Student: aksgu350 (Akshay Gurudath)

Student: kespa139 (Keshav Padiyar Manuru)

TDDD41/732A75: Association Analysis -1

Dataset:

In this exercise we are using Iris dataset, it consists of 4 feature variables - Sepal length and width, Petal length and width. We also have classs (Iris-setosa, Iris-versicolor and Iris-virginica) associdated with respective features. There are 150 data points where each class has 50 datapoints each.

Approach to Association Analysis

We loaded the iris dataset into weka, as we will be using Apriori algorithm for association analysis we discretized the 4 continuous feature variables into sets of bins. Then we created a new cluster variable using a clustering algorithm. Finally, using the association algorithm we generated association rules for the previously generated clusters where the clusters were consequents and features were antecedents.

Experiment 1: Number of Bins = 3 and Number of Clusters = 3; Clustering Algorithm: Simple K-Means

Following are the parameters used for Apriori Algorithm:

Parameter	Value
car	true
delta	0.05
lowerBoundMinSupport	0.1
metricType	Confidence
minMetric	0.75
numRules	50

Best rules found:

Cluster1

1. petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' 48 ==> cluster=cluster1 48 conf:(1)

Cluster2

1. petallength='(4.933333-inf)' petalwidth='(1.7-inf)' 40 ==> cluster=cluster2 40 conf:(1)

Cluster3

1. petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 50 ==> cluster=cluster3 50 conf:(1)

Here we notice that, there are simple rules formed with max confidence i.e 1. Also, we noticed that all of these rules are having very high support. Therefore, the advantage of this setting is that it forms simple associations with high confidence and support.

Experiment 2: Number of Bins = 4 and Number of Clusters = 3; Clustering Algorithm: Simple K-Means

Following are the parameters used for Apriori Algorithm:

Parameter	Value
car	true
delta	0.05
lowerBoundMinSupport	0.1
metricType	Confidence
minMetric	0.9
numRules	50

Best rules found:

Cluster1

1. petallength='(-inf-2.475]' petalwidth='(-inf-0.7]' 50 ==> cluster=cluster1 50 conf:(1)

Cluster2

1. petallength='(3.95-5.425]' petalwidth='(1.3-1.9]' 33 ==> cluster=cluster2 33 conf:(1)
2. sepallength='(5.2-6.1]' petallength='(3.95-5.425]' 32 ==> cluster=cluster2 32 conf:(1)

Cluster3

1. petallength='(5.425-inf)' petalwidth='(1.9-inf)' 19 ==> cluster=cluster3 19 conf:(1)
2. sepallength='(6.1-7]' petalwidth='(1.9-inf)' 18 ==> cluster=cluster3 18 conf:(1)

Again we notice that, there are simple rules formed with max confidence = 1. However, because of more bins there is lesser support for some of these rules.

Experiment 3: Number of Bins = 3 and Number of Clusters = 4; Clustering Algorithm: Simple K-Means

Following are the parameters used for Apriori Algorithm:

Parameter	Value
car	true
delta	0.05
lowerBoundMinSupport	0.1
metricType	Confidence
minMetric	0.75
numRules	50

Best rules found:

Cluster1

1. petallength='(4.54-5.13]' 29 ==> cluster=cluster1 25 conf:(0.86)

Cluster2

1. petalwidth='(1.06-1.3]' 21 ==> cluster=cluster2 18 conf:(0.86)
2. sepallength='(5.38-5.74]' 27 ==> cluster=cluster2 23 conf:(0.85)

Cluster3

1. sepalwidth='(2.96-3.2]' **class=Iris-virginica 21 ==> cluster=cluster3** 18 conf:(0.86)

Cluster4

1. petallength='(-inf-1.59]' petalwidth='(-inf-0.34]' 33 ==> cluster=cluster4 33 conf:(1)
2. sepallength='(4.66-5.02]' petalwidth='(-inf-0.34]' 17 ==> cluster=cluster4 17 conf:(1)
3. sepalwidth='(2.96-3.2]' petalwidth='(-inf-0.34]' 16 ==> cluster=cluster4 16 conf:(1)

In this case, since the number of clusters are more than the true number of classes, we see that in cluster 3 there are no rules formed only with the feature variables. This suggests that for this additional unnatural cluster association rules cannot be obtained from available feature variables. In addition, for remaining clusters the rules have been formed either with lower confidence or lower support.

Experiment 4: Number of Bins = 3 and Number of Clusters = 3; Clustering Algorithm: Hierarchical Clustering

Following are the parameters used for Apriori Algorithm:

Parameter	Value
car	true
delta	0.05
lowerBoundMinSupport	0.1
metricType	Confidence
minMetric	0.75
numRules	50

Following are the parameters used for Hierarchical Clustering Algorithm:

Parameter	Value
numClusters	3
linkType	Complete

Best rules found:

Cluster1

1. petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 50 ==> cluster=cluster1 50 conf:(1)
2. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' 36 ==> cluster=cluster1 36 conf:(1)
3. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]' 36 ==> cluster=cluster1 36 conf:(1)

Cluster2

1. sepalwidth='(-inf-2.8]' petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' 27 ==> cluster=cluster2 27 conf:(1)

Cluster3

1. sepallength='(5.5-6.7]' petallength='(4.933333-inf)' petalwidth='(1.7-inf)' 24 ==> cluster=cluster3 24 conf:(1)
2. sepallength='(5.5-6.7]' sepalwidth='(2.8-3.6]' petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' 18 ==> cluster=cluster3 18 conf:(1)

When compared bulding association rules using Simple K-Means and Hierarchical cluastering algorithm, we noticed one stark difference. The association rules for hierarchical clustering use 3 to 4 item sets compared with k-means which uses one or two itemsets. Therefore the rules generated for heirarchical clusters are