Knowledge-Based Systems 40 (2013) 36–49

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A data mining approach to knowledge discovery from multidimensional cube structures

Muhammad Usman *, Russel Pears, A.C.M. Fong

Auckland University of Technology, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 19 July 2012
Received in revised form 4 November 2012
Accepted 23 November 2012
Available online 10 December 2012

Keywords:
Data cubes
OLAP analysis
Data mining
Ranked paths
Principal Component Analysis
Multiple Correspondence Analysis

ABSTRACT

In this research we present a novel methodology for the discovery of cubes of interest in large multi-dimensional datasets. Unlike previous research in this area, our approach does not rely on the availability of specialized domain knowledge and instead makes use of robust methods of data reduction such as Principal Component Analysis and Multiple Correspondence Analysis to identify a small subset of numeric and nominal variables that are responsible for capturing the greatest degree of variation in the data and are thus used in generating cubes of interest. Hierarchical clustering was integrated with the use of data reduction in order to gain insights into the dynamics of relationships between variables of interests at different levels of data abstraction. The two case studies that were conducted on two real word datasets revealed that the methodology was able to capture regions of interest that were significant from both the application and statistical perspectives.

1. Introduction

Knowledge discovery aims to extract valid, novel, potentially useful, and ultimately understandable patterns from data [7]. Recently, the integrated use of data mining and Online Analytical Processing (OLAP) has received considerable attention from researchers and practitioners alike, as they are key tools used in knowledge discovery from large data cubes [8,10,21,23,24,34,35,37]. A variety of integrated approaches have been proposed in the literature to mine large data cubes for discovering knowledge. However, a number of issues remain unresolved in that previous work [26,25,16,22], especially on the intelligent data analysis front.

Firstly, the prior work assumed that data analysts could identify a set of candidate data cubes for exploratory analysis based on domain knowledge. Unfortunately, situations exist where such assumptions are not valid. These include high dimensional datasets where it may be very difficult or even impossible to predetermine which dimensions and which cubes are the most informative. In such environments it would be highly desirable to automate the process of finding the dimensions and cubes that hold the most interesting and informative content.

Secondly, reliance on domain knowledge tends to constrain the knowledge discovered to only encapsulate known knowledge, thus excluding the discovery of unexpected but nonetheless interesting knowledge [14]. Another related issue is that it restricts the application of these methodologies to only those domains where such domain knowledge is available. However, a knowledge discovery system should be able to work in ill-defined domains [20] and other domains where no background knowledge is available [36].

This motivated us to formulate a generic methodology for data cube identification and knowledge discovery that is applicable across any given application domain, including those environments where limited domain knowledge exists. High dimensional and high volume datasets present significant challenges to domain experts in terms of identifying data cubes of interest. The presence of mixed data in the form of nominal and numeric variables present further complications as the interrelationships between nominal and numeric variables have also to be taken into account. A methodology that assists domain experts in identifying dimensions and facts of interest is highly desirable in these types of environments.

In this paper, we address these issues by proposing a knowledge discovery methodology that utilizes a combination of machine learning and statistical methods to identify interesting regions of information in large multi-dimensional data cubes. We utilize hierarchical clustering to construct data cubes at multiple levels of data abstraction. At each level of data abstraction, we apply well-known dimension reduction techniques such as Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA)

* Corresponding author. Address: WT 405A, AUT Tower, Private Bag 92006, 2-14 Wakefield St., Auckland 1010, New Zealand. Tel.: +64 9 921 9999x8953; fax: +64 9 921 9944.

E-mail addresses: muhammad.usman@aut.ac.nz, usmanspak@yahoo.com (M. Usman).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.11.008

in order to identify the most informative dimensions and facts present in data cubes. PCA was designed to operate with numeric data, whereas MCA works with nominal data. Each of these techniques on its own can operate on appropriate partitions of the original dataset and produce its own sets of variables that are most informative. The research challenge then becomes the integration of the partition containing candidate numeric variables with the other partition containing the most informative nominal variables.

We make two main contributions in this paper. Firstly, we generate cubes at different levels of data abstraction and study the effect of abstraction level on information content. Secondly, at each level of data abstraction we identify the most significant interrelationships that exist between numeric and nominal variables, thus enabling the cubes of interest to be identified.

The rest of the paper is organized as follows. In the next section, we review previous work in the area of mixed data analysis and assistance towards intelligent exploration of data cubes. In Section 3 we give an overview of the two main statistical approaches used in this paper, namely PCA and MCA. Section 4 presents an overview of our proposed methodology and illustrates the methodological steps with a hypothetical example. Our real world case studies presented in Section 5 show that the knowledge discovered is significant from both the application and statistical perspectives. Finally, we summarize our research contributions in Section 6 and outline directions for future research.

