

COMP37332

WEKA: Clustering tutorial¹

In this tutorial, you will explore clustering techniques. In the first part you will look at partitioning, and in the second you will explore hierarchical clustering. Weka provides several clustering implementations, including

- *SimpleKMeans* (a partitioning method using k-means) and
- *Cobweb* (hierarchical clustering).

Task A: k-means clustering (*SimpleKmeans*)

SimpleKMeans algorithm automatically handles a mixture of categorical and numerical attributes. The algorithm automatically normalises numerical attributes when doing distance computations. It uses Euclidean distance measure to compute distances between instances and clusters.

A1. Load the weather dataset (weather.arff) into Weka.

A2. Select the Cluster panel and choose *SimpleKMeans* as clustering method. The default number of clusters is 2. Click on Start – the clustering result should be in the output window. The result window shows the centroid of each cluster as well as statistics on the number and percentage of instances assigned to different clusters. Cluster centroids are the mean vectors for each cluster (so, each dimension value in the centroid represents the mean value for that dimension in the cluster). Thus, centroids can be used to characterize the clusters – how would you interpret the two clusters?

A3. Visualise the clustering results by right-clicking the result set on the left "Result list" panel and selecting "Visualize cluster assignments". This pops up the visualization window. Note that the visualisation is a 2 dimensional plot. You can choose the cluster number and any of the other attributes for each of the three different dimensions available (x-axis, y-axis, and colour). Different combinations of choices will result in a visual rendering of different relationships within each cluster. Inspect each of the entries of the plot (varying the x- and y- axes).

A4. Save the results by clicking the Save button in the visualisation pop-up window. Open the resulting arff file (in an editor). Note that in addition to the *instance_number* attribute, WEKA has also added *Cluster* attribute to the original data set. Manually check the two clusters – does the clustering make sense to you (for example, taking the *play* attribute as the main feature for grouping the instances)?

A5. Change the number of clusters to 3 (click on *SimpleKMeans*), and analyse the output. Which clustering is better? What is the "within cluster sum of squared errors"? Find out what the *seed* parameter is about. Why is it important for k-means clustering? Repeat the experiment with different seed values and compare the results.

A6. Repeat the experiment with the data in iris.arff (run *SimpleKmeans* with $k = 3$). In this data set, we have *class* values assigned to each instance, so we can evaluate the clustering results by comparing the actual class of the instances in a given cluster (i.e. whether the clusters correspond to the known classes). This is facilitated through the option "Classes to clusters evaluation". In this mode Weka ignores the class attribute (as selected by the user) and generates the clustering. During the evaluation phase it

¹ Tailored after: <http://maya.cs.depaul.edu/~Classes/Ect584/Weka/k-means.html> and other resources.

assigns classes to the clusters, based on the majority value of the class attribute within each cluster. Then it computes the classification error, based on this assignment and shows the corresponding confusion matrix. Analyse the results for the iris data – what is the error rate? What class was difficult to characterise? Visualise the results (using different attributes for Y-axis) – the instances denoted as squares have been “incorrectly” clustered.

A7. Load the data from bank.arff (600 customer instances) in Weka. Analyse the data set and attributes. Repeat the experiment with this data by entering 6 as the number of clusters. Analyse the centroids – do the six clusters make sense? Visualise the results – select the cluster number as the x-axis (*X: Cluster (Nom)*), the instance number (assigned by WEKA) as the y-axis (*Y: Instance_number (Num)*), and the "sex" attribute as the colour dimension (*Colour: sex (Nom)*). This will result in a visualisation of the distribution of males and females in each cluster. Which clusters are dominated by males, and which are dominated by females? Repeat the experiment with 4 clusters – do they provide a better partitioning of the data set?

Task B: hierarchical clustering (Cobweb)

B1. Load the weather data and run *Cobweb*. Make sure you select the option *saveInstanceData* to *True* (click on *Cobweb*) and option *Use training set* (for the cluster mode). The output presents the resulting dendrogram (in a textual form) that can be visualised² by right-click on the experiment in the Result list (*Visualize tree*). Manually draw the dendrogram using the following interpretation

- node N or leaf N represents a sub-cluster whose parent cluster is N. So, a sub-tree



means that cluster 2 is a daughter (and a leaf node) of cluster 1, and cluster 1 is a daughter of cluster 0, etc. The clustering tree structure is shown as a horizontal tree, where sub-clusters are aligned at the same column. For example, cluster 1 has cluster 2 and cluster 3 as its sub-clusters (and maybe more, not represented here).

- The root cluster is 0. Each line with node 0 defines a sub-cluster of the root.
- The number in square brackets after node N represents the number of instances in the parent cluster N (so, cluster 1 for example, contains 5 original instances). Clusters with [1] at the end of the line are instances. To link instances to clusters, save the results in the arff format – for example, line

```
5,rainy,65,70,TRUE,no,cluster2
```

means that cluster2 is instance number 5 (numbers assigned by Weka). Analyse the dendrogram – compare the results to those obtained in A4.

B2. Examine and load the flag data-set (flagdata.arff) that represent flag attributes of some European countries. Repeat the experiment using *Cobweb* and setting the parameter C to 0.4. Visualise the dendrogram and right-click on cluster 8: it has 3 instances that should appear on the plot. By clicking on each of the instances, you can see that these flags belong to Malta, Cyprus and Portugal. Does this cluster make sense? Find out what are the clusters in which the UK flag has been clustered.

² The visualisation is not the best feature of Weka – the clustering results can be uploaded to visualisation software, for example XGobi (<http://www.research.att.com/areas/stat/xgobi/>).