* Q7.7 – The price of each item in a store is non-negative. The store manager is only interested in rules of certain forms, using the constraints given in (a)–(b). For each of the following cases, identify the kinds of **constraints** they represent and briefly discuss how to mine such association rules using **constraint-based pattern mining**.

1. Containing at least one Blu-ray DVD movie.

This would represent a monotonic constraint (rule constraint), as well as a weak succinct constraint. Once the itemset reaches a count of at least one Blu-ray DVD, then any additional tests on this itemset is unneeded, because all supersets would always satisfy the criteria.

1. Containing items with a sum of the prices that is less than $150.

This would represent an antimonotonic constraint (rule constraint). Each item added to the itemset would push it closer to going over the $150 criteria. Mining this type of association rule would be enacted at each iteration of a Apriori algorithm to check if it is no longer meeting the user-defined constraint. Once it is no longer meeting the constraint, all supersets will no longer meet the constraint either.

1. Containing one free item and other items with a sum of the prices that is at least $200.

The one free item represents a succinct constraint and the sum of at least $200 represents a monotonic constraint (rule constraint). The succinct constraint is precounting prunable, so item sets can be pruned prior to support counting since a specific formula can be used to check if the itemset contains any free items. Therefore, there is no need for iterative checks for this constraint. The monotonic constraint would be mined similar to part (a). Once the itemset reaches the $200 minimum, then any additional tests would be redundant, because all supersets would also meet the critera.

1. Where the average price of all the items is between $100 and $500.

Both of these constraints represent a convertible constraint (rule constraint). They can be antimonotonic or monotonic based on how the items are arranged and added to the itemset. The items would have to be added in either price-ascending or price-descending order. This would make one of the constraint antimonotonic and the other constraint monotonic (based on which price order the user decided on), which could then be used for efficient constraint-based pattern mining.

* Q7.10 - Association rule mining often generates a large number of rules, many of which may be similar, thus not containing much novel information. Design an efficient algorithm that **compresses** a large set of patterns into a small compact set. Discuss whether your mining method is robust under different pattern similarity definitions.

Once the frequent patterns have been mined, clustering can be used to compress the pattern and get a set of smaller representative pattern models. First, we would perform lossless compression of the data set by finding the set of closed frequent patterns. Next, we would the pattern distance formula to calculate the distances between the frequent patterns and k-means to defined clusters; however, γ-clusters can be utilized to overcome the weaknesses of k-means method of clustering and finding the frequent representative pattern that meets the minimum support value.

* Q7.11 - Frequent pattern mining may generate many superfluous patterns. Therefore, it is important to develop methods that mine compressed patterns. Suppose a user would like to obtain only *k* patterns (where *k* is a small integer). Outline an efficient method that generates the ***k* most representative patterns**, where more distinct patterns are preferred over very similar patterns. Illustrate the effectiveness of your method using a small data set.

For this type of problem, we are looking for k-summarized pattern strategy since it finds the most representative patterns. A user would define how many patterns they are looking for, k, and then algorithm would take into account objective significance measurements (i.e. support, confidence, etc.) and/or subjective significance measurements, as well as the redundancy values between patterns. In this example, more distinct patterns are preferred, so the method should put more weight on the redundancy measures to minimize the amount of redundancy and find the pattern that represents the most centered pattern. The redundancy-aware top-k pattern may also be useful in this situation if the user would make some trade-offs in redundancy for significance.

 

Method looking for significant or similar patterns

Method looking for unique and representative patterns

* Q7.12 - It is interesting to generate **semantic annotations** for mined patterns. Section 7.6.1 presented a pattern annotation method. Alternative methods are possible, such as by utilizing type information. In the DBLP data set, for example, authors, conferences, terms, and papers form multi-typed data. Develop a method for automated semantic pattern annotation that makes good use of typed information.

As discussed in the textbook, there are three main tasks we wish to accomplish in semantic annotations for mined patterns: selecting context units and assigning them weight, design similarity measures for patterns, extract the most significant context indicators, representative transactions, and semantically similar patterns.

Therefore, the method for an automated semantic pattern annotation for typed information would:

* + Select desired context units, typically these are chosen to be frequent patterns within the data set.
  + Select the important and nonredundant patterns via compression, clustering, or other means on the data to get the most representative patterns.
  + Assign a weighting function to the data set via the mutual information computation.
  + Significant context indicators can be selected by cosine similarity on pairs of context vectors.
  + Representative transactions would be extracted by treating every transaction as a context vector and using the cosine similarity to rank the transactions semantic similarity.
  + Semantically similar patterns would be found by ranking each frequent pattern and similarity via cosine similarity to the model and context of the data set.
  + These outputs can then be displayed in a table or other format for the user to view and better understand the mined patterns.

