온라인 계층적 군집화 기법을 활용한 양방향 헬스케어 프레임워크

Ibrahim Musa Ishag Musa, 정석호, 신동문, 이경민, 이동규, 손교용, 류근호
충북대학교 데이터베이스연구실
e-mail: { ibrahim, sukhojung, mastershin216, min9709, dglee, gysohn, khryu}@dblab.chungbuk.ac.kr

An Interactive e-HealthCare Framework Utilizing Online Hierarchical Clustering Method

Ibrahim Musa Ishag Musa, Sukho Jung, DongMun Shin, Gyeong Min Yi, Dong Gyu Lee, Gyoyong Sohn, Keun Ho Ryu
Database/Bioinformatics Laboratory, Chungbuk National University

Abstract

As a part of the era of human centric applications people started to care about their well being utilizing any possible mean. This paper proposes a framework for real time on-body sensor health-care system, addresses the current issues in such systems, and utilizes an enhanced online divisive agglomerative clustering algorithm (EODAC); an algorithm that builds a top-down tree-like structure of clusters that evolves with streaming data to rationally cluster on-body sensor data and give accurate diagnoses remotely, guaranteeing high performance, and scalability. Furthermore it does not depend on the number of data points.

1. Introduction

Nowadays healthcare systems are becoming an essential part of hum[1]an's life. In the biomedical literature, many researchers have been done in this area [1], [2] yet to the best of our knowledge. Our proposed framework and the online clustering method are one of the few models for real-time healthcare. The problem of clustering on-body sensor data can be viewed as grouping together those sensors (variables) that produce the same values over time. This way makes doctors interpret peoples activities over time and help them advise and take actions upon their patients.

The most promising work in Healthcare system is Framingham's ten years risk system which evaluates the risk of having Heart disease according to given metrics. But with the invention of the Body Sensor Networks (BSNs) which has been widely deployed for monitoring and measuring diagnosing measurements; a new type of time series data became a challenge in the biomedical field [3]. The main task in e-healthcare is to diagnose patient's case in real time which can save live. Many classification techniques have been applied for diagnosing with the unsupervised classification being the most accurate [4]. Clustering sensor data needs to be done in different ways. One of which is to consider a variable clustering rather than point clusters, to organize sensor values in a medically understandable manner.

Hierarchical clustering has three advantages over other clustering methods even in stationary data. Those advantages are firstly it does not require user involvement in specifying the number of clusters, as it happens in other clustering methodologies for example in partitional clustering. Secondly, most of hierarchical methods do not require whole data to be available at once as BIRCH [7]. Thirdly it has a linear time complexity with respect to the size of input [8]. Variable clustering is useful for many applications especially for applications where the data sets are of high dimensionality, like power distribution records, monitoring real world phenomenon, and on body sensor data that produces a massive streams of times series data resembling various characteristics of human's body that help doctors trace their patients remotely and advice them to change their bad habits or even send them medical prescriptions.

2. Proposed Framework

In this paper we propose a framework for online mining streaming time series data, and the framework consists of three parts. Firstly the continuously coming data from on-body sensors are preprocessed missing values and transform categorical data into numeric to allow for applying the similarity measure, secondly our algorithm is applied for clustering the data in real time and the cluster structure is stored in a database to allow for incoming user queries from Medical centers and user's own mobile phones. Those kinds of query allow for asking about the shape of clusters in a specific point of time. Since the architecture is employing a

이 논문은 2009 년 교육과학기술부로부터 지원을 받아 수행된 연구(지역거점연구단육성사업/충북 BIT 연구중심대학육성사업단)와 2009 년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. R11-2008-014-02002-0)

geometric time frame, queries are fine grained if required for a recent time and tolerated approximations for the past. Lastly the final structure is visualized to the end user. In addition, patient's locations have to be detected for better healthcare. Thus another issue arises which is how to preserve patients privacy in such frame work.

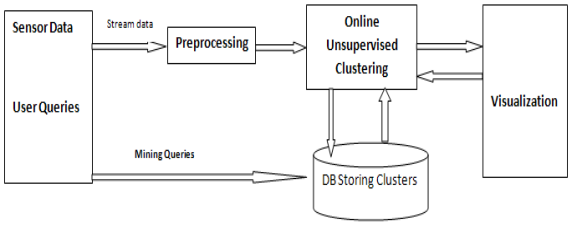

Figure 1 Interactive e-Healthcare framework.

