Essay Big Data

Jasper Robeer
3802337

Utrecht University
February 19, 2017

1 Terminology

The discovery of **frequent itemsets** is one of the major families of techniques for characterizing data. To explain how frequent itemsets work, we will take a look at the *market-basket* model first. The market-basket model and frequent itemsets problem originated in the analysis of true market baskets. Retailers wanted to learn what items are frequently bought together. In the *market-basket* model we look at the many-to-many relationship between the *items* and the *baskets*. The *frequent itemsets* problem follows from this model. We are concerned with finding the sets of items that appear in many of the same baskets. To give a more formal definition: we have a number s, which is known as the *support threshold*. The *support* for a set of items I is the number of baskets for which I is a subset. We say that a set of items I is frequent if its supports is s or more.

Association Rules are a collection of if-then rules. They are often used to represent the information of frequent itemsets. For an *association rule* $I \to j$: I is a set of items and j is an item. The implication of such a rule is that if all of the items in I appear in some basket, then j is likely to appear as well. We define likely as the *confidence* of the association rule $I \to j$. The *confidence* is the ratio of the support for $I \cup \{j\}$ to the support for I. More intuitively, the confidence of the rule is the fraction of the baskets with all of I that also contain j. The *interest* of an association rule $I \to j$ is the difference between its confidence and the fraction of baskets that contain j. A high *positive interest* means that the presence of I in a basket somehow causes the presence of j. A high *negative interest* means that the presence of I somehow discourages the presence of j.