model filling, which can use the panel data with normal distribution to construct different missing proportions. Data set, then apply the corresponding model to the data set for missing value filling.

3.1.2 Noise data processing

The noise data generally refers to the random error generated by the measured data, including erroneous data and outliers. These errors can be caused by equipment failures, data transmission problems, or inaccurate calculations. Error data can be processed by noise filtering. Common noise filtering methods include regression method, mean smoothing method, outlier analysis, wavelet method, etc. [5]. For the real world noise elimination problem Hao Shuang [6] et al. summarized the data repair methods, including data cleaning based on integrity constraints, rule-based data cleaning and data cleaning combined with man and machine. Chen Jianming [7] proposed an algorithm based on EEMD data preprocessing and DNN speech enhancement for the problem of poor speech transmission quality under noise background. The EEMD decomposition and feature extraction are used to distinguish the noise component and the speech component, which can be excluded. Interference noise information. Zhang Xijun [8] et al. combined the characteristics of trajectory big data non-stationarity, and adopted a two-dimensional discrete wavelet method to complete the deconvolution and data compression, in order to deal with traffic trajectory big data.

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0020

3.1.3 Inconsistent data processing

The types of databases used in mining vary, which leads to different data sets. The diversity of data storage forms leads to inconsistent data in the real world. Inconsistent data is caused by a violation of the original integrity rules for some reason. If it is not processed, it can directly lead to over-fitting on the classification result, which will affect the final visualization. Wang Hejun [9] accurately classifies the inconsistent data in the dataset through the improved classification algorithm. For the high-dimensional inconsistent dataset, the feature selection algorithm is used to reduce the dimension of the data, and then the influence of the influencing data feature influence factor is used to influence the feature. Larger subset classification modeling. Inconsistent data Zhang Anzhen [10] designed an inconsistent data detection and repair algorithm based on Hadoop parallel platform, and detected inconsistent data sets according to the conditional function dependence in the data dependence principle.

3.2 Data integration

In order to minimize redundant data in the data set, data from multiple databases is usually combined into one database to reduce resource waste. The most common methods in data integration are entity identification and data collision detection processing.

3.2.1 Entity recognition

Massive data sets often contain heterogeneous data sources. An important task of data integration is to identify records from multiple different data sources that represent the same object. This process is entity identification. Zhang Fulin [11] proposed a token-independent block-based block algorithm based on the advantages and disadvantages of existing block algorithms. At the same time, it proposes a block-based intra-block redundancy based on Meta-blocking technology. A pruning algorithm based on cumulative weights, which greatly increases the speed of entity recognition at the expense of a small amount of benefit.

3.2.2 Data collision detection

For entities in the real world, different databases use different rules, so there will be different forms of difference. Data conflicts usually include semantic level conflicts and pattern level conflicts. We need to detect and resolve these conflicts during data processing. Wen Jing [12] proposed different solutions for semantic conflicts at the pattern level and semantic conflicts at the data level, using owl to express the relationship between ontology and the mapping between ontology and database schema and solve the pattern level. Semantic conflict, based on the semantic conflict ontology DLCO extended database schema, proposes detection algorithms based on ontology and underlying database semantic model to solve conflicts at the data level.

3.3 Data transformation

In general, data from different databases will be different due to different database rules. Data transformation is needed to transform data into the same type. In addition, the data transformation is still to reduce the dimension of the existing high-dimensional complex data, and obtain the characteristic representation of the data when the number of effective variables is reduced. The general data transformation dimension reduction model proposed by Wu Xinling et al. [13] compresses p primitive variables into p' variables through data transformation. The corresponding data transformation examples prove that this method uses the fewest variables to capture the largest. The amount of data is possible. Li Hongli [14] used the knowledge of related rough sets to construct a granulating model based on tolerance relationship and compatible particles to quickly obtain any attribute column, which can realize rapid reduction of attributes.

