Chapter 37
Integration of Data Mining and Statistical Methods for Constructing and Exploring Data Cubes

Muhammad Usman
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Pakistan

ABSTRACT

In high dimensional environments, the sheer size and volume of data poses a number of challenges in order to generate meaningful and informative data cubes. Data cube construction and exploration is a manual process in which analysts are required to visually explore the complex cube structure in order to find interesting information. Data cube construction and exploration has been dealt separately in the literature and in the past there has been very limited amount of work done which would guide the data warehouse designers and analysts to automatically construct and intelligently explore the data cubes. In the recent years, the combined use of data mining techniques and statistical methods has shown promising results in discovering knowledge from large and complex datasets. In this chapter, we propose a methodology that utilizes hierarchical clustering along with Principal Component Analysis (PCA) to generate informative data cubes at different levels of data abstraction. Moreover, automatically ranked cube navigational paths are provided by our proposed methods to enhance knowledge discovery from large data cubes. The methodology has been validated using real world dataset taken from UCI machine learning repository and the results show that the proposed approach assists in cube design and intelligent exploration of interesting cube regions.

INTRODUCTION

The extensive use of computers and Information technology has made data collection a routine task in a variety of fields, continuously increasing data repositories can contribute significantly towards future decision making only if appropriate knowledge discovery mechanisms are applied on large datasets. Data

DOI: 10.4018/978-1-5225-5191-1.ch037

Mining (DM) and Data warehousing (DW) are the two main constituents of the knowledge discovery process. DM permits targeted mining of large datasets in order to discover hidden trends, patterns and rules while DW provides for the interactive exploration and multi-dimensional analysis of summarized data. The two strands of research share a common set of objectives such as information/knowledge extraction from large datasets, support for decision making and use of background knowledge for additional information extraction, both DM and DW have progressed rapidly in their independent ways.

Yet, there is significant potential value in integrating these disciplines. However, little research has been carried out in the integration of the two disciplines (Usman and Pears 2010). Undeniably, it is a challenging task as the techniques employed in each of the disciplines are quite different from each other. This highlights a crucial problem in integrating these incompatible techniques in a seamless manner to extract valuable knowledge from the rapidly growing data repositories. According to (Han and Kamber 2006) when a DM system works in an environment that requires it to communicate with other information system components such as DW, possible integration schemes include no coupling, loose coupling, semi-tight coupling and tight-coupling. Tight coupling means that the data mining system is smoothly (seamlessly) integrated into data warehouse system. Such smooth integration is highly desirable because it facilitates efficient implementation of data mining algorithms, high system performance and an integrated information processing environment (Han and Kamber 2006). However, implementation of such an integrated system is nontrivial and extensive research is required in this area.

Recently, there has been an increase in adapting DM and advanced statistical techniques such as Cluster Analysis and Principal Component Analysis (PCA) for knowledge discovery (Messaoud, Boussaid et al. 2004). With similar objectives, various methods and techniques, researchers have been attracted towards DW design and multi-dimensional modeling, with increasing attention being paid to the integration of mining techniques with data warehousing for data-driven knowledge discovery. However, little progress has been made so far to integrate the outcomes of data mining techniques with data warehouse to perform analytical operations. The reason for this difficulty is that data mining results need to be modelled in the form of a multidimensional schema to support interactive queries for the exploration of data.

Multidimensional modelling is a complex task, which requires domain knowledge, solid warehouse modelling expertise and deep understanding of data structures and their attributes. Certainly, in real world scenarios, data warehouse designers possess modelling expertise but lack the domain knowledge and detailed understanding of semantic relationships among the data attributes, this lack of knowledge is the prime reason for a poor warehouse design that in turn has a dramatic effect on knowledge discovery and decision making process. Additionally, multi-dimensional modelling techniques require multiple manual actions to discover measures and relevant dimensions from the dataset, such manual discovery actions become a bottleneck in the knowledge discovery process even if the human data warehouse designer tries to resolve these problems, the designer can still end up with generating an incorrect design if he/she doesn't understand the underlying relationships among the data items. In data warehouse, the choice of the attributes that are to be considered as measures and dimensions heavily influences the data warehouse effectiveness.

Data mining techniques such as clustering and pattern visualization can assist the human data warehouse designer in understanding and visualizing complex data structures. Clustering is one of the most widely researched areas in the DM discipline; data miners have traditionally used clustering as a method of segmenting data in order to recognize different groups inherent in large collections of data. However, a difficult barrier in the efficient clustering of data is the presence of mixed numeric and nominal variables which are present in real-world data sets. Numerous algorithms and techniques have been proposed in the