**Tour of Data Mining Algorithms**

Implementation Project

The Apriori algorithm has been fully implemented. The implementation is composed of an Apriori class, a HashTree class, and classes that make up the components of the hash tree, including Item and ItemSet classes. The implementation works primarily by making calls to the HashTree class, in order to manipulate the hash trees storing frequent itemsets and their occurrence counts. The Apache Commons Math library was used for a fast implementation of the binomial coefficient function. This function is used when visiting a leaf node with a transaction to determine whether or not all possible k-itemsets would exceed the bucket size limit. The Apriori implementation has been successfully tested using the Belgian retail market data set. The algorithm is very computation intensive, taking a significant amount of time to form 4-itemset trees and above.

The x-means algorithm is finishing up implementation with a visual representation of the clusters. The kd tree was difficult to implement but was better than just using a list of points that are in the data set. The x-means algorithm is also very computation intensive; measuring a lot of distances takes quite a bit of time.

Certain aspects of the Apriori algorithm’s hash tree behavior were unclear from the documentation, with further research failing to produce further clarification. As a result, some decisions and assumptions were made in order to complete the implementation. One such issue had to do with how an itemset should be added to the hash tree. The algorithm sets a limit on the number of itemsets a leaf node may store (max bucket size). When the bucket limit has been reached and another itemset is to be added to that bucket, the algorithm is to convert the node into an interior hash node, and split the itemsets in the bucket among new child nodes. To determine which child nodes to assign the itemsets to, the itemsets are hashed on their next unhashed item. But what happens when there is no next unhashed item? Our implementation handles this by simply exceeding the bucket size limit in this case.

Another point of uncertainty was regarding whether or not a sepearate hash tree should be created for itemsets of a particular length. In other words, should there be a 1-itemset tree, a 2-itemset tree, and so on? For our implementation, this was the route we went, resulting in an array of hash trees.

The x-means algorithm implementation started with implementing k-means first, which worked well. Then we implemented the kd tree to speed up the process of finding clusters. While testing k-means we used a 2-Gaussian distribution with a k of 2 with the k means. This worked well but it was hard to know for sure without visualizing it. The Weka helped in figuring that out. There was a small issue when figuring out when to reevaluate the centroids. A few implementations sent the program into an infinite loop. Limiting the number of loops helped.

The command line interface for our project has also been implemented fully for Apriori. The Apache Commons CLI library was used to create the command line interface. When calling the program from the command line, the algorithm and input file must be specified, along with algorithm-specific options. Apriori-specific options are the minimum support for an itemset to be considered frequent (required), the number of children per node of a hash tree, and the maximum bucket size of a bucket node. General optional options may be specified, such as specifying a specific name for the output file, or specifying the delimiter used to separate attributes in the input file. A usage and help message is also provided when the help option is given, or when invalid or incomplete options are specified.

For the ID3 algorithm, we have just begun implementing it using a unique binary tree.

**References**

Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." *Proc. 20th int. conf. very large data bases, VLDB*. Vol. 1215. 1994.

Coffman Jr, Edward G., and James Eve. "File structures using hashing functions." *Communications of the ACM* 13.7 (1970): 427-432.

*Association Analysis: Basic Concepts and Algorithms* from Introduction to Data Mining by Tan, Steinbach, Kumar.

*An intoductory tutorial on kd-trees* by Andrew W. Moore published with Carnegie Mellon University

*Data Mining: Concepts and Techniques*, 3rd Ed, by Jiawei Han.