Pattern Recognition - Lecture 8
More with Clustering

Dr. Dewan Md. Farid

Associate Professor, Department of Computer Science & Engineering
United International University, Bangladesh

April 10, 2018

Similarity-Based Clustering

A similarity-based clustering method (SCM) is an effective and robust clustering approach based on the similarity of instances, which is robust to initialise the cluster numbers and efficient to detect different volumes of clusters. SCM is a method for clustering a data set into most similar instances in the same cluster and most dissimilar instances in different clusters. The instances in SCM can self-organise local optimal cluster number and volumes without using cluster validity functions.

Similarity between Instances

Let's consider $sim(x_i, x_l)$ as the similarity measure between instances x_i and the lth cluster center x_l. The goal is to find x_l to maximise the total similarity measure shown in Eq. 1.

$$J_s(C) = \sum_{l=1}^{k} \sum_{i=1}^{N} f(sim(x_i, x_l)) \tag{1}$$

Where, $f(sim(x_i, x_l))$ is a reasonable similarity measure and $C = \{C_1, \cdots, C_k\}$. In general, the similarity-based clustering method uses feature values to check the similarity between instances. However, any suitable distance measure can be used to check the similarity between the instances.

Algorithm 1 Similarity-based Clustering

Input: $X = \{x_1, x_2, \cdots, x_N\}$ // A set of unlabelled instances.
Output: A set of clusters, $C = \{C_1, C_2, \cdots, C_k\}$.
Method:

```
1:  C = ∅;
2:  k = 1;
3:  C_k = {x_1};
4:  C = C ∪ C_k;
5:  for i = 2 to N do
6:     for l = 1 to k do
7:        find the lth cluster center x_l ∈ C_l to maximize the similarity
          measure, sim(x_i, x_l);
8:     end for
9:     if sim(x_i, x_l) ≥ threshold_value then
10:       C_l = C_l ∪ x_i
11:    else
12:       k = k + 1;
13:       C_k = {x_i};
14:       C = C ∪ C_k;
15:    end if
16: end for
```

SCM - An Example

Viewer

Relation: weather.symbolic-weka.filters.unsupervised.attribute.Remove-R5

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal
1	sunny	hot	high	FALSE
2	sunny	hot	high	TRUE
3	overcast	hot	high	FALSE
4	rainy	mild	high	FALSE
5	rainy	cool	normal	FALSE
6	rainy	cool	normal	TRUE
7	overcast	cool	normal	TRUE
8	sunny	mild	high	FALSE
9	sunny	cool	normal	FALSE
10	rainy	mild	normal	FALSE
11	sunny	mild	normal	TRUE
12	overcast	mild	high	TRUE
13	overcast	hot	normal	FALSE
14	rainy	mild	high	TRUE

Undo OK Cancel

Figure: Weather Nominal Data.

Nearest Neighbor (NN) Clustering

Instances are iteratively merged into the existing clusters that are closest. In NN clustering a threshold, t, is used to determine if instances will be added to existing clusters or if a new cluster is created. The complexity of the NN clustering algorithm is depends on the number of instances in the dataset. For each loop, each instance must be compared to each instance already in a cluster.

Thus, the time complexity of NN clustering algorithm is $O(n^2)$. We do need to calculate the distance between instances often, we assume that the space requirement is also $O(n^2)$.

Algorithm 2 Nearest Neighbor Clustering

Input: $D = \{x_1, x_2, \cdots, x_n\}$ // A set of instances.
A // Adjacency matrix showing distance between instances
Output: A set of C clusters.
Method:

1: $C_1 = \{x_1\}$;
2: $C = \{C_1\}$;
3: $k = 1$;
4: **for** $i = 2$ to n **do**
5: find x_m in some cluster C_m in C so that $dis(x_i, x_m)$ is the smallest;
6: **if** $dis(x_i, x_m) \leq$t, $threshold\_value$ **then**
7: $C_m = C_m \cup x_i$
8: **else**
9: $k = k + 1$;
10: $C_k = \{x_i\}$;
11: $C = C \cup C_k$;
12: **end if**
13: **end for**

Euclidean Vs. Manhattan distance

The distance between the two points in the plane with coordinate (x,y) and (a,b) is given by:

$$EuclideanDistance, (x, y)(a, b) = \sqrt{(x-a)^2 + (y-b)^2} \quad (2)$$

$$ManhattanDistance, (x, y)(a, b) = |x-a| + |y-b| \quad (3)$$

Ensemble Clustering

Ensemble clustering is a process of integrating multiple clustering algorithms to form a single strong clustering approach that usually provides better clustering results. It generates a set of clusters from a given unlabelled data set and then combines the clusters into final clusters to improve the quality of individual clustering.

- No single cluster analysis method is optimal.
- Different clustering methods may produce different clusters, because they impose different structure on the data set.
- Ensemble clustering performs more effectively in high dimensional complex data.
- It's a good alternative when facing cluster analysis problems.

