CSE 601 – Data Mining and Bioinformatics (Fall 2017)

Project 1:

Dimensionality Reduction & Association Analysis

Part 2: Association Analysis

Team Members:

1. **Siddharth Pateriya** spateriy@buffalo.edu **Person #: 50206348**
2. **Hrishikesh Saraf** hsaraf@buffalo.edu **Person #: 50205927**
3. **Ajay Gandhi** agandhi3@buffalo.edu **Person #: 50207403**

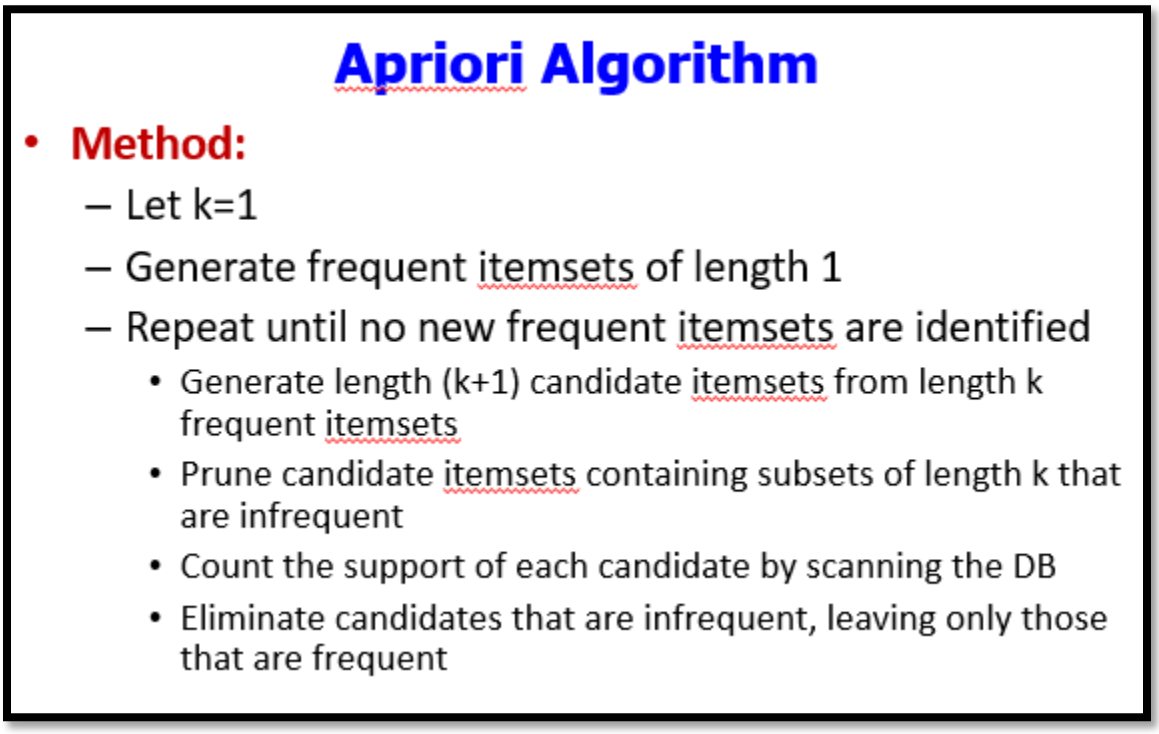
Part 2: Association Analysis

Apriori Algorithm:

- The Apriori algorithm is a data mining algorithm which is used for identifying attributes which occur together frequently and to generate/learn association rules by using these frequent item sets, over transactional databases.

- The process includes finding frequent items starting from frequent item list of length one and extending them to larger and larger item sets until no more frequent item sets can be found for given support values.

- The key principle of Apriori is as follows:
  "If an itemset is frequent, then all of its subsets must also be frequent".
  Using this principle, we are able to prune candidates in the early stages which are not frequent, so as to avoid using those sets to form larger item sets.



The flow of the algorithm is as follows:

1. First, we generate frequent item sets of length 1. To do this, we need to traverse the whole database and for each attribute, we see if the count of the attribute divided by

the number of total rows is greater than or equal to the given minSupport. If the support of the attribute is less than minSupport, it is not included in the frequent item sets of length 1.

