BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE,
Hyderabad Campus

DATA MINING REPORT: ASSIGNMENT 1

NAME:	ID NO.:
Anirudh Ravi	2011C6PS575H
Philip Joseph Thomas	2011A3PS272H
Satya Narayan Das	2011A7PS219H

Submitted in partial fulfilment of requirements of

CS F415 (Data Mining)

April 2015

Contents

Introduction to Association Analysis

Association analysis has been broadly used in many application domains. One of the best known is the business field where the discovering of purchase patterns or associations between products is very useful for decision making and for effective marketing.

Some examples of recent applications are finding patterns in biological databases, extraction of knowledge from software engineering metrics or obtaining user's profiles for web system personalization.

Association analysis is considered an unsupervised technique, so it has been applied in knowledge discovery tasks. Recent studies have shown that knowledge discovery algorithms, such as association rule mining, can be successfully used for prediction in classification problem.

Association rule mining discovers the frequent patterns among the itemsets. It aims to extract interesting associations, frequent patterns, and correlations among sets of items in the data repositories.

Objectives:

Association rules are the {if/then} type of statements that help uncover relationships between seemingly unrelated data in a transactional database. The database need not be transactional. It can be relational database or other information repository as well. This is a popular method of discovering interesting variables in large databases. Association rules mining falls under the domain of Data Mining.

It is quite evident that by forming association rules, one tries to find patterns in the database. However, finding patterns does not gives us the information needed by the customer. These patterns need to be translated into association rules to enable a layman to get an insight in the data. Association rules are something that can be easily understood by the customer. Finding association rules have its own benefits.

Association rule mining can help make decisions in marketing and business decisions. Now, for instance, if in a certain sales outlet of a certain company, Cheetos and Kurkure are being sold along with other general stuff. So, through association rule mining, on can discover a new rule; for example, that the customers buying Cheetos and Kurkure, have a higher

tendency of purchasing Lays as well. In addition to the above example, association rule mining is used in many application-based areas including web usage mining, intrusion detection, continuous production, bioinformatics, etc.

The choice of Algorithm

The main parameters to be taken into account by humans to choose the proper data mining technique in a real application are:

- The main goal of the problem to be solved
- The structure of the available data

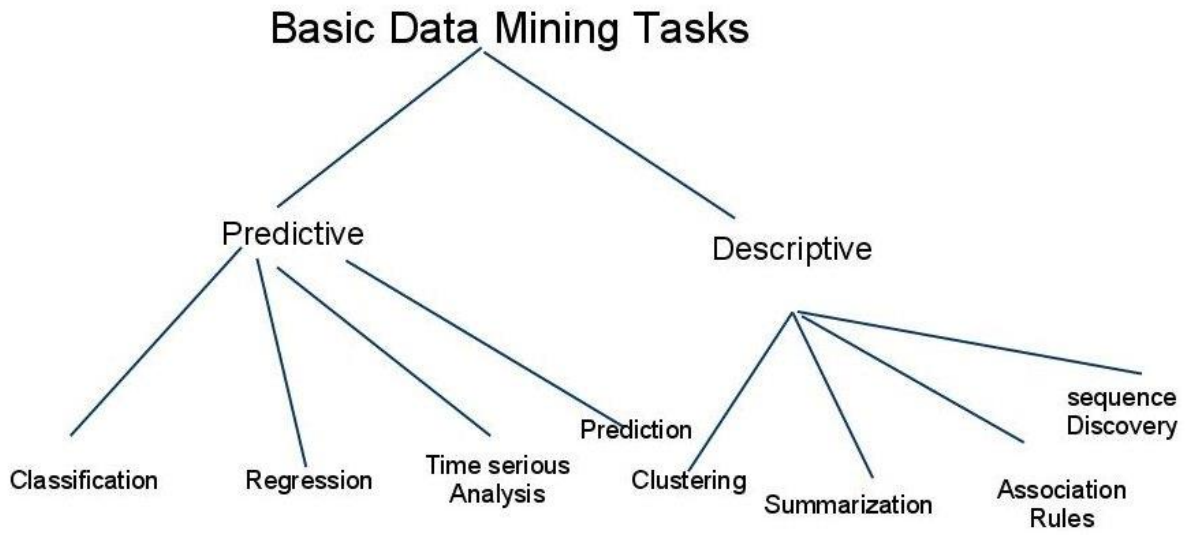

fig. 1

According to that, we elaborate the classification displayed in Fig. 1. Which includes some of the most popular data mining techniques useful for environmental scientists.

