Tools Pros and Cons of Clustering Algorithms using Weka Tools

Ayyoob.MP
Assistant Professor
Dept.of Computer Science
Sullamusallam Science College
Areacode-Kerala

ABSTRACT
Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. This paper analyzes three major clustering algorithms: K-Means, Hierarchical clustering and Density based clustering. The performance of these three clustering algorithms is compared using the clustering toolkit Weka.

General Terms
Data mining algorithms, Weka tools, K-means algorithms, Hierarchical clustering and Density based clustering.

1. INTRODUCTION
Data mining and knowledge discovery in databases have been an active area of research lately. Data mining is defined as the application of specific algorithms for extracting patterns from data [1].

Three of the major data mining techniques are regression, classification and clustering.

The open source clustering toolkit Weka is used for analyzing the algorithms (K-means algorithms, Hierarchical clustering and Density based clustering).

2. WEKA
Weka is considered as a landmark system in the history of the data mining among machine learning research communities [2].The toolkit has gained widespread adoption and survived for an extended period of time. The toolkit is developed at the University of Waikato, New Zealand. The acronym stands for Waikato Environment for Knowledge Analysis. Weka is platform-independent open source toolkit.

The GUI Chooser consists of four buttons:

1) Explorer: It provides an environment to explore data with WEKA.
2) Experimenter: It provides an environment to perform experiments and conducting statistical tests between learning schemes.
3) Knowledge Flow: This interface supports essentially the same functions as the Explorer but with drag and drop options. One advantage is that it supports incremental learning.
4) Simple CLI: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

3. CLUSTERING ALGORITHMS AND TECHNIQUES
The different kinds of clustering algorithms produce different sorts of representations of clusters. One way of imaging clusters is as disjoint sets. Find out the instance space and divide it into sets such that each part of the instance space is in just one cluster, this dividing the clusters to form a disjoint set.

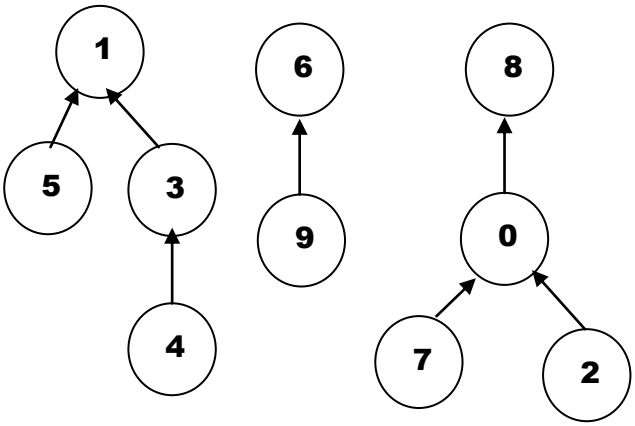

Fig 1: Disjoint set Clustering.

Another type of set is overlapping sets; here the clusters overlap, then the probabilistic assignment of instances is required.

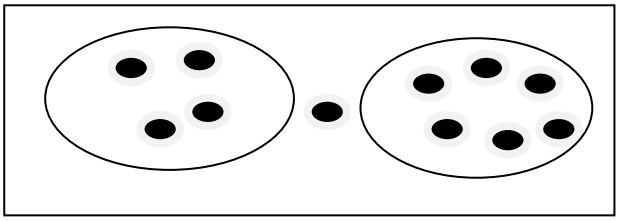

Fig 2: Overlapping sets.

Table 1. Probabilistic cluster

	1	2	3
A	0.3	0.1	0.2
B	0.5	0.4	0.3
C	0.2	0.1	0.6
D	0.6	0.4	0.5
E	0.6	0.2	0.4

Table [1] shows five instances A, B, C, D, E and three clusters. Instance A has a probability of 0.3 to belong to cluster 1, probability of 0.1 to belong to cluster 2 and probability of 0.2 to belong to cluster 3 and so on.

The third one is hierarchical clustering method.

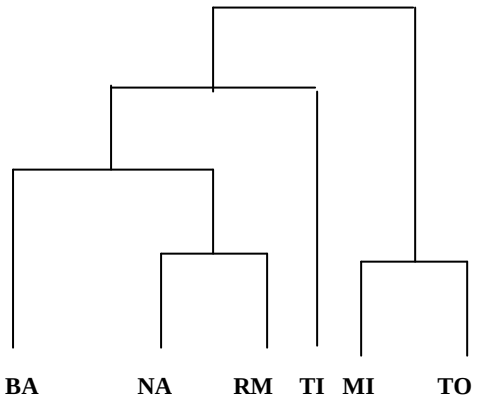

Fig 3: Hierarchical clustering.