2. Next, using the 1-length frequent item sets, we generate a 'candidate' list of length 2 sets. For example, if the 1 length frequent attributes are A,B,C, then the candidate list of length 2 would be AB, AC, BC. From these candidate sets, we prune(delete) the ones which have support less than minSupport. We then proceed to generate candidate sets of length n+1, and pruning them, until there are no more new frequent item sets being formed.

3. While generating candidate lists, if we need to generate n-length candidate sets, we combine two n-1 length sets, and we combine them if and only if the first (n-2) elements of the n-1 length sets are matching. Otherwise we don't combine the two sets.

4. When pruning, we check the support of each n-length combination, and see if it is greater than or equal to minSupport or not. If not, then we discard that frequent item set.

Rule Generation

- After we have the frequent item sets generated, we need to form association rules with them, which satisfy the given confidence criterion.

- For a rule "X -> (Y, Z)" , it's confidence C is defined as Support(X union Y,Z) divided by Support(X). That is, Confidence equals the support of (BODY union HEAD) divided by support of (BODY).

- To generate rules, we create subsets from each frequent item set, and check the confidence of each subset, and if it satisfies the minConfidence criteria.

- For example, a frequent set ABC will have the following subsets: A->B A->C B->C B->A C->A C->B

- We check confidence for each of the rules above, and those that have confidence greater than or equal to minConfidence are stored as association rules.

Output of Association Analysis:

Number of Frequent Itemsets:

For Minimum Support 30:

Number of Frequent Sets of length: 1 are 196
Number of Frequent Sets of length: 2 are 5340
Number of Frequent Sets of length: 3 are 5287
Number of Frequent Sets of length: 4 are 1518
Number of Frequent Sets of length: 5 are 438
Number of Frequent Sets of length: 6 are 88
Number of Frequent Sets of length: 7 are 11
Number of Frequent Sets of length: 8 are 1
Total: 12879

For Minimum Support 40:

Number of Frequent Sets of length: 1 are 167
Number of Frequent Sets of length: 2 are 753
Number of Frequent Sets of length: 3 are 149
Number of Frequent Sets of length: 4 are 7
Number of Frequent Sets of length: 5 are 1
Total: 1077

For Minimum Support 50:

Number of Frequent Sets of length: 1 are 109
Number of Frequent Sets of length: 2 are 63
Number of Frequent Sets of length: 3 are 2
Total: 174

For Minimum Support 60:

Number of Frequent Sets of length: 1 are 34
Number of Frequent Sets of length: 2 are 2
Total: 36

For Minimum Support 70:

Number of Frequent Sets of length: 1 are 7
Total: 7

Association Rules Based on Templates:

Based on Support = 50% and Confidence = 70%

Template	Query based on Template	Number of Rules Generated
Template 1 – 1	RULE HAS ANY OF (G59_Up)	26
Template 1 – 2	RULE HAS NONE OF (G59_Up)	91
Template 1 – 3	RULE HAS 1 OF (G59_Up,G10_Down)	39
Template 1 – 4	BODY HAS ANY OF (G59_Up)	9
Template 1 – 5	BODY HAS NONE OF (G59_Up)	108
Template 1 – 6	BODY HAS 1 OF (G59_Up, G10_Down)	17
Template 1 – 7	HEAD HAS ANY OF (G59_Up)	17
Template 1 – 8	HEAD HAS NONE OF (G59_Up)	100
Template 1 – 9	HEAD HAS 1 OF (G59_Up, G10_Down)	24
Template 2 – 1	SizeOf(RULE) >= 3	9
Template 2 – 2	SizeOf(BODY) >= 2	6
Template 2 – 3	SizeOf(HEAD) >= 1	117
Template 3 – 1	BODY HAS ANY OF (G10_Down) OR HEAD HAS 1 OF (G59_Up)	24
Template 3 – 2	BODY HAS ANY OF (G10_Down) AND HEAD HAS 1 OF (G59_Up)	1
Template 3 – 3	BODY HAS ANY OF (G10_Down) OR SizeOf(HEAD) >= 2	11
Template 3 – 4	BODY HAS ANY OF (G10_Down) AND SizeOf(HEAD) >= 2	0
Template 3 – 5	SizeOf(BODY) >= 1 OR SizeOf(HEAD) >= 2	117
Template 3 – 6	SizeOf(BODY) >= 1 AND SizeOf(HEAD) >= 2	3