Binary data flow visualization on factorial axes

Alfonso Iodice D'Enza[1], Francesco Palumbo[2]

[1] Dipartimento di Matematica e Statistica
Universitá di Napoli Federico II
Complesso Universitario di Monte S. Angelo, via Cinthia
I-80126, Napoli, Italy
iodicede@unina.it
[2] Dipartimento di Istituzioni Economiche e Finanziarie
Universitá di Macerata
via Crescimbeni, 20
I-62100, Macerata, Italy
palumbo@unimc.it

Abstract. *Data streams* are one of the most relevant new data sources, they refer to flows of data that come at a very high rate. Let us consider a stock-exchange market, where n different stocks with p considered attributes (e.g. price, quantity, seller/buyer id, ...) are negotiated all day long. The distinguishing feature in data streams analysis is that the focus is on *transient relations*. The present paper proposes a visualization tool exploiting Multidimensional Data Analisis (MDA) techniques to represent the evolving association structures among attributes over different time-frames. The general aim is to detect the stability of the deviation from indipendence in the occurrence of an observed set of attributes stored as binary stream.

1 Introduction

In recent years, enhancements in monitoring activities and collecting data determined the need for a different approach in knowledge extraction: new data are produced at a faster rate than the capability of analyzing them.
Information mining through traditional data mining systems becomes often inadequate. *Data streams* are one of the most relevant new data sources, they refer to flows of data that come at a very high rate. Let us consider a stock-exchange market, where n different stocks are negotiated: at each interval time-unit (seconds or minutes) a $n \times p$ array is added to the database, with p indicating the number of considered attributes (e.g. price, quantity, seller/buyer id, ...). These features make data streams leading to data structure unusual in the data analysis and statistical data mining framework. New and more appropriated techniques should be taken into account to usefully extract knowledge without storing the data for a long term [12].

The most relevant changing feature in data mining systems dealing with high-speed data streams is the necessity of analyzing data in a single pass: iterative procedures will lead to unfeasible solutions. Further ideal features of a data stream mining system are described in the proposal by [4].

Data streams mining, in a wide sense, can be considered an evolution of data mining. The development of data mining techniques has been feeded in the last decade by statisticians and computer scientists. Similarly, the data stream analysis should be enhanced with the contribution of researchers from different areas.

Data stream analysis techniques can be roughly divided into: *i*) data-based, indicating techniques aiming to summarize or reduce the amount of streams to be analyzed; *ii*) task-based techniques, facing the crucial problem of adapting existing algorithms to the new computational costs;