FINGERPRINT: Summarizing Cluster Evolution in Dynamic Environments

Eirini Ntoutsi, Institute for Informatics, Ludwig-Maximilians University of Munich, Germany
Myra Spiliopoulou, University of Magdeburg, Germany
Yannis Theodoridis, University of Piraeus, Greece

ABSTRACT

Monitoring and interpretation of changing patterns is a task of paramount importance for data mining applications in dynamic environments. While there is much research in adapting patterns in the presence of drift or shift, there is less research on how to maintain an overview of pattern changes over time. A major challenge is summarizing changes in an effective way, so that the nature of change can be understood by the user, while the demand on resources remains low. To this end, the authors propose FINGERPRINT, an environment for the summarization of cluster evolution. Cluster changes are captured into an “evolution graph,” which is then summarized based on cluster similarity into a fingerprint of evolution by merging similar clusters. The authors propose a batch summarization method that traverses and summarizes the Evolution Graph as a whole and an incremental method that is applied during the process of cluster transition discovery. They present experiments on different data streams and discuss the space reduction and information preservation achieved by the two methods.

Keywords: Change Detection, Change Monitoring, Change Summarization, Cluster Evolution, Cluster Summarization, Data Streams, Dynamic Environments

INTRODUCTION

Data streams are used in many modern applications and impose new challenges for the data management systems because of their size and high degree of variability. One of the challenges is the efficient detection and monitoring of changes in the underlying population. For example, changes in the patterns known to a network intrusion detection system may indicate that intruders test new attacks and abandon old, already known (and blocked) intrusion patterns.

In general, monitoring of change is essential for applications demanding long-term prediction and pro-action.

Cluster models are commonly used as a tool for studying the dynamics of a population. In recent years actually, due to the dynamic nature of data, it has been recognized that clusters upon the data of many real applications are affected by changes in the underlying population of customer transactions, user activities, network accesses or documents. A lot of research has been devoted in adapting the clusters to the changed population. Recently, research has expanded to encompass tracing and understanding of
the changes themselves, as means of gaining insights on the population; see for example the survey of Spiliopoulou (2011) on the evolution of social networks. Understanding change is also important when taking strategic decisions: Consider, for example, a business analyst who studies customer profiles; understanding how such profiles change over time would allow for a long-term proactive portfolio design instead of reactive portfolio adaptation. While much research has been recently devoted to pattern change detection, little work has been done on the efficient maintenance of the pattern changes.

The maintenance and summarization of pattern changes upon a stream is a new problem. Summarization of data (rather than patterns); however, has been studied extensively: Popular summarization methods include histograms and wavelets, and there is much work on the efficient maintenance of these structures and on the adaptation of their contents when data change; however, these methods do not show how the data change nor do they maintain the changes themselves. There is also research on storing, modifying and querying patterns in inductive or conventional databases (e.g., Bartolini, Ciaccia, Ntoutsi, Patella, & Theodoridis, 2004); however, those approaches have not been designed for patterns over streams and, although there is provision for modifying patterns when new data arrive, there are no solutions on the efficient maintenance of changes over time. Finally, there are methods for pattern change detection (e.g., Aggarwal, 2005; Mei & Zhai, 2005; Spiliopoulou, Ntoutsi, Theodoridis, & Schult, 2006), in which different types of change can be identified and highlighted; however, the efficient long-term maintenance of the changes over an “infinite” stream is not considered.

Evolution is a permanent characteristic of streamed data, thus long-term perusal requires a space-efficient accommodation of the evolving patterns and a representation that highlights remarkable changes while suppressing trivial pattern perturbations. In this study, we propose a graph representation of pattern changes/ transitions and two algorithms that condense this graph into a “fingerprint” - a structure in which similar patterns are efficiently summarized, subject to an information loss function.

The rest of the paper is organized as follows: Related work is discussed in the upcoming section. We then present our graph model for the representation of cluster transitions. The criteria for the summarization of cluster changes and the actual summarization methods are presented afterwards. Experiments are presented in the next section. Finally, the last section concludes our work.

RELATED WORK

Relevant to our work is the work on data summarization, stream clustering and change detection. We review these areas hereafter and point out how we differentiate.

Summarization Methods

Summarization for a set of transactions with categorical attributes is studied by Chandola and Kumar (2005). In one of their methods, they derive summaries by clustering the transactions, extracting the feature/value pairs supported by all transactions in a cluster and using them as cluster summaries. They do not address the issue of cluster change upon a stream, but propose two metrics that characterize the output of the summarization algorithm, “compaction gain” and “information loss”. Quite naturally, our metrics are similarly motivated and have similar names. However, they summarize static data using clusters, while we summarize evolving clusters upon an “infinite” data stream.

Summarization and change are considered by Ipeirotis, Ntoulas, and Gravano (2005), who study changes of database content summaries. They define as “content summary” for a database a set of keywords, weighted on their importance within the database. Meta-search services use such summaries to select appropriate databases, towards which they issue keyword-based queries. The reliability of such a summary deteriorates as the contents of the database change over time. So, the authors propose methods to quantify and detect summary changes. This
A Novel Hybrid Algorithm Based on K-Means and Evolutionary Computations for Real Time Clustering
Taha Mansouri, Ahad Zare Ravasan and Mohammad Reza Gholamian (2014). *International Journal of Data Warehousing and Mining* (pp. 1-14).
[www.igi-global.com/article/a-novel-hybrid-algorithm-based-on-k-means-and-evolutionary-computations-for-real-time-clustering/116890?camid=4v1a](www.igi-global.com/article/a-novel-hybrid-algorithm-based-on-k-means-and-evolutionary-computations-for-real-time-clustering/116890?camid=4v1a)

SQL-Based Fuzzy Query Mechanism Over Encrypted Database
[www.igi-global.com/article/sql-based-fuzzy-query-mechanism-over-encrypted-database/117159?camid=4v1a](www.igi-global.com/article/sql-based-fuzzy-query-mechanism-over-encrypted-database/117159?camid=4v1a)