Data protocol

Massive data sets are very inconvenient for mining work. The data specification is to streamline the data with the largest specification in the case of minimizing the loss of the original data amount, and greatly improve the efficiency of mining for the data set after the specification. Wang Weijun [15] et

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0020

al. proposed a dimension reduction method based on the cumulative proportion of feature combinations and a dimension reduction method based on clustering algorithm. The former concentrates information on some major feature combinations, and selects the first few cumulative frequencies greater than the threshold. K's sorting method is combined to achieve the purpose of dimensionality reduction. The latter is to divide the data with similar information into one class by the difference between the clustering methods to achieve the purpose of dimensionality reduction.

4. Conclusion

Data preprocessing is an important part of data mining work. Now data mining technology has been widely used in various fields, and the preprocessing has attracted the attention of scholars. More and more researches in this field have been carried out. To use various methods of data preprocessing to complete the analysis of data, the platform of big data analysis and decision should be gradually improved, and the optimization algorithm of data processing should be continuously improved, so as to better complete the data analysis and processing and provide better decision support for people.

Acknowledgements

First of all, I want to thank my mentor zhen fresh-from-stats-class teacher, zhen gave me a lot of inspiration when the teacher in the selected topic, guide me out when I meet with difficulties, also want to thank the classmates, they complete the whole article offers help to me, finally I would also like to thank the related literatures, the authors, their research gave me help on writing.

References

[1] Yang Xiugang.Overview of Data Mining Algorithms[J].Science and Technology Economics Guide,2019,27(05):166.

[2] Liu Li,Xu Yusheng,Ma Zhixin.Overview of Data Preprocessing Technology in Data Mining[J].Journal of Gansu Science and Technology,2003(01):117-119.

[3] Zheng Yashi. Research and Application of Data Cleaning in Multi-Radar Data Fusion Algorithm[D].Beijing University of Posts and Telecommunications,2018.

[4] Yin Zhentao. Research on Random Missing Value Filling and Its Effect[D].Shanghai Normal University,2018.

[5] Zhao Yifan, Bian Liang, Cong Xin. A Review of Data Cleaning Methods[J].Software Guide, 2017, 16(12): 222-224.

[6] Hao Shuang, Li Guoliang, Feng Jianhua, Wang Ning.Overview of Structured Data Cleaning Technology[J].Journal of Tsinghua University(Science and Technology),2018,58(12):1037-1050.

[7] Chen Jianming,Liang Zhicheng.Study on speech enhancement algorithm based on EEMD data preprocessing and DNN[J].Journal of Ordnance Equipment Engineering,2019,40(06):96-103.

[8] Zhang Xijun,Yuan Zhanting,Zhang Hong,Gao Weijun,Zhang Enzhan.Study on the preprocessing method of traffic trajectory big data[J].Computer Engineering,2019,45(06):26-31.

[9] Wang Hepeng. Research on classification algorithm of massive inconsistent data[D].Harbin Institute of Technology, 2017.

[10] Zhang Anzhen, Men Xueying, Wang Hongzhi, Li Jianzhong, Gao Hong.Hadoop-based inconsistent data detection and repair algorithm based on Hadoop[J].Journal of Computer Science and Technology,2015,9(09):1044-1055.

[11] Zhang Fulin. Research on Entity Recognition Technology for Heterogeneous Big Data Integration[D].Beijing University of Posts and Telecommunications,2018.

[12] Wen Jing. Research on Data Conflict Detection and Solution in Data Integration[D].Shandong University,2010.

[13] Wu Xinling,Wu Guoqing.Dimension Reduction Method Based on Data Transformation[J].Journal of Wuhan University(Natural Science Edition),2006(01):73-76.

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0020

[14] Li Hongli. Research on the filling of incomplete and inconsistent data and its attribute reduction algorithm [D]. Guangxi University, 2018.

[15] Wang Weijun, Jing Hao, Gou Na, Yang Jinhao.Study on feature set analysis method based on multi-selection problem data protocol[J].Journal of University of Electronic Science and Technology of China(Social Sciences Edition),2018,20(01):66-70.