Ensemble clustering (con.)

Generally three strategies are applied in ensemble clustering:

1. Using different clustering algorithms on the same data set to create heterogeneous clusters.
2. Using different samples/ subsets of the data with different clustering algorithms to cluster them to produce component clusters.
3. Running the same clustering algorithm many times on same data set with different parameters or initialisations to create homogeneous clusters.

The main goal of the ensemble clustering is to integrate component clustering into one final clustering with a higher accuracy.

Subspace Clustering

The subspace clustering finds subspace clusters in high-dimensional data. It can be classified into three groups:

1. Subspace search methods.
2. Correlation-based clustering methods
3. Biclustering methods.

A subspace search method searches various subspaces for clusters (set of instances that are similar to each other in a subspace) in the full space. It uses two kinds of strategies:

- Bottom-up approach - start from low-dimensional subspace and search higher-dimensional subspaces.
- Top-down approach - start with full space and search smaller subspaces recursively.

Subspace Clustering (con.)

A correlation-based approach uses space transformation methods to derive a set of new, uncorrelated dimensions, and then mine clusters in the new space or its subspaces. It uses PCA-based approach (principal components analysis), the Hough transform, and fractal dimensions.
Biclustering methods cluster both instances and features simultaneously, where cluster analysis involves searching data matrices for sub-matrices that show unique patterns as clusters.

Weka 3: Data Mining Software in Java

Weka (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, **clustering**, association rules, and visualization.
Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Clustering Algorithms in Weka 3

1. SimpleKMeans - Cluster using the k-Means method.
2. XMeans - Extension of k-Means.
3. DBScan - Nearest-neighbor-based clustering that automatically determines the number of clusters.
4. OPTICS - Extension of DBScan to hierarchical clustering.
5. HierarchicalClusterer - Agglomerative hierarchical clustering.
6. MakeDensityBasedCluster - Wrap a clusterer to make it return distribution and density.
7. EM - Cluster using expectation maximization.
8. CLOPE - Fast clustering of transactional data.
9. Cobweb - Implements the Cobweb and Classit clustering algorithms.
10. FarthestFirst - Cluster using the farthest first traversal algorithm.
11. FilteredClusterer - Runs a clusterer on filtered data.
12. sIB - Cluster using the sequential information bottleneck algorithm.

Weka GUI Chooser

Figure: Weka GUI Chooser.

