Research on the Key Technology of Mixed Attribute Data Cluster based on Sparse Representation

Huifang Xue

Xi'an International University, Xi'an, China

Keywords: sparse, mixed attribute, data cluster

Abstract: Data mining is one of the most important tools to provide assistance in management decision. As the application range of data mining is enlarging, the processed data of data mining have gradually transformed from single numeric type or classified-type data to mixed attribute type. The research on mixed attribute data mining is a hot topic, wherein, data cluster is an important part. Traditional clustering algorithm only aims at single numeric attribute or classified attribute, while more and more researches show that most real data are presented in mixed attribute, which makes it difficult in processing most traditional clustering algorithms. Therefore, designing efficient clustering algorithm which can process mixed attribute data has become an attractive topic in cluster analysis.

1. Introduction

Data refer to the records to describe the information characteristic of things, such as commodity sales data, patient information data, financial expenditure and income data, all daily information can be called as data. In the current big data era, under the backgrounds of increasing means for humans to know the world and comprehensive integration of computer technology in people's daily life, the data scale which can be cognized by humans has presented a growth trend of geometrical progression. The global commercial activities have generated numerous data, including stock trading; the scientific research work has also produced a large number of data from experiment records, engineering observation and environment monitoring; while social media such as microblog, WeChat, blog and forum have already become the important data sources. Therefore, how to analyze and process the numerous data better and transform them to be the direct knowledge resources is always the topic needing to be researched. In 1995, the Association for Computing Machinery had proposed related theories and frameworks of data mining that was the process of Abstracting potential, effective, useful, novel, explainable and understandable from numerous, incomplete, noisy and fuzzy data. Data mining is the most important step in the database knowledge discovery process, which is a complicated process Abstracting unknown and valuable mode from a large amount of data. It involves the fields related to mode recognition, statistics, machine learning and data visualization, has higher requirement on related technologies and tends to be difficult to cope with the actual demand of data explosion.

2. Mixed attribute data cluster

At present, the existing data clustering algorithms are limited by the data featured with continuous attribute, in addition, some algorithms can only handle the limitations of the data featured with classified attribute. In real world, most data have two kinds of attributes at the same time, if only one kind of attribute is processed, data information will be lost under the situation of mixed attribute thus to affect the quality of data mining. It is still a challenging task to realize clustering of mixed data. However, there are still a large number of researches on single-type attribute, therefore, it is a natural choice to promote these good clustering algorithms to solve the mixed attribute data cluster. In specific, there are three methods to cope with the mixed-type data clustering algorithms:

Firstly, disperse the numeric-attribute data of mixed-type data to be classified attribute data only,

  DOI: 10.25236/iwmecs.2019.020

and then conduct clustering by using the clustering algorithms suiTable to classified attribute data only. The numeric discretization method mainly include overall discretization and partial discretization, which can also be divided into non-supervision discretization and supervision-based discretization. Wherein, the same-frequency split method and same-width split method are most widely used in the former one. Wherein, the same-frequency split method is to divide attribute value domain into several intervals, and then the users determine the quantity of object in each interval. The same-width split method refers to that the users determine the quantity of intervals, and then they divide the attribute value domain into the intervals with same width. But the above two methods exist with certain defect, wherein, the same-frequency split method is sensitive to the objects with misjudged ClassID. For example, one certain object is close to the characteristic value of one kind of data, but the ClassID doesn't belong to this kind of data, which will cause wrong classification result. The same-width split method is sensitive to abnormal value, when one certain abnormal value is far from most data points, the abnormal value will be split to independent interval in order to guarantee the split width to be same, which will be easy to destroy the decision structure. In general, the numeric attribute discretization process will be easy to cause information loss and error thus to largely influence the effect of data mining.

Secondly, transform the classified attribute date in the mixed-type data to be numeric attribute data only, and then conduct clustering by using the clustering algorithms suiTable to numeric attribute data only and conduct processing by using the similarity measurement function suiTable to numeric date only. The most common similarity measurement function method is Euclidean distance, but the difficulty of this method lies in transforming the classified data to be numeric data correctly.

Thirdly, solve the mixed data clustering problem through redesigning the clustering cost function of existing clustering algorithms, wherein, the new clustering cost function can measure continuous attribute and classified attribute at the same time. For instance, combine with the above two methods, adopt different similarity measurement methods for numeric data and classified-type data. For the two numeric data with concentrated data, adopt Euclidean distance to be the similarity measurement; for the classified-type data, use sign function for measurement, when the two data are completely same, the distance will be 0, otherwise, the distance will be 1. Finally, accumulate the similarity measurement results of different types of data.

