



Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

Name:	Swarup Satish Kakade
Roll No:	17
Class/Sem:	TE/V
Experiment No.:	5
Title:	Using open-source tools Implement Association Mining Algorithms.
Date of Performance:	
Date of Submission:	
Marks:	
Sign of Faculty:	



Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

Aim: To implement Apriori Algorithm on a large dataset using Open-source tool WEKA.

Objective: To make students well versed with open-source tools like WEKA to implement Apriori algorithm.

Theory:

- Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently an itemset occurs in a transaction.
- A typical example is a Market Based Analysis. Market Based Analysis is one of the key techniques used by large relations to show associations between items.
- It allows retailers to identify relationships between the items that people buy together frequently.
- Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Support Count () – Frequency of occurrence of a itemset.

Here $(\{\text{Milk, Bread, Diaper}\}) = 2$

Frequent Itemset – An itemset whose support is greater than or equal to minsup threshold.

Association Rule – An implication expression of the form $X \Rightarrow Y$, where X and Y are any 2 itemsets.

Example: $\{\text{Milk, Diaper}\} \Rightarrow \{\text{Beer}\}$

- WEKA contains an implementation of the Apriori algorithm. The algorithm works only with discrete data.
- It can identify statistical dependencies between groups of attributes.
- Apriori algorithm can compute all rules that have a given minimum support and exceed a given confidence.
- Clicking on the "Associate" tab will bring up the interface for the association rule algorithms.



Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

- The Apriori algorithm which we will use is the default algorithm selected. However, in order to change the parameters for this run (e.g., support, confidence, etc.) we click on the text box immediately to the right of the "Choose" button. Note that this box, at any given time, shows the specific command line arguments that are to be used for the algorithm.
- WEKA allows the resulting rules to be sorted according to different metrics such as confidence, leverage, and lift. We can also change the default value of rules (10) to be 20; this indicates that the program will report no more than the top 20 rules. The upper bound for minimum support is set to 1.0 (100%) and the lower bound to 0.1 (10%).
- Apriori in WEKA starts with the upper bound support and incrementally decreases support (by delta increments which by default is set to 0.05 or 5%). The algorithm halts when either the specified number of rules are generated, or the lower bound for min. support is reached. Once the parameters have been set, the command line text box will show the new command line. We now click on start to run the program. This results in a set of rules. The panel on the left ("Result list") now shows an item indicating the algorithm that was run and the time of the run. You can perform multiple runs in the same session each time with different parameters. Each run will appear as an item in the Result list panel. Clicking on one of the results in this list will bring up the details of the run, including the discovered rules in the right panel. In addition, right-clicking on the result set allows us to save the result buffer into a separate file. Note that the rules were discovered based on the specified threshold values for support and lift. For each rule, the frequency counts for the LHS and RHS of each rule is given, as well as the values for confidence, lift, leverage, and conviction. In most cases, it is sufficient to focus on a combination of support, confidence, and either lift or leverage to quantitatively measure the "quality" of the rule. However, the real value of a rule, in terms of usefulness and action ability, is subjective and depends heavily on the particular domain and business objectives.

OUTPUT:

```
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Choose Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Start Stop
Result list (right-click for ...)
15:03:25 - Apriori

Associator output
=== Run information ===
Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: supermarket
Instances: 4627
Attributes: 217
[11 of attributes omitted]
=== Associator model (full training set) ===

Apriori
=====
Minimum support: 0.15 (694 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 44
Size of set of large itemsets L(2): 380
Size of set of large itemsets L(3): 910
Size of set of large itemsets L(4): 633
Size of set of large itemsets L(5): 105
Size of set of large itemsets L(6): 1

Best rules found:
```



Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

```
Weka Explorer
Preprocess  Classify  Cluster  Associate  Select attributes  Visualize
Associate
Choose  Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S 1.0 -c 1
Start  Stop
Result list (right-click for ...)
150525 - Apriori

Associate output
=== Run information ===
Scheme:  weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S 1.0 -c 1
Relation:  supermarket
Instances:  4627
Attributes:  217
[List of attributes omitted]
=== Associate model (full training set) ===

Apriori
=====
Minimum support: 0.15 (694 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:
Size of set of large itemsets L(1): 44
Size of set of large itemsets L(2): 300
Size of set of large itemsets L(3): 910
Size of set of large itemsets L(4): 633
Size of set of large itemsets L(5): 105
Size of set of large itemsets L(6): 1

Best rules found:
1. biscuits=> frozen food=> fruit=> total-high 768 ==> bread and cake=> 723  <conf:(0.92)> lift:(1.27) lev:(0.03) [155] convr:(3.35)
2. baking needs=> biscuits=> fruit=> total-high 760 ==> bread and cake=> 696  <conf:(0.92)> lift:(1.27) lev:(0.03) [149] convr:(3.28)
3. baking needs=> frozen food=> fruit=> total-high 770 ==> bread and cake=> 705  <conf:(0.92)> lift:(1.27) lev:(0.03) [150] convr:(3.27)
4. biscuits=> fruit=> vegetables=> total-high 815 ==> bread and cake=> 746  <conf:(0.92)> lift:(1.27) lev:(0.03) [159] convr:(3.26)
5. party snack food=> fruit=> total-high 854 ==> bread and cake=> 779  <conf:(0.91)> lift:(1.27) lev:(0.04) [144] convr:(3.15)
6. biscuits=> frozen food=> vegetables=> total-high 797 ==> bread and cake=> 725  <conf:(0.91)> lift:(1.26) lev:(0.03) [151] convr:(3.06)
7. baking needs=> biscuits=> vegetables=> total-high 772 ==> bread and cake=> 701  <conf:(0.91)> lift:(1.26) lev:(0.03) [145] convr:(3.01)
8. biscuits=> fruit=> total-high 954 ==> bread and cake=> 866  <conf:(0.91)> lift:(1.26) lev:(0.04) [179] convr:(3)
9. frozen food=> fruit=> vegetables=> total-high 834 ==> bread and cake=> 787  <conf:(0.91)> lift:(1.26) lev:(0.03) [156] convr:(3)
10. frozen food=> fruit=> total-high 949 ==> bread and cake=> 877  <conf:(0.91)> lift:(1.26) lev:(0.04) [179] convr:(2.92)
```

Conclusion:

1) Explain the main steps involved in the Apriori algorithm.

The Apriori algorithm works by:

1. **Identifying Frequent Itemsets:** Find itemsets that meet the minimum support threshold from the dataset.
2. **Pruning:** Eliminate infrequent itemsets that do not meet the support requirement.
3. **Generating Association Rules:** Create rules from the frequent itemsets based on minimum confidence and lift metrics.
4. **Repeating:** Continue the process iteratively until no new frequent itemsets can be identified.

2) What are the key parameters in the Apriori algorithm and how do they affect its performance?

The key parameters in the Apriori algorithm are:

1. **Minimum Support:** Determines the frequency needed for an itemset to be considered frequent. Higher values reduce the number of itemsets and speed up processing, but may miss interesting patterns.
2. **Minimum Confidence:** Sets the threshold for the reliability of association rules. Higher values produce stronger rules but fewer of them, while lower values can yield many rules with less significance.
3. **Minimum Lift:** Measures the strength of an association compared to random chance. It helps prioritize relevant rules.

Adjusting these parameters affects computational efficiency, the number of itemsets/rules generated, and



Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

the relevance of the findings.