APPLICATION OF DYNAMIC CLUSTERING ALGORITHM IN MEDICAL SURVEILLANCE

Zhuohui Ren and Cong Wang

Department of Software Engineering, Beijing University of Posts and Telecommunications, BeiJing City, China

ABSTRACT

The traditional medical analysis is based on the static data, the medical data is about to be analysis after the collection of these data sets is completed, but this is far from satisfying the actual demand. Large amounts of medical data are generated in real time, so that real-time analysis can yield more value. This paper introduces the design of the Sentinel which can realize the real-time analysis system based on the clustering algorithm. Sentinel can realize clustering analysis of real-time data based on the clustering algorithm and issue an early alert.

KEYWORDS

Algorithms, Data Mining , Cluster, Data stream, Medical

1. INTRODUCTION

With the arrival of big data era, Medical big data has gradually entered the people's vision, Medical big data refers to all the big data related to medical and life health. According to the source of medical big data can be broadly divided into biological big data, clinical big data and health big data[1].This The potential value of medical data is enormous. For example, public health departments can conduct comprehensive disease surveillance in the monitoring of infectious diseases through a nationwide electronic medical record database, and analysis the characteristics of the spread of illness through data mining.

In the field of health care, most of the data can be seen as streaming data, such as out-patient records, electronic medical records and so on. These data increase by the of time and the numbers of people, It has the characteristic of continuity. Because of its real-time nature, it plays an important role in disease surveillance. For example, mining of outpatient records can dynamically detect diseases that increase in a large amount over a certain period of time, for example, sudden infectious diseases or collective poisoning. Unlike traditional databases that contain static data, data stream are inherently continuous and unconstrained, and there are many problems when working with such data. In addition, the result of data analyse is very unstable and constantly generating new patterns. Static pattern mining techniques proved inefficient when working with data streams. With the deepening of information technology in the medical field, the ability of generating data is rapidly increasing. Mining useful information from these streams has become an inevitable task.

Natarajan Meghanathan et al. (Eds) : NLP, JSE, CST, SIP, ARIA - 2018
pp. 83– 86, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80208

2. RELATED WORK

Sudipto Guha proposed a clustering algorithm based on stream data [2].In his algorithm, the idea of divide and conquer is adopted, the data flow is divided into multiple segments, and the segments is separately clustered to obtain the first cluster center. When the first cluster center reaches a certain number, the second segment of data is introduced to cluster to get the second cluster center. As data continues to flow in, this process will continue. At each time point, the system only needs to maintain m i-th layer center points. This division of mind is very efficient for the analysis of streaming data. Since only a limited number of data needs to be saved at each time point in the system, the storage and memory shortage due to the large incoming stream data are avoided. Because the stream data analysis is a dynamic process, Most of the algorithms are based on the needs of the application to choose the time as a standard, select a period of time to analysis. According to the selected timing range can be divided into snapshot model, landmark model And sliding window model. Landmark model and sliding window model are more used.

As an important algorithm in data mining, the main goal of clustering is to classify the internal relations between data into a large category and distinguish each category as much as possible, which is an extension in taxonomy. According to the different basic principles of clustering can be divided into, division clustering, hierarchical clustering ,density-based clustering, model-based clustering and grid-based clustering [3].

With the extension time or space will produce a wide range of data, and data mining is to extract valuable information from these complex types of data. These complex types of data can be divided into spatial data, timing data, web data, text data [4]. From its process, dynamic data mining can be divided into several stages such as dynamic data collection, data processing, data mining, and mining evaluation [5]. In general, data mining and mining evaluation are closely integrated. Dynamic data mining needs better handling of real-time data and the impact of real-time data on analysis results. The main problems of k-mean algorithm in dealing with dynamic data mining are as follows: Since the initial value of k is fixed means that it can not be changed after it is selected, that makes k-means algorithm unsuitable for mining of dynamic data. Therefore, in the k-means algorithm for dynamic data clustering algorithm, the improvement mainly focuses on the selection and dynamic adjustment of the k value [6], which can be mainly divided into two directions:1, in the process of dynamic data acquisition of data preprocessing, according to the predetermined strategy to adjust the size of k;2 In the data mining process according to the data mining results and predetermined criteria, the data results are dynamically adjusted, and then update k value. The difference between the two methods is that the former is adjusted in the data processing stage and the latter is adjusted in the data mining stage.

The algorithms based on the first idea are: K-means clustering algorithm based on KD tree[6]. The KD tree represents a k-dimensional storage structure that stores data separately at each node in a well-spaced space. Since the initial cluster centers in the k-means algorithm are randomly selected, they can not reflect the true distribution of the data. In order to distribute the actual reaction data as much as possible, it is better to distribute the initial center points more evenly. The basic idea of clustering using KD tree is as follows: Firstly, the advantages of KD tree are used to divide the spatial extent of data set and the data of the corresponding interval is stored. This will effectively improve the effect of the initial center point selection. Using KD tree to divide the space and preprocess the data, we can know the distribution of the data truly. Then according to the partitioned interval, the initial center point is chosen directionally. Finally, the clustering operation is carried out. The algorithm can better find the k-value and the clustering center point, but the computational cost is larger when the clustering operation is re-performed. Compared with the first method of dynamically adjusting the k value, the second method is based on The computation overhead caused by the local dynamic adjustment of the clustering result and the result evaluation index will be significantly reduced. For dynamic data sets, it is obviously

inefficient to re-execute the clustering algorithm on the updated new data set to update the clustering results accordingly, so it is very important to adopt incremental clustering algorithm effectively .

