Research On Large outliers in the data set data mining algorithm

Jinhai Zhang

Marine college of Shandong jiaotong university, Weihai , Shandong,China

Keywords: Data mining, outlier detection, outlier analysis, clustering, classification

Abstract. Main purpose of outliers mining is from a large number of, incomplete, there are all kinds of data, the found hidden in one of the people is not known in advance but potentially valuable information or knowledge. Outlier is a data: deviate significantly from other data, it does not meet the general patterns or behavior. Outlier data mining has been widely used in the stock market, telecommunications, financial services, intrusion detection, weather forecasting and many other fields. Outliers may be "noise", but it may also be significant events. In practice, in some applications, those rare events are likely to have more value than events that occur frequently. Unfortunately, the outlier data mining is a very important and meaningful work.

Introduction

With the popularization of computers, the rapid development of database technology and the wide use of a variety of database management systems, after years of effort, all areas are accessible through your own database management system accumulated huge amounts of data. And these huge data, timeliness and complexity far exceeding the current information processing capabilities.

People look forward to make use of these data, since they get the answer they wanted, turning data into knowledge in order to obtain useful information, summary or forecast. But on the other hand, database systems with different color just to stay in the input data, queries, data and statistics, there is no way for effective data analysis, then hidden in this large amount of data we are interested in some of the information cannot be found, of course, cannot be said to anticipate future trends.

Data mining is the world-wide focus on investment research, one of the top ten technologies. Of course, the study of data mining technology without the development of related technologies and applications:

- improving your computer's performance and advanced architecture development;
- the development of databases, data warehouses, and information technology;
- statistics and artificial intelligence methods in research and application of data analysis.

Practical needs and the development of related technologies and database technology to manage data storage, through machine learning method to analyze the data, and then dig out useful information hidden in the data, it was this idea, formed in the deep concern of the people of a new field of research: knowledge discovery in databases. In order to avoid the serious waste of resources and data mining technologies. How to make full use of these large amounts of data, discovering the hidden message, and summed up the potential of knowledge and/or rules has become a data mining is active in the areas of research topics.

Outlier data mining based on knowledge

As one of the basic tasks in data mining, outlier detection focus not consistent with most of the objects in the dataset, or significant deviation from the General object of a small portion of the data. This fraction of the data can affect the analysis of the data should be identified and removed before the analysis. But this view is not comprehensive, outliers may be due to measurement error, computer input error, human error, or there may be inherent data variation. Former bad data in the analysis of data mining should be modified or deleted, otherwise it will affect the analysis results. But for real bias inherent data, these values may be for some people is the difference between noise and information which may be useful for other people, and in fact in some practical applications, rare events have more value than ordinary events.

Outlier data mining tasks can be divided into two subtasks: the first is to determine what kind of data is in the DataSet objects that we are looking for: stray; then is to find an effective way to detect these our first step to identify outliers. First a task determine away from group data actually is not a simple of work, must specific problem specific analysis, from different of angle will away from group data and general data for compared, significantly difference Yu general data, in different situation Xia away from group data is significantly different of, in different of situation Xia same away from group data by contains of meaning is completely different of, like Dang processing of object is a time sequence data collection Shi, away from group data defined on compared tricky, because they has may hidden in season, and cycle or trend of changes in the.

Outlier data mining and data warehousing data mining development has a close relationship with the data warehouse, data warehouse technology development is one of the reasons for promoting the development of data mining techniques. Data mining can be seen as advanced stages of OLAP in data warehouse. But is a more advanced data analysis techniques, data mining, data warehouse summary-much more in-depth and detailed analysis. Data warehouses are not essential for data mining, which means that data mining is not necessary to have a data warehouse supports.

Data mining is conducted in real-world data, incomplete, noisy and inconsistent data in the real world of large databases or common features of the data warehouse. Cannot be made directly on the data such as data mining, the corresponding processing technology was born. Including data preprocessing, data sampling techniques and the clustering of data.

Outlier mining algorithm overview

With growing awareness of the importance of outliers mining and its increasingly wide range of applications, a growing emphasis on outlier mining of outliers mining methods are endless, most known mainly in the following ways:

Outlier data mining based on statistical methods: assuming the data sets that meet specific distribution models such as the normal distribution, distribution according to the inconsistency tests on each point in the dataset if the distribution does not meet, they think it is an outlier.

