Data Pattern Tutor for AprioriAll and PrefixSpan

Mohammed Alshalalfa
University of Calgary, Canada

Ryan Harrison
University of Calgary, Canada

Jeremy Luterbach
University of Calgary, Canada

Keivan Kianmehr
University of Calgary, Canada

Reda Alhajj
University of Calgary, Canada

INTRODUCTION

Data mining can be described as data processing using sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large pre-existing databases (Agrawal & Srikant 1995; Zhao & Sourav 2003). From these patterns, new and important information can be obtained that will lead to the discovery of new meanings which can then be translated into enhancements in many current fields.

In this paper, we focus on the usability of sequential data mining algorithms. Based on a conducted user study, many of these algorithms are difficult to comprehend. Our goal is to make an interface that acts as a "tutor" to help the users understand better how data mining works. We consider two of the algorithms more commonly used by our students for discovering sequential patterns, namely the AprioriAll and the PrefixSpan algorithms. We hope to generate some educational value, such that the tool could be used as a teaching aid for comprehending data mining algorithms. We concentrated our effort to develop the user interface to be easy to use by naïve end users with minimum computer literacy; the interface is intended to be used by beginners. This will help in having a wider audience and users for the developed tool.

BACKGROUND

Kopanakis and Theodoulidis (2003) highlight the importance of visual data mining and how pictorial representation of data mining outcomes are more meaningful than plain statistics, especially for non-technical users. They suggest many modeling techniques pertaining to association rules, relevance analysis, and classification. With regards to association rules they suggest using grid and bar representations for visualizing not only the raw data but also support, confidence, association rules, and evolution of time.

Eureka! is a visual knowledge discovery tool that specializes in two dimensional (2D) modeling of clustered data for extracting interesting patterns from them (Manco, Pizzuti & Talia 2004). VidaMine is a general purpose tool that provides three visual data mining modeling environments to its user: (a) the meta-query environment allows users through the use of "hooks" and "chains" to specify relationships between the datasets provided as input; (b) the association rule environment allows users to create association rules by dragging and dropping items into both the IF and THEN baskets; and (c) the clustering environment for selecting data clusters and their attributes (Kimani, *et al*., 2004). After the model derivation phase, the user can perform analysis and visualize the results.

MAIN THRUST

AprioriAll is a equential data pattern discovery algorithm. It involves a sequence of five phases that work together to uncover sequential data patterns in large datasets. The first three phases, Sorting, L-itemset, and Transformation, take the original database and prepare the information for AprioriAll. The Sorting phase begins by grouping the information, for example a list of customer transactions, into groups of sequences with customer ID as a primary key. The L-itemset phase then scans the sorted database to obtain length one itemsets according to a predetermined minimum support value. These length one itemsets are then mapped to integer value, which will make generating larger candidate patterns much easier. In the Transformation phase, the sorted database is then updated to use the mapped values from the previous phase. If an item in the original sequence does not meet minimum support, it is removed in this phase, as only the parts of the customer sequences that include items found in the length one itemsets can be represented.

After preprocessing the data, AprioriAll efficiently determines sequential patterns in the Sequence phase. Length K sequences are used to generate length K+1 candidate sequences until K+1 sequences can no longer be generated (i.e., K+1, is greater than the largest sequence in the transformed database. Finally, the Maximal Phase prunes down this list of candidates by removing any sequential patterns that are contained within a larger sequential pattern.

Although this algorithm produces the desired results, it has several disadvantages. The potential for huge sets of candidate sequences is a major concern (Pei, *et al*., 2001). This results in an immensely large amount of memory space being used, especially when databases contain several large sequences. Another disadvantage is the time required to process large datasets since the algorithm requires multiple passes over the database. Additionally, AprioriAll has some difficulty mining long sequential patterns.

PrefixSpan requires preprocessing the database into a database consisting of sequences. The initial step scans the database and finds all length-1 prefixes that meet the minimum support. Next, the search space is partitioned into chunks corresponding to each length-1 prefix found in the previous step. Finally, the subsets of each of the prefixes can be mined by constructing corresponding projected databases; then each will be mined recursively. The final result is a compiled table consisting of the prefixes, postfixes, and all sequential patterns generated by the algorithm.

The major cost of PrefixSpan is the cost of construction of all the projected databases. In its worst case, it will construct a projected database for every sequential pattern (Pei, *et al*., 2001). "If the number and/or the size of projected databases can be reduced, the performance of sequential data mining can be improved substantially." (Pei, *et al*., 2001). One solution is the Bi-Level Projection. The Bi-Level Projection is represented in a lower triangular matrix, which can be used to generate the candidate sequential patterns.

The data mining technology is becoming increasingly popular and attractive as prediction technique, especially for advanced and emerging challenging applications. To compensate for this trend, it is crucial that more time be spent in the teaching process of many data mining algorithms. However, current resources that could serve as a learning aid lack the necessary tools to successfully present the material in an easy to understand manner. Taking a closer look at both AprioriAll and PrefixSpan, we found that they are not as straight forward as one might have liked. Needless to say, our experience with beginners tells that someone who decides to learn these algorithms may be left frustrated and confused. On a positive note, a visual interface which could be used as a tutor is an ideal solution to overcome the frustration and confusion. Compiling as many of the current resources and ideas regarding these algorithms into one user friendly visual interface seems like a good start. We decided to start from the bottom up. This means a systematical completion of the entire design process. With some basic background in Human Computer Interaction principles, we implemented a design that is best suited to be a potential ideal learning tool for both algorithms.

While designing our interface we implemented several features that allow for easier learning. The first implemented feature is to provide to the user all the steps of the algorithm while giving the user the option to selectively choose what step they want to display. Another feature of our interface is the clear textural description that couples with the selected step. Our system allows for some aspects of the interface to be interactive and better demonstrate how an algorithm works. Our interface also allows for a slideshow type feature to consecutively traverse all steps in an algorithm. A feature that we felt quite useful is to include