The main part is the unsupervised clustering method which is an enhanced version of the Online Divisive Agglomerative Clustering (ODAC) proposed by [5], our enhancement is in how to split a cluster, because ODAC's problem was that it sometimes ends up giving an inaccurate clusters because of strict splits, and even in The Semi-Fuzzy version of ODAC [6] there is still one problem which is the duplicate clusters that results from assigning some variables to tow clusters.

The algorithm builds a tree-like top down variable hierarchy of clusters by measuring the similarity between two time series using Pearson's Correlation coefficient between time series, given by following formula.

$$corr\ (a,b) = \frac{P - \frac{AB}{n}}{\sqrt{A_2 - \frac{A^2}{n}}\sqrt{B_2 - \frac{B^2}{n}}} \qquad (1)$$

$$A = \Sigma a_i, B = \Sigma b_i, A_2 = \Sigma a_i^2 B_2 = \Sigma b_i^2, P = \Sigma a_i b_i$$

And take splitting decisions based on rooted normalized one minus correlation supported by statistical Hoefding bound.

$$rnomc = \sqrt{\frac{1 - corr\ (a,b)}{2}} \qquad (2)$$

And it is not just split clusters but also sometimes detect a convergence and decides to reaggregate towards parents by monitoring cluster diameter, a second and third maximum dissimilarity. The maximum number of clusters obtained in one split is four. Figure 2 explains the idea.

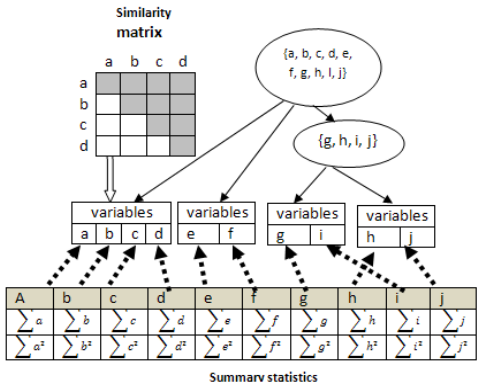

Figure 2. E-ODAC and our proposed splitting enhancement

As shown in this figure, cluster split and reaggregate based on summary statistics calculated at leaf nodes and the new coming streams are fed to the equivalent clusters at leaf node, thus reducing computation time and memory space used.

3. Conclusion

In this paper we proposed a novel framework for e-healthcare system utilizing on-body sensor data that consists of three parts; preprocessing to remove outliers and resolve missing data, the Extended Online Divisive Agglomerative Clustering method, and a database back end for storing the resulting cluster structure, and finally data visualization. The most important part is the online multi-divisive approach that uses our proposed E-ODAC to perform the tasks. Future work is to add moving object algorithm to detect patient's location and help save life.

References

[1] Bernardo G. et al. "ECG Data Provisioning for Telehomecare Monitoring" ACM 2008.

[2] Georgios G et al. "ECG signal Recording Processing and Transmission Using a Mobile Phone" PETRA 2008.

[3] Gaurav N. Pradhan, and Balakrishnan Prabhakaran "Storage, Retrieval, and Communication of Body Sensor Network Data", *TUTORIAL SESSION, ACM*, Vancouver, British Columbia, Canada, 2008, pp. 1161-1162.

[4] Nishizwa, H., et al., "Hierarchical Clustering Method for Extraction of knowledge from large amount of data " Optical Review 1999.

[5] Pedro Pereira Rodrigues, Jo˜ao Gama and Joao Pedroso."Hierarchical Clustering of Time Series data streams" IEEE Transactions on Knowledge and Data Engineering, Piscataway, NJ, USA, 2008, pp. 615-627.

[6] Pedro Pereira Rodrigues, and Jo˜ao Gama. "Semi-fuzzy Splitting in Online Divisive-Agglomerative Clustering", Lecture Notes in Computer Science (LNCS), 2007, pp 133-144.

[7] Zhang,T R. Ramakrishnan, and Livny, M., "An Efficient Data Clustering Method for very Large Databases", processing of ACM International Conference on Management of Data., Birch, 1996. pp. 103-114.

[8] Mohamed, A. N., Leckie, C., and Udaya, P. "An Efficient Clustering Scheme to Exploit Hierarchical Data in Network Traffic Analysis". IEEE Transactions on Knowledge and Data Engineering, 20(6). 2008, pp.752-767.