Outlier data mining method based on distance: if part of the data in a DataSet object data with the object distance is greater than and refer to objects as outliers.

Outlier data mining method based on deviation: determine if a data is the basis for outliers is produced by removing the data in the data through the data changes and impacts.

Rule-based outlier mining approach: using data mining technique (such as Association Rules), once a sample set of rules has been established, then uses these rules to verify the outlier.

Outlier data mining based on clustering method for data clustering method using clustering more distance from the clustering of data is more likely to be outliers.

Density-based outlier mining method: a method of outlier mining based on distance set up on the basis of density refers to the number of data objects within a range, based on this data to determine if a point is an outlier.

Distance based method: if part of the data in a DataSet object data object is greater than the distance between the objects as outliers.

Outlier data mining work consists of two parts: outlier detection and outlier analysis. Outlier detection currently is the hot spot, as has been done about; from the results of the data analysis is based on outlier detection for further analysis and study, and useful rules or conclusions drawn. Outlier detection was not the ultimate goal, the ultimate goal is the second part, the analysis and research of these outliers cause and possibly hidden knowledge.

Detected following the outliers, you first need to determine what these data are from the Group and analysis of its causes, finally judge the value of a hidden or law.

Outlier data mining based on two clustering algorithm

Cluster, and then to the next step. However, their clustering is not the same, comparative analysis is given below:

First, the purpose of clustering is different. Division algorithm based on cluster's goal is to reduce (), by calculating the Division will not be able to contain the outlier removed so as to reduce the amount of computation; algorithm based on clustering, clustering is a class in order to calculate, instead of the entire data set.

Second: the selection of clustering methods are not the same. Based on partitioning algorithm is selected under an efficient clustering algorithm for large data sets, it only needs to scan data sets and for the data set has a linear time complexity for clustering data elements with similar behavior is very effective.

Third: the cluster number is different. For partition algorithm, each partition contains data that are small, so divide the number so that each Division the last contains a data clustering algorithm based on, in order to ensure the point of nearest neighbors in the class, then the class contains data that is far greater than, so the fewer number of clustering.

Then integrated the advantages of both, it has been based on two clustering algorithm KNN outliers mining.

First of all, we do an experiment in the synthesis of a two-dimensional data set respectively based on clustering algorithm and based on secondary clustering algorithm to calculate. We set the number of neighbors k to 10, you need to find the outlier number n set to 12, finally pinpointing the 12 outliers, as shown in Figure 1

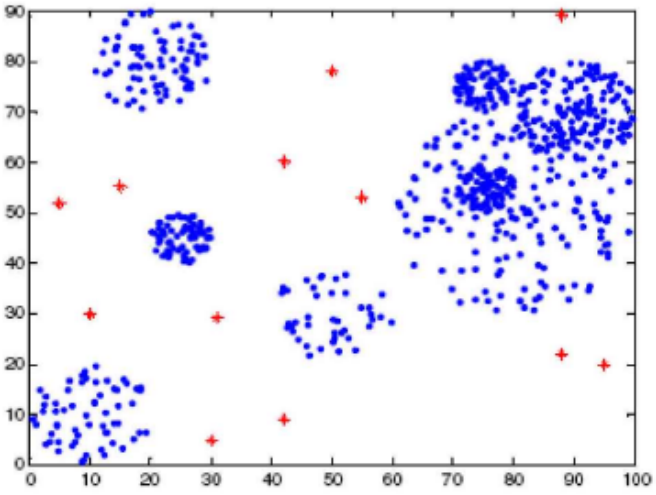

Fig.1. Outlier detection results on two data results

In the second experiment, we used simulated data sets, data on the probability of this data set is the same. Data n size from between 100000 and 500000. Compared with the partition algorithm, execution time is shown in Figure2.

