Clustering for monitoring distributed data streams

Barouti, Maria . University of Maryland, Baltimore County, ProQuest Dissertations Publishing, 2016. 10159957.

ABSTRACT

Data mining is a challenging research area of computer science with profound applications in database industries and resulting market needs. Data mining is the computational process of discovering patterns in big data sets. This process enables us to extract valuable information from large data by involving methods at the intersection of different topics such as machine learning, statistics, and artificial intelligence. Over the last years there has been a growing interest in data analysis research by monitoring data streams in a distributed system. In this study we propose to monitor arbitrary threshold functions over distributed data streams while minimizing communication overhead. To illustrate this further, assume that we have a number of sensors that are spread in the space and we would like to monitor the average of their measurements while minimizing communication between the sensors. Each sensor represents a node that produces time varying vectors derived from the stream of measurements. Thus we are interested to check if a function evaluated at the vectors' average at each time is greater than zero while communication between the nodes is minimized.

Motivated by recent contributions based on geometric ideas, and after reviewing some well known clustering algorithms we present an alternative approach that combines system theory techniques, clustering and statistical approaches. Our approach enables monitoring values of an arbitrary threshold function over distributed data streams through a set of constraints applied independently on each stream and/or clusters of streams. The clusters are designed to evolve in time and to adapt themselves to the data stream. A correct choice of clusters yields a reduction in communication load. Unlike many clustering algorithms that attempt to collect together similar data items, monitoring requires clusters with dissimilar vectors canceling each other as much as possible. In particular, subclusters of a good cluster do not have to be good. This novel type of clustering dictated by the problem at hand requires development of new algorithms and/or modification of the existing ones, and this thesis is a step in this direction.

We report experiments on real-world data with a newly devised clustering algorithm. The experiments detect instances where communication between nodes is required, and show that the clustering approach reduces communication load. Last but not least, we indicate new future directions and discuss possible methodologies that can be involved into my future research agenda.

DETAILS

Subject:	Mathematics; Computer science
Classification:	0405: Mathematics; 0984: Computer science
Identifier / keyword:	Pure sciences Applied sciences Clustering Data mining Distributed data streams Feature selection Monitoring Optimization
Number of pages:	66

Publication year:	2016
Degree date:	2016
School code:	0434
Source:	DAI-B 78/02(E), Dissertation Abstracts International
Place of publication:	Ann Arbor
Country of publication:	United States
ISBN:	9781369148848
Advisor:	Kogan, Jacob Malinovsky, Yaakov
Committee member:	Kogan, Jacob; Lo, James; Malinovsky, Yaakov; Nicholas, Charles; Park, Junyong
University/institution:	University of Maryland, Baltimore County
Department:	Mathematics, Applied
University location:	United States -- Maryland
Degree:	Ph.D.
Source type:	Dissertations &Theses
Language:	English
Document type:	Dissertation/Thesis
Dissertation/thesis number:	10159957
ProQuest document ID:	1826021656
Document URL:	http://proxy-bc.researchport.umd.edu/login?url=https://search.proquest.com/docview/1826021656?accountid=14577
Copyright:	Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works.
Database:	ProQuest Dissertations &Theses Global,Dissertations &Theses @ UMBC

LINKS

Linking Service