Chapter 3
A Parameterized Framework for Clustering Streams

Vasudha Bhatnagar
University of Delhi, India

Sharanjit Kaur
University of Delhi, India

Laurent Mignet
IBM, Indian Research Lab, India

ABSTRACT

Clustering of data streams finds important applications in tracking evolution of various phenomena in medical, meteorological, astrophysical, seismic studies. Algorithms designed for this purpose are capable of adapting the discovered clustering model to the changes in data characteristics but are not capable of adapting to the user's requirements themselves. Based on the previous observation, we perform a comparative study of different approaches for existing stream clustering algorithms and present a parameterized architectural framework that exploits nuances of the algorithms. This framework permits the end user to tailor a method to suit his specific application needs. We give a parameterized framework that empowers the end-users of KDD technology to build a clustering model. The framework delivers results as per the user's application requirements. We also present two assembled algorithms G-kMeans and G-dbscan to instantiate the proposed framework and compare the performance with the existing stream clustering algorithms.

INTRODUCTION

Data streams pose special challenges to mining algorithms, not only because of the huge volume of on-line data streams and its computation (Henzinger, Raghavan & Rajagopalan, 1998; Babcock, Babu, Datar, Motwani & Widom, 2002; Carney, Cetintemel, Cherniack, Convey, Lee, Seidman et al., 2002; Domingos and Hulten, 2000), but also because of the fact that data in streams may show temporal correlations. Such temporal correlations help in disclosing important data trends in XML document clustering (Rusu, Rahayu & Taniar, 2008), multimedia communication and programming support for ubiquitous distributed computing environment (Aggarwal, 2007).

Clustering is considered as one of the most popular and effective techniques for discovering similarity trends in data streams. Compactness of representation, fast incremental processing of new data points, insensitivity to order of input records have been identified as basic requirements in stream clustering algorithms (Henzinger, Raghavan & Rajagopalan, 1998; Barb´ara, 2002; Orlowska, Sun & Li, 2006).

The problem of incremental clustering is addressed in Zhang, Ramakrishnan & Livny (1996) and inspired clustering of data streams. The importance of the problem is evident from the large body of work (Aggarwal, Han, Wang & Yu, 2003; Motoyoshi, Miura & Shioya, 2004; Park & Lee, 2004) that has evolved over a relatively short period of time since the earliest attempt to address the problem of stream clustering (Guha, Mishra, Motwani & O'Callaghan, 2000).

The algorithms that have been developed for stream clustering have either an on-line or a batch component for processing incoming data, to maintain synopsis. A mechanism is used to highlight the evolving nature of data in stream. Clustering is done using varied approaches based on distance (*k-means* or *k-median*), density estimation, statistical methods (e.g. co-variance, skewness etc.) and connected component analysis.

Motivation

One of the reasons for the fallen-short-of-anticipated growth curve of KDD technology is that the end-user is forced to use the mining algorithms provided by the data mining packages and has no say in designing the algorithm. The current KDD technology is limited by the adhoc approach for solving individual problems (Yang & Wu, 2006). The need for a unified framework for integrating different data mining tasks has been recognized recently (Yang & Wu, 2006).

Motivated by the above observation, we propose a parameterized framework for stream clustering. The framework empowers the end-user to choose the features of the algorithm to suit their business requirements in terms of nature of inputs, outputs, availability of resources etc.. The proposed component-based architecture of stream clustering algorithms advocates development of a data-mining environment where the user can match the application needs with the features of the components and assemble the algorithm. The approach overcomes the rigidity prevalent in the use of data mining environments, where the match between the available algorithmic features and desired functionality is sometime less than satisfactory. This work lays the theoretical foundation for the unified framework by parameterizing an algorithm based on application requirements.

Outline of the Paper

The paper is divided into five sections. Section "Comparison of Stream Clustering Algorithms" studies different approaches used in stream clustering algorithms, and a systematic comparison vis-à-vis the nature of input, output, processing and functionality is presented. The study leads to a component based architectural framework underlying all stream clustering algorithms, which is discussed in Section "Generic Architecture for Stream Clustering Algorithms". Based on this framework, subsection "Architectural Framework" proposes a scheme to assemble designer algorithms by selecting appropriate components to suit the user's specific needs. Section "Realization of the Framework" instantiates the proposed framework by laying down hypothetical user requirements and assembling two algorithms *G-kMeans* and *G-dbscan.* Experimental evaluation of the two algorithms is also presented in the same section.

Related Content

Devising Parametric User Models for Processing and Analysing Social Media Data to Influence User Behaviour: Using Quantitative and Qualitative Analysis of Social Media Data

Jonathan Bishop (2017). *Social Media Data Extraction and Content Analysis (pp. 1-41).*

www.irma-international.org/chapter/devising-parametric-user-models-for-processing-and-analysing-social-media-data-to-influence-user-behaviour/161957

Statistical Entropy Measures in C4.5 Trees

Aldo Ramirez Arellano, Juan Bory-Reyesand Luis Manuel Hernandez-Simon (2018). *International Journal of Data Warehousing and Mining (pp. 1-14).*

www.irma-international.org/article/statistical-entropy-measures-in-c45-trees/198971

Preparing for New Competition in the Retail Industry

Goran Klepac (2010). *Data Mining in Public and Private Sectors: Organizational and Government Applications (pp. 245-266).*

www.irma-international.org/chapter/preparing-new-competition-retail-industry/44292

Introduction to Data Mining Techniques via Multiple Criteria Optimization Approaches and Applications

Yong Shi, Yi Peng, Gang Kouand Zhengxin Chen (2007). *Research and Trends in Data Mining Technologies and Applications (pp. 242-275).*

www.irma-international.org/chapter/introduction-data-mining-techniques-via/28427

An Envisioned Approach for Modeling and Supporting User-Centric Query Activities on Data Warehouses

Marie-Aude Aufaure, Alfredo Cuzzocrea, Cécile Favre, Patrick Marceland Rokia Missaoui (2013). *International Journal of Data Warehousing and Mining (pp. 89-109).*

www.irma-international.org/article/envisioned-approach-modeling-supporting-user/78288