Clustering lab

Yi Yang (yiyan338) Helena Llorens Lluís (hllor282)

March 2025

1 Introduction

In this lab we use the data mining toolkit Weka to Cluster a given dataset and use association analysis to describe the clusters obtained.

2 Dataset

The dataset used in this lab is Iris dataset. It consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample, they are the length and the width of sepal and petal.

3 Methodology

3.1 SimpleKMeans

SimpleKMeans is a clustering algorithm that partitions a dataset into a predefined number of clusters (k) based on feature similarity. It starts by randomly selecting k initial centroids. Each data point is then assigned to the nearest centroid using Euclidean distance, which calculates the straight-line distance between points in a multi-dimensional space. After assignment, the centroids are updated as the mean of all points in their respective clusters. This process of assignment and update is repeated iteratively until the centroids stabilize or a maximum number of iterations is reached. The objective is to minimize the total within-cluster variance, ensuring that data points within each cluster are as close as possible to their centroid.

3.2 EM

EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. EM can decide how many clusters to create by cross validation, or you may specify apriori how many clusters to generate.

3.3 Apriori

Apriori Algorithm is a foundational method in data mining used for discovering frequent itemsets and generating association rules. Its significance lies in its ability to identify relationships between items in large datasets which is particularly valuable in market basket analysis. The Apriori Algorithm operates through a systematic process that involves several key steps: Identifying Frequent Itemsets,Creating Possible item groups,Removing Infrequent Item groups, Generating Association Rules

4 Experimental Results

4.1 Kmeans,3 clusters, 3 bins

Using the SimpleKmeans method with 3 clusters and 3 bins, we observed that Cluster 3 generated the most rules. We then applied the following settings to derive rules for the clusters.

Table 1: 3 clusters,3 bins, Kmeans

Parameter	**Value**
Delta	0.05
Lower Bound Min Support	0.01
Metric Type	Confidence
Min Metric	0.3
Num Rules	300
Upper Bound Min Support	1.0

Here are the association rules for the 3 clusters:

Table 2: Association Rules for Cluster 1

Antecedent	**Confidence**	**Lift**	**Leverage**	**Conviction**
petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]'	1	2.73	0.2	30.4
sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]'	1	2.73	0.14	20.9
sepalwidth='(-inf-2.8]' petalwidth='(0.9-1.7]'	1	2.73	0.13	19.63
sepalwidth='(-inf-2.8]' petallength='(2.966667-4.933333]'	1	2.73	0.13	19
sepallength='(5.5-6.7]' petalwidth='(0.9-1.7]'	0.97	2.66	0.15	12.03

Table 3: Association Rules for Cluster 2

Antecedent	**Confidence**	**Lift**	**Leverage**	**Conviction**
petallength='(4.933333-inf)' petalwidth='(1.7-inf)'	1	3.33	0.19	28

Table 4: Association Rules for Cluster 3

Antecedent	Confidence	Lift	Leverage	Conviction
petallength='(-inf-2.966667]'	1	3	0.22	33.33
petalwidth='(-inf-0.9]'	1	3	0.22	33.33
petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]'	1	3	0.22	33.33
sepallength='(-inf-5.5]' petallength='(-inf-2.966667]'	1	3	0.21	31.33
sepallength='(-inf-5.5]' petalwidth='(-inf-0.9]'	1	3	0.21	31.33
sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]'	1	3	0.21	31.33
sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]'	1	3	0.16	24
sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]'	1	3	0.16	24
sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]'	1	3	0.16	24
sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]'	1	3	0.16	24
sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]'	1	3	0.16	24
sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]'	1	3	0.16	24

4.2 Kmeans,3 clusters, 10 bins

Using the SimpleKmeans method with 3 clusters and 10 bins, then applied the following settings to derive rules for the clusters.

Table 5: 3 clusters,10 bins, Kmeans

Parameter	Value
Delta	0.05
Lower Bound Min Support	0.01
Metric Type	Confidence
Min Metric	0.5
Num Rules	50
Upper Bound Min Support	1.0

Here are the association rules for the 3 clusters:

Table 6: Association Rules for Cluster 1

Antecedent	Confidence	Lift	Leverage	Conviction
sepallength='(5.74-6.1]'	0.86	2.09	0.07	3.23
sepalwidth='(2.48-2.72]'	0.86	2.09	0.07	3.23

Table 7: Association Rules for Cluster 2

Antecedent	Confidence	Lift	Leverage	Conviction
sepalwidth='(2.72-2.96]'	0.88	3.2	0.1	4.36

Table 8: Association Rules for Cluster 3

Antecedent	Confidence	Lift	Leverage	Conviction
sepallength='(4.66-5.02]' petalwidth='(-inf-0.34]'	1	3.19	0.08	11.67

4.3 EM,3 clusters, 3 bins

Using the EM method with 3 clusters and 3 bins, then applied the following settings to derive rules for the clusters

Table 9: 3 clusters,3 bins, EM Algorithm

Parameter	**Value**
Delta	0.05
Lower Bound Min Support	0.01
Metric Type	Confidence
Min Metric	0.3
Num Rules	300
Upper Bound Min Support	1.0

Here are the association rules for the 3 clusters.