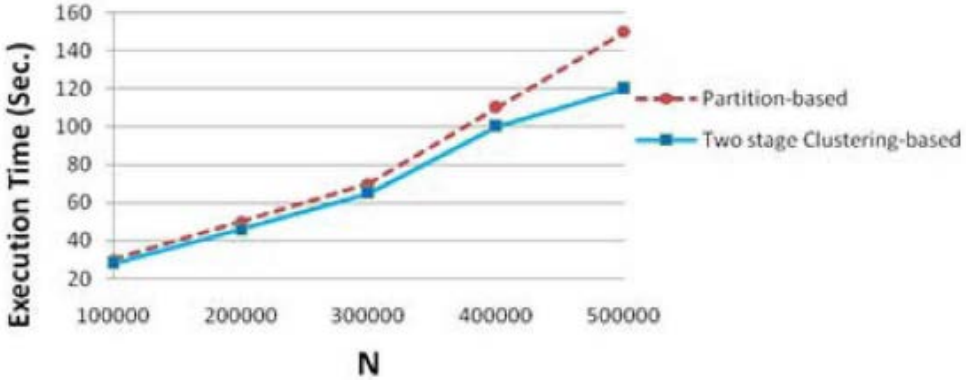

Fig.2. The experimental results for total n

From figure in the can see, based on two times poly class of algorithm with based on divided of algorithm phase compared, although two algorithm in total of data volume increased of when implementation time are almost is linear growth, but based on two times poly class of algorithm growth of range smaller, implementation time less, growth of slope than based on divided of to small some, which experiment results description, based on two times poly class of algorithm in increases of when, implementation time growth is is considerable of. This is because we will have a great deal of data clustering two, which both reduces I/O reduces the amount of computation.

Experiments in order to verify the efficiency and scalability of the algorithm for d, we used simulated data sets, data on the probability of this data set is the same. From 10000 to 50000 n the size of the data volume, dimensions ranging from 10 to 30, the execution time as shown in Figure 3 (the horizontal axis represents the data sets the amount of data, the vertical axis represents time unit: seconds).

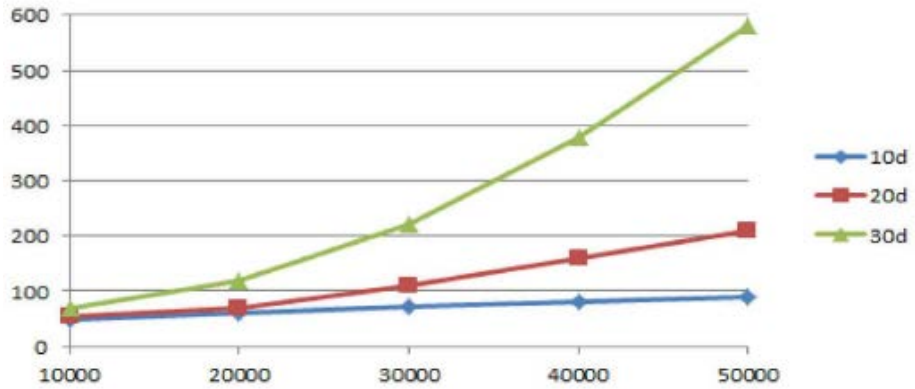

Fig.3. Two clustering algorithms for dimension of scalability test results

We do research in theory and experiment on the consolidated data to verify the validity of the algorithms presented in this paper. This based on, we will this algorithm with in a buy navigation website forum registered user property of mining analysis, we select has four a property, and on data for has and processing, last will paper of algorithm for calculation user in the of away from Group points, last get has optimistic of results; while, to compared, we also will based on once poly class of algorithm used to mining (this method has was proved effective), last of results difference is unlikely to, and paper of algorithm by spent time more short.

This study was not found "large mode", but a "small mode" problem, which is that of outlier data mining. The so-called outlier, refers to the abnormal compared with large amounts of data relatively isolated data mode. In previous research, outlier is often as a byproduct of clustering many clustering algorithms are as a noise or being directly discard outliers, but these data may represent the development trend of the whole or part of the difference in characteristics, is of great value.

References

[1] Shao Fengjing, in chungchong. principles of data mining and algorithms ... China waterpower press, 2003

[2] Wang Yung-ching. principles and methods of artificial intelligence., XI ' an Jiaotong University Press 1999.

[3] Fan Ming, Meng Xiaofeng, translated. data mining concepts and technologies. mechanical industry publishing house, 2001.

[4] Shi Donghui, Cai Qingsheng, Ni Zhiwei. disaggregated data on methods of outliers mining based on rule research. of computer research and development 2007.

[5] Ming Fan, Fan Hongjian. Introduction to data mining. the people's posts and telecommunications Publishing House 2006.

[6] Chen zonghai. process system modeling and simulation. University of science and technology of China Press, 2009.

[7] V. Cherkassky, Y.Q. Ma. Practical Selection of SVM Parameters and Noise Estimation for SVM Regression, Neural Networks, 2004,17(1): 113~126