Resource Constrained Data Stream Clustering with Concept Drifting for Processing Sensor Data

Gansen Zhao, School of Computer Science, South China Normal University, Guangzhou, China

Zhongjie Ba, School of Software, Sun Yat-sen University, Guangzhou, China

Jiahua Du, School of Computer Science, South China Normal University, Guangzhou, China

Xinming Wang, School of Computer Science, South China Normal University, Guangzhou, China

Ziliu Li, Microsoft Search Technology Center Asia, Beijing, China

Chunming Rong, Centre of Innovation Technology, University of Stavanger, Stavanger, Norway

Changqin Huang, School of Information Technology in Education, South China Normal University, Guangzhou, China

ABSTRACT

Wireless sensors and mobile devices have been widely deployed as data collecting devices for monitoring real world systems. A large amount of stream data is generated in real-time, which has to be processed in real-time as well. One of the common processing operations is clustering that automatically groups the elements of a stream into a number of clusters in general. Elements of the same cluster have maximum similarity and elements of different clusters have minimum similarity. This paper proposes an on-demand framework (SRAStream) based on the concept drifting detection mechanism. The concept drifting detection algorithm is used to measure the distance of the new clusters for the current data and that of the existing clusters. Only when a concept drifting occurs will the re-clustering be performed to identify new clusters. SRAStream thus avoids the unnecessary computation intensive re-clustering calculation. Experiments suggest that the proposed framework does work well and improve the processing speed greatly in data streams clustering.

Keywords: Concept Drifting, Data Mining, Stream Data

Figure 1. Clustering real-time data generated by mobile devices
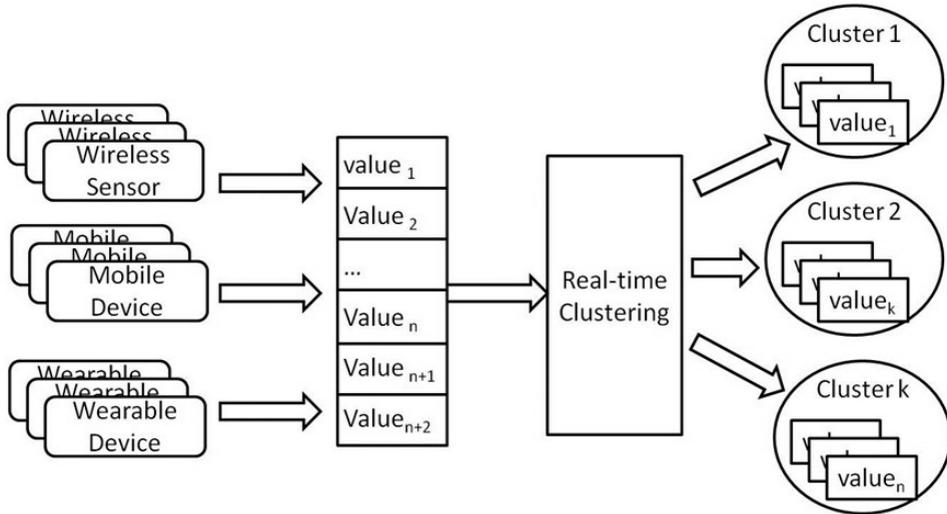
1. INTRODUCTION

Cyber physical systems deploy a great number of sensors and various mobile devices to monitor real world systems. Each sensor and device senses and measures objects in a regular interval, generating new values from time to time, hence resulting in a continuous data stream(shown in Figure 1). In such a scenario data streams need to be processed to identify events and extract useful information. For example, in the context of smart cities, mining user daily behavior data stream is able to identify the daily commuting patterns of users.

Clustering is a process of grouping a set of objects. In general, clustering puts similar objects in the same group and less similar objects in different groups, where objects of the same group have the most similarity and objects of different groups have the least similarity. Clustering stream data is one of the most common data processing operations for mining real time data in many scenarios.

Data streams are different from static data sets in several perspectives (Aggarwal & Yu, 2008).

- Data arrives as one or more continuous and unbounded data streams, requiring a great amount of resource to process while most processing systems have limited amount of resource for use.
- Data streams usually are not available for random access and most items in a data stream may be processed only once during the entire computation. This is due to the great volume and some data may be abandoned after being processed.
- The data arrival rate of a data stream varies from time to time and it could be difficult to predict.

There are a few major challenges of data streams clustering, including but not limited to the follows:

- Data streams are dynamic with new data comes continuously. It is not possible for performing the clustering operation after receiving all data. Hence it is necessary to be able to cluster without having the whole data and update the cluster result continuously as new data arrive.

Related Content

Spatial Clustering in SOLAP Systems to Enhance Map Visualization
Ricardo Silva, João Moura-Pires and Maribel Yasmina Santos (2012). *International Journal of Data Warehousing and Mining (pp. 23-43).*
www.igi-global.com/article/spatial-clustering-solap-systems-enhance/65572?camid=4v1a

Mining Association Rules in Data Warehouses
Haorianto Cokrowijoyo Tjioe and David Taniar (2005). *International Journal of Data Warehousing and Mining (pp. 28-62).*
www.igi-global.com/article/mining-association-rules-data-warehouses/1755?camid=4v1a

A New Approach for Supervised Dimensionality Reduction

Yinglei Song, Yongzhong Li and Junfeng Qu (2018). *International Journal of Data Warehousing and Mining (pp. 20-37).*

www.igi-global.com/article/a-new-approach-for-supervised-dimensionality-reduction/215004?camid=4v1a

Mining Spatio-Temporal Trees

Wynne Hsu, Mong Li Lee and Junmei Wang (2008). *Temporal and Spatio-Temporal Data Mining (pp. 209-226).*

www.igi-global.com/chapter/mining-spatio-temporal-trees/30268?camid=4v1a