Weka Explorer

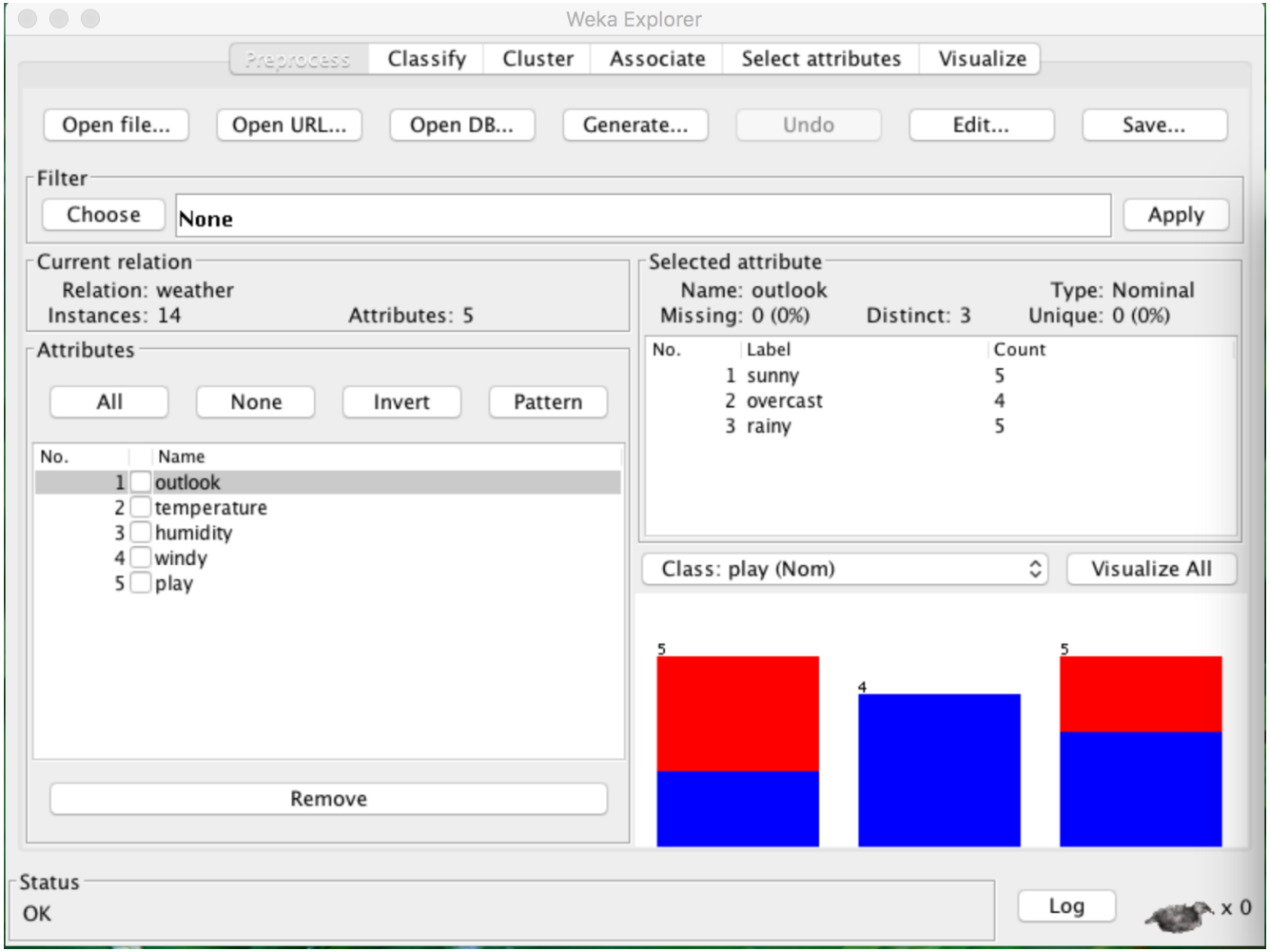

Figure: Weka Explorer.

Clustering using Weka

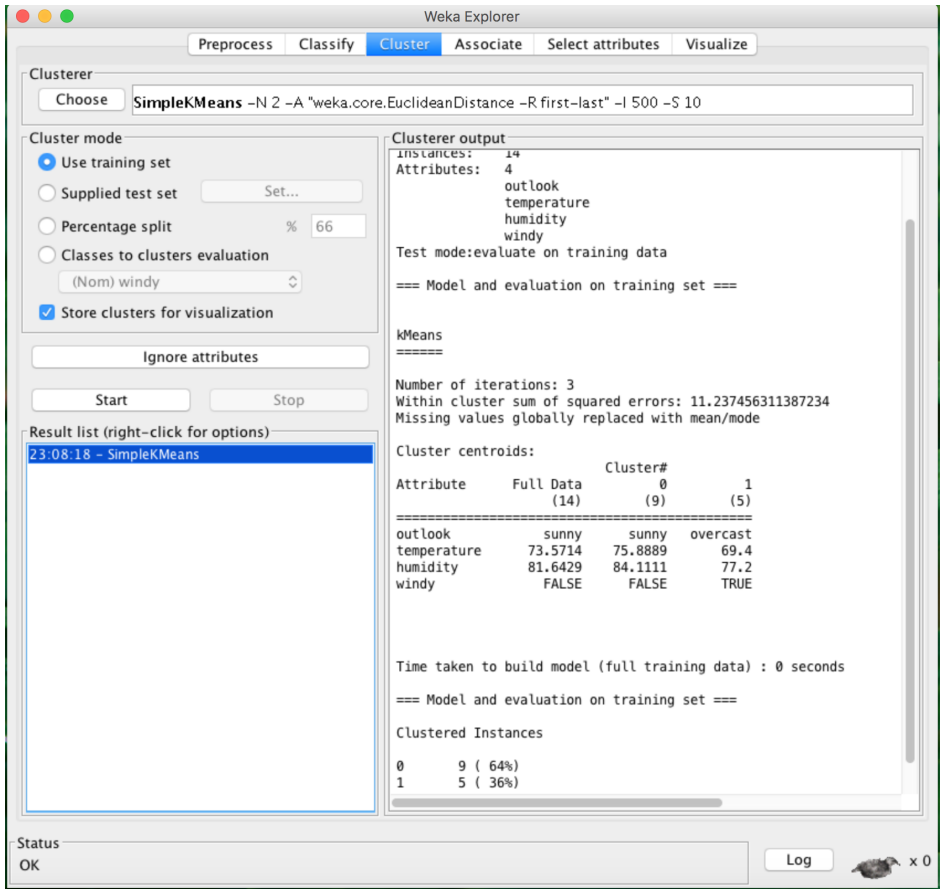

Figure: Cluster - Weka Explorer.

Reference Books

1. Data Mining Concepts and Technique, by Jiawei Han, Micheline Kamber, and Jian Pei (Third Edition)
2. Data Mining Practical Machine Learning Tools and Techniques, by Ian H. Witten, Eibe Frank, and Mark A. Hall (Third Edition)
3. Data Mining Knowledge Discovery and Applications, Edited by Adem Karahoca
4. Mining Complex Data, by Djamel A. Zighed, Shusaku Tsumoto, Zbigniew W. Ras, and Hakim Hacid

*** THANK YOU ***