2. Related work

We review two themes of research that are most relevant to the current research. Firstly, we discuss previous research in identifying data cubes that hold the greatest information content. Secondly, we discuss research carried out so far in identifying relationships between mixed data types.

Sarawagi et al. [26] explored methods for guiding the user towards interesting cube regions. The authors focused on identifying regions within the data where cells contained values that were significantly different from the expected threshold value calculated via a regression model. This work was extended further by Sarawagi [25], whereby differences in cell values across regions were used to find surprising information in unexplored areas of a data cube based on the concept of maximum entropy.

According to Kumar et al. [16], the work done in [25,26] defined surprises in a rigid manner, implying that users cannot view them differently according to their needs. Furthermore, the discovered surprises are not easy to understand and interpret by merely scanning high dimensional data presented in a large number of rows and columns. Kumar et al. [16] overcame these limitations by providing an algorithm for detecting surprises defined by users and establishing the concept of cube navigation using the detected surprises. Our research exhibits some similarities with [16,25] in the sense that we also assist the user by providing candidate interesting cube regions for exploration concealed at multiple levels of data abstraction. Moreover, we provide greater flexibility to the users by providing ranked paths (interesting regions) for discovering significant information at various levels of data granularity.

In the medical field, statistical methods were applied by [22] to improve disease diagnostics. The authors proposed the integration of OLAP cube exploration with parametric statistical techniques to find significant differences in facts by identifying a small set of discriminating dimensions. However, this work is limited in terms of understanding the interrelationships between the significant facts and dimension variables. Additionally, the work can only be utilized after the construction of a data cube. There is no facility for the user to construct a constrained data cube having important dimensions and facts beforehand for further cube exploration, as proposed in this paper. Furthermore, their application of statistical tests requires domain knowledge in order to pinpoint regions where the significant fact differences are suspected.

Moreover, as Koh et al. [14] argue, heavy reliance on domain specific information only leads to the discovery of known patterns that fit a preconceived template and has a danger of inhibiting the discovery of unknown hidden patterns present in the data. Motivated by this, they proposed a generic solution for discovering informative rules by automatically assigning item weights based on the total number of items and strength of interactions between them in a transactional database. In order to evaluate the informative rules generated by their proposed method, they utilized PCA to capture the amount of variance by each rule term (the actionable component of the rule). The higher the variance captured for a rule term, the greater the significance of the rule is as a whole. Our work is similar as we also provide a solution without reliance on domain specific information. However, our proposal differs from their work as we provide interesting cube regions instead of informative association rules. Additionally, our use of PCA is not to evaluate the results but to rank the numeric variables in order of significance. This provides more flexibility for the user to choose the variables (facts) of his/her own choice.

To date a large number of approaches have been proposed for finding interesting information from large collections of data. However, these approaches mostly target a specific data type. In the real world, datasets have a mix of numeric and nominal variables, often involving high cardinality nominal variables, thus challenging the analytical capability of the methods employed. To tackle this long-standing problem a variety of clustering algorithms have been proposed [2–4,18,6,11,12,19] ranging from hierarchical clustering, k-means clustering, fuzzy clustering and incremental clustering algorithms. However, none of these approaches have been integrated with statistical approaches to provide assistance towards the discovery of interesting information, as proposed in this paper. The main consequence of not using statistical methods in the past is that it led users to the discovery of previously known patterns because the data exploration process heavily relied on user's subjective knowledge. In the real world, it is extremely hard for novice users, and even for experts, to have a clear idea of the underlying data in a large multi-dimensional space. Statistical methods such as PCA and MCA help in constraining a large multi-dimensional space by filtering out the less informative dimensions and retaining the important ones. This allows users to have meaningful statistical information that they can use together with any specialized domain knowledge that may be relevant to identifying cubes of interest.

We close this section by presenting the overall limitations of previous research related to the problem that we are examining. It is evident that limited research has been conducted in the area of finding interrelationships between numeric and nominal variables that are increasingly becoming common in real-world datasets. Moreover, the predominant statistical techniques for analysis of such variables lack integration with machine learning methods for finding interrelationships between variables.

3. Fundamentals of statistical techniques adopted

In this section, we present an overview of the principles governing PCA and MCA in view of the central importance of these techniques in our proposed methodology.

3.1. Principal Component Analysis

Principal Component Analysis (PCA) is a popularly used statistical technique that has been applied to a wide variety of applications