Among them, the algorithm based on the second idea has a two-point k-means clustering algorithm [7]. The main idea is to adjust the clustering result locally instead of the global adjustment according to the threshold in the process of data clustering, which can effectively improve the efficiency of the algorithm and does not affect the final result of the clustering.

3. PROBLEM SETUP

The main problem in the processing of streaming data is that streaming data is a sequence of data sequences that is massively and continuously arriving[8]. When the clustering algorithm is applied to streaming data, it is mainly necessary to consider the real-time performance and the scale unpredictability of the streaming data.
Sentinel's main process is as following steps:

Step 1. Monitoring data cache, if the cache data to meet the conditions to step two.

Step 2. Cached data submitted to the data analysis module, analysis module is used for data analysis, and based on the results, update corresponding parameters.

Step 3. Data early warning module to update the data to determine, greater than the predetermined value issued a alert.

Step 4. Return to step one.

Data caching is mainly for real-time streaming data processing, according to Sudipto Guha' algorithms in the treatment of streaming data, the idea of data segmentation processing, application cache technology can be very good to achieve this. The process of caching is to segment the data base on time line. Data analysis moudle is mainly based on dynamic clustering algorithm. In the data analysis of a block need to use the relevant information in the database, the information is the system needs long-term maintenance. The content of this information includes the number of clusters, the center of each cluster, and the data set that belongs to each cluster. The data processing flow train of thought is as following steps:

Step 1 According to the cluster center stored in the system, the data will be assigned to the corresponding cluster.

Step 2 pairs of clusters of data are calculated and compared with the threshold, according to the comparison result,then adjust of clusters.

Step 3 According to the results of the adjustment, update the relevant records in the database.

In the local adjustment of the cluster, the main reference is the intra-cluster similarity and inter-cluster similarity. The inter-cluster similarity is defined as the mean of the data in the cluster. The similarity between clusters is defined as the distance between the centers of two adjacent clusters. If the similarity between clusters in a cluster is greater than the threshold, the k-means algorithm for k = 2 is performed on the cluster[9]. If the cluster similarity between two clusters is greater than the threshold, the two clusters are merged.

4. CONCLUSION AND FUTURE WORK

There are some places in the system design that deserve further study, mainly for setting the threshold of division and consolidation. The setting of the threshold determines the quality of the splitting and merging[10]. At the same time, the setting of the threshold has a great relationship with the selection of data types. How to find out a suitable model to adapt the model to more types of The data set will be very necessary. Relevant researchers can conduct in-depth research based on different subjects in medical data, and build a better model to make the algorithm better adapt to various data mining.

REFERENCES

[1] Meng Qun, Bi Dan, Zhang Yiming et al .Chinese Journal of Health Information Management, 2016, 13 (6): 547-552..

[2] Sudipto Guha. Asymmetric k-center is log^*n-hard to Approximate[J]. Journal of the Acm, 2013, 52(4):538-551.

[3] Wang Juan.Study on Evolutionary Clustering Algorithm in Dynamic Data Mining [D]. Nanjing University of Aeronautics and Astronautics, 2012.

[4] Zhang Yufeng, Zeng Yitang, Hao Yan.Study on Intelligent Strategy of Logistics Information Based on Dynamic Data Mining [J] .Library Science, 2016 (5): 46-49.

[5] Wang Lunwen, Feng Yanqing, Zhang Ling.A Review of Constructive Learning Methods for Dynamic Data Mining [J] .Microsoft Microcomputer Systems, 2016, 37 (9): 1953-1958.

[6] Tang C, Ling C X, Zhou X, et al. Proceedings of the 4th international conference on Advanced Data Mining and Applications[C]// International Conference on Advanced Data Mining and Applications. Springer-Verlag, 2008:15-15.

[7] Lunwen Wang, Yanqing Feng, Ling Zhang.A Review of Constructive Learning Methods for Dynamic Data Mining [J] .Microsoft Microcomputer Systems, 2016, 37 (9): 1953-1958.

[8] Xiujin Shi, Yanling Hu,et al.Privacy Protection of Dynamic Set-valued Data Publishing Based on Classification Tree [J] .Computer Science, 2017, 44 (5): 120-124.

[9] Guangcong Liu,Tingting Huang,Haiin Chen,et al.An improved dichotomous K-means clustering algorithm [J].Computer Applications and Software, 2015 (2): 261-263.

[10] Zhu Y T, Wang F Z, Shan X H, et al. K-medoids clustering based on MapReduce and optimal search of medoids[C]// International Conference on Computer Science & Education. IEEE, 2014:573-577.

AUTHORS

Ren Zhuohui, Male, 1989, renzh@bupt.edu.cn, Master's degree of Beijing University of Posts and Telecommunications. Main research areas include privacy protection and data mining

Wang Cong, Female, 1958, Professor and doctoral tutor at Beijing University of Posts and Telecommunications. Main research areas include intelligent control and Wisdom-Web information security.