3. Key technology of mixed attribute data cluster based on sparse representation

3.1 Imputation method based on statistical theory

3.1.1 Mean value imputation method

The mean value imputation method refers to calculating the non-missing mean value of each numeric attribute variable (interval variable) in all objects, and taking all mean values as the imputed value of all missing attribute values; for the non-numeric attribute (binary attribute, ordinal attribute and classification attribute), calculate the mode with highest frequency in other objects to make up for the missing data. This method bases on the situation of the attribute value obeying to the normal distribution or approximately normal distribution, if the distribution is skewed, use the medium value for imputation. The mean value imputation method is characteristic by simpleness and easiness, besides, the univariates such as mean value and total quantity can effectively reduce the deviation of the point estimation, while its defect is obvious as well. Firstly, the imputation result changes the variable distribution information in the sample unit, because the imputed value of the missing data is replaced by the mode or median, which causes the distribution situation to be restricted by the mean value calculated from the observation data; secondly, the imputation result will cause underestimation on variance from mean value and total quantity, because the deviation of the sample unit will decline owing to multiple appearances of same numeric value, because the mean value imputation is only suiTable to simple point estimation not complicated analysis on variance estimation.

3.1.2 Conditional mean imputation method

The largest difference of conditional mean imputation method from the mean value imputation method lies in that the obtained mean value is not from all real examples but the missing data containing classification data information, which is obtained from mean value calculation of attribute values of similar real examples, and belongs to the imputation method using label information to supervise missing data.

3.2 Imputation method based on machine learning

3.2.1 Imputation method based on K Nearest Neighbor (KNN) and related algorithm

In view of missing data imputation, the mature K Nearest Neighbor method with good robustness is widely used based on machine learning. In the machine learning, K Nearest Neighbor method persists in the basic principle: If one certain sample belongs to one certain category among k nearest (namely nearest in characteristic space) samples in characteristic space, the sample will belong to the sae category. Based on KNN algorithm, the k nearest neighbors are all objects having been classified correctly. This method only bases on one or several class labels in the nearest sample to determine the category of the sub-samples.

The missing data imputation method based on KNN was firstly proposed by KNNimpute in researching the missing data imputation related to gene, wherein, the specific idea is as follows: 1) Calculate the distance between the object having missing data and the object having complete data without missing attribute value, wherein, different similarity measurement functions can be used; 2) Select k data objects with minimum distance, use the distance weighted mean value of the k objects in each missing attribute to estimate the missing value.

3.2.2 Missing data imputation method based on K mean value and other clustering processes

In addition to imputation estimation on the missing data base on K Nearest Neighbor, various cluster methods have also been applied in missing data imputation process by scholars. KMI, similar to KNNimpute, divides the pending data to be complete data and missing data. Firstly, it conducts K-Means clustering on the complete data, and then distribute the missing data objects containing missing attribute value to the nearest class according to the distance between the class center and missing object, afterwards, it conducts estimation on missing attribute value based on the cluster center. This method can effectively decline the calculation complexity of KNNimpute and estimate all missing values of one certain missing object.

4. Conclusion

In summary, data mining and knowledge discovery have been widely used in intelligent commerce and daily life, while cluster, as one of the most important tasks for data mining, has been the hot topic for researchers. Most clustering algorithms in data mining are established based on single-type data. But in reality, most data include numeric attribute and classification attribute. Owing to large difference between the numeric attribute and classification attribute, the clustering algorithm which can process mixed-attribute data is urgently needed.

References

[1] He Yan. Research on Bayes' non-parametric modelling method in statistical sparse learning and its application [D]. Zhejiang University, 2012.

[2] Feng Xiaodong. Research on non-supervision mining of high-dimensional data based on sparse representation [D]. University of Science and Technology Beijing, 2015.

[3] Xiao Liang. Research on data classification and cluster algorithms and corresponding application based on sparse representation [D]. National University of Defense Technology, 2012.

[4] Yang Guoliang, Xie Naijun, Wang Yanfang, Liang Liming. Non-supervision feature selection based on low-order and sparse scoring [J]. Computer engineering and science, 2015, 37(04):649-656.