Table 10: Association Rules for Cluster 1

Antecedent	**Confidence**	**Lift**	**Leverage**	**Conviction**
petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]'	1	2.68	0.2	30.08
sepallength='(5.5-6.7]' petalwidth='(0.9-1.7]'	1	2.68	0.16	23.81
sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]'	1	2.68	0.14	20.68
sepalwidth='(-inf-2.8]' petalwidth='(0.9-1.7]'	1	2.68	0.13	19.43
sepalwidth='(-inf-2.8]' petallength='(2.966667-4.933333]'	1	2.68	0.13	18.8
petalwidth='(0.9-1.7]'	0.98	2.63	0.22	16.92

Table 11: Association Rules for Cluster 2

Antecedent	**Confidence**	**Lift**	**Leverage**	**Conviction**
petallength='(-inf-2.966667]'	1	3	0.22	33.33
petalwidth='(-inf-0.9]'	1	3	0.22	33.33
petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]'	1	3	0.22	33.33
sepallength='(-inf-5.5]' petallength='(-inf-2.966667]'	1	3	0.21	31.33
sepallength='(-inf-5.5]' petalwidth='(-inf-0.9]'	1	3	0.21	31.33
sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]'	1	3	0.21	31.33
sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]'	1	3	0.16	24
sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]'	1	3	0.16	24
sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]'	1	3	0.16	24
sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]'	1	3	0.16	24
sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]'	1	3	0.16	24
sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]'	1	3	0.16	24

Table 12: Association Rules for Cluster 3

Antecedent	**Confidence**	**Lift**	**Leverage**	**Conviction**
petallength='(4.933333-inf)' petalwidth='(1.7-inf)'	1	3.41	0.19	28.27

5 Summary and Conclusion

5.1 Experiment Overview

In this experiment, we used the K-Means and EM clustering algorithms and the Apriori method experimented with different numbers of bins to observe the changes in the experimental results.

5.1.1 K-Means Algorithm

- **3 clusters, 3 bins.**

  The generated association rules are quite general, with high confidence and lift.
  For example, the rule for Cluster 1, petallength="(2.966667-4.933333]" petalwidth="(0.9-1.7]", has a confidence of 1 and a lift of 2.73.
  Cluster 3 generated the most rules, indicating that the data points in this cluster have stronger associations.

- **3 clusters, 10 bins.**

  Due to the increase in the number of bins, the data is divided into finer intervals, resulting in more specific rules.

  The confidence and lift have decreased, but the rules are more targeted. For example, the rule for Cluster 2, sepalength='(2.72-2.96]', has a confidence of 0.88 and a lift of 3.2.

  Cluster 3 still generated a large number of rules, indicating that the data points in this cluster have strong associations even with finer binning.

5.1.2 EM Algorithm

- **3 clusters, 3 bins.**

  The EM algorithm, based on probabilistic models, can handle the overlap in data distribution.

  The generated association rules are similar to those of K-Means, but the EM algorithm tends to produce rules with higher confidence. For example, the rule for Cluster 2, petallength='(-inf-2.966667]', has a confidence of 1 and a lift of 3.

  The rule for Cluster 3, petallength='(4.933333-inf)' petalwidth='(1.7-inf)', has a confidence of 1 and a lift of 3.41, indicating strong associations in the high-value intervals for this cluster.

5.2 Discussion of Different Experimental Variations

5.2.1 Different Clustering Algorithms (K-Means / EM)

- **K-Means**

  Generates more general rules, suitable for preliminary analysis.

- **EM**

  Probabilistic model-based clustering algorithm, capable of handling overlapping data distributions.
  Generates rules with higher confidence, suitable for handling complex data distributions.

5.2.2 Different Number of Bins (3 bins /10 bins)

- **3 bins**

  Data is divided into wider intervals, generating more general rules with higher confidence and lift.

  Suitable for preliminary analysis, quickly identifying strong associations in the data.

- **10 bins**

  Data is divided into narrower intervals, generating more specific rules, but with decreased confidence and lift.
  Suitable for detailed analysis, uncovering more specific associations.