Here the instances NA and RM reside at the bottom level. These clusters join together at the next level up to form a single higher level clusters. The process repeats until it reaches the topmost level where all the instances merge into a single big cluster. This representation is known as dendrogram.

There are many algorithms for clustering. Figure [4] shows three major clustering methods and their approach for clustering.

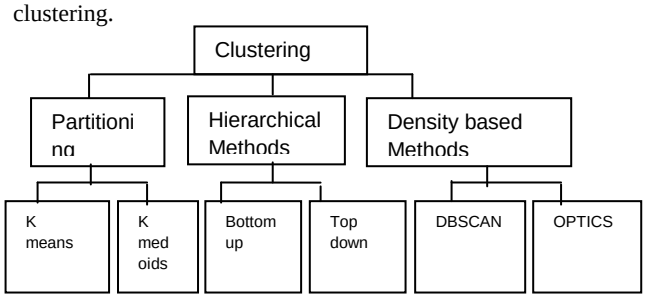

Fig 4: Clustering Algorithms.

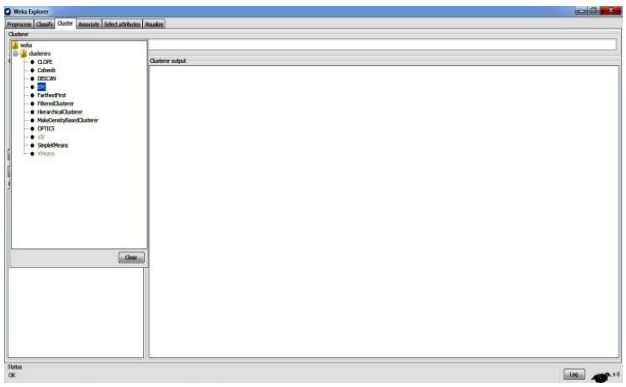

Fig 5: Various clustering algorithms in WEKA.

3.1 K-means Clustering

The term "*k*-means" was first used by James Mac Queen (1967)[3],though the idea goes back to 1957 [4]. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn't published until 1982. K-means is a widely used partitioned clustering method in the industries. The K-means algorithm is the most commonly used partitioned clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time.

Working of Algorithm [5].

K-Means Algorithm: In this method the cluster's center is represented by the mean value of the objects in the cluster.

Input: k: the number of clusters. D: a data set containing n objects.

Output: A set of k clusters.

Method:

1) Arbitrarily choose k objects from D as the initial cluster centers.
2) Repeat.
3) (Re) assign each object to the cluster to which the object is most similar based on the mean value of the objects in the cluster.
4) Update the cluster means.
5) Until no change.

Advantages

With a large number of variables, when k is small K-Means can computationally faster than hierarchical clustering [6].

1) K-means can produce tighter clusters compared to hierarchical clustering.

Disadvantages

1) It is sometimes difficult to compare the quality of the clusters produced.
2) It is difficult to fix the actual value of k.
3) It does not work well with non-globular clusters.

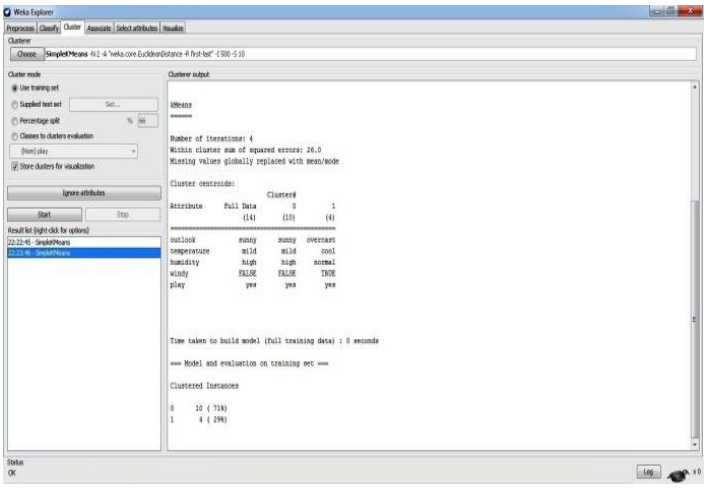

Fig 6: Result of k-means clustering.

All the three algorithms tested on the sample data set weather. Nominal which is available as part of the Weka toolkit.

The figure [6] shows the results of k-means clustering obtained using the dataset weather. Nominal. The result will be saved in ARFF file format which can also be opened in the MS Excel.

3.2 Hierarