An Intuitive Framework for Understanding Changes in Evolving Data Streams

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
charu@watson.ibm.com

Abstract

Many organizations today store large streams of transactional data in real time. This data can often show important changes in trends over time. In many commercial applications, it may be valuable to provide the user with an understanding of the nature of changes occuring over time in the data stream. In this poster, we discuss the process of analysis of the significant changes and trends in data streams in a way which is understandable, intuitive and friendly to a user.

1 Introduction, Motivation and Overview

Many databases today are created by continuous activity over long periods of time, and are therefore databases which grow continuously over time. Such dynamically updated databases are referred to as data streams. The volume of the transactions in a typicial data stream such as a telecommunication network may be as high as a few million per day. This data may often exhibit changes over time; a process which is referred to as data evolution. By understanding the bature of such changes a user may be able to convert in into valuable decisions for many commercial applications. Therefore, it is useful to develop tools and techniques which would provide a visual and diagnostic overview of the key characteristics in the data which have changed over time in a fast and user friendly way.

The large number of applications which generate teams of data has lead to a recent interest in this area of research. An important area of interest is that of the incremental maintenance of data mining models over evolving data. In particular, the problem of change analysis has attracted considerable attention because of its implications on the ability of maintain effective data mining models. A recent paper [1] discusses ways of change quantification in terms of the different models and algorithms. The focus and motivation of this paper is quite different from and orthogonal to the work in [1]; the latter is focussed on the effects of data change on data mining models and algorithms, whereas this poster is focussed on the problem of measuring and understanding *data* change directly rather than measuring the effects on data mining *models*.

In many cases, the process of measurement of summary changes can be of great use in a number of real applications. For example, a supermarket application may require us to find the important trends in the volume of transactional activity. For an end user, it may be valuable to be able to track such changes in real time. This can often be a difficult task, since the volume of information in a data stream can be considerable.

The basic approach is to determine the rate of change of data densities at each spatial location by using a computation by a single scan over the records. To this effect, we use differential kernel density estimation methods over different windows in time. In addition, the resulting computations can be converted into different kinds of visual profiles by plotting the density distributions of these computations. This rate of change can be aggregated over the entire data space in order to provide a measurement of the level of change over the entire data set. In many cases, this gives the user a physical understanding of the level of changes in the entire data set. By using windows of different sizes, it is possible to measure both short-term and long-term trends in the data.

We show how to convert these computations into graphical format. This graphical format can be useful to an analyst who wishes to get a broad understanding of the amount and level of changes in the different regions. In some cases, the nature of the change in a given region can be classified which provides an even better understanding of the change.

References

[1] V. Ganti, J. Gehrke, R. Ramakrishnan, W.-Y. Loh. A Framework for Measuring Differences in Data Characteristics. *ACM PODS Conference*, 1999.