The higher level division is taking into account the basic distinction between having and not a reference variable to be explained (response variable). Left part of the tree refers to non-supervised methods, i.e. without response variable, in where the main goal is a better cognition of the target phenomenon and description is enough as a result. Whereas right part of the tree refers those supervised models oriented to re-cognition, where a response variable is to be explained and prediction is pretended.

At a second level, for methods oriented to description, the main division regards the interest of describing relationships between objects (rows of data matrix), which are labelled as descriptive methods or describing relationships between variables (columns of data matrix), labelled as associative methods. For methods oriented to prediction, here the main distinction regards the nature of the response variable: while discriminant methods explain or predict qualitative variables, the classical predictive methods refer to quantitative response variables. Because of variety, discriminant models include a further level of subdivision.

Rule-based reasoning methods group methods providing explicit knowledge model, which can be expressed by formal rules or not, to be applied for further prediction; in case-based reasoning methods the predictive model is implicit in historical data; the third option is a mixture between prior explicit knowledge model and iterative refinements based on future data (Bayesian learning).

Hash-Tree

Hash trees can be computed in parallel. If you have two blocks of data to hash, you can use two processors to compute the hash twice as fast. This only works if your hash speed is lower than your IO speed, which is unlikely. Hash trees can be computed from hashes of individual blocks, or from hashes of larger sections that are aligned correctly. This is important.

For example, if I want to send you a file, I can break it up into chunks of 1 MB and send you each chunk with its SHA-256 hash. If the hash for any of the individual chunks is incorrect, then you can ask for that chunk again. At the end, I can sign the tree hash for the file and send you the signed hash. You can verify the hash just by hashing each of the block hashes (which you already verified), which is a lot faster than rehashing the entire file.

A tree hash is advantageous any time that you want to compute the hash of both a portion of a file and the entire file. The "extra work" for computing a tree hash is actually minimal. Yes, it does require computing extra hashes -- but only O (1) extra work.

FP-Growth Algorithm:

FP-Growth that stands for Frequent Pattern Growth Algorithm is a scalable technique for mining frequent itemset from a database. Frequent Pattern Growth adopts a divide and conquer strategy.

Below are some of the valid reasons for choice of the algorithm for association rule generation-

Technique:

The algorithm constructs conditional frequent pattern tree and conditional pattern base from data set(database) which satisfy minimum support.

Memory Utilization:

Due to compact structure and no candidate generation requires less memory.

Execution Time:

Execution time required for FP-Growth is less than Apriori Algorithm.

Number of Scans:

FP-Growth algorithm scans the database only twice.

Groceries Data Set

Description

The Groceries data set contains 1 month (30 days) of real-world point-of-sale transaction data from a typical local grocery outlet. The data set contains **9835 transactions** and the **items** are aggregated to **169** categories.

Source

The data set is provided for arules by Michael Hahsler, Kurt Hornik and Thomas Reutterer.

Rules Generated

The rules generated by using both **hash tree** and **fp growth** algorithm are as given below:

7 11 -> 5 Confidence: 30.978261

7 5 -> 11 Confidence: 41.379310

11 5 -> 7 Confidence: 53.395785

7 11 -> 42 Confidence: 30.978261

7 42 -> 11 Confidence: 47.401247

11 42 -> 7 Confidence: 48.927039

Summary

Many algorithms for generating association rules were presented over time. Some well-known algorithms are Apriori, Éclat and FP-Growth, but they only do half the job, since they are algorithms for mining frequent itemsets. Another step needs to be done after to generate rules from frequent itemsets found in a database.

One limitation of the standard approach to discovering associations is that by searching massive numbers of possible associations to look for collections of items that appear to be associated, there is a large risk of finding many spurious associations. These are collections of items that co-occur with unexpected frequency in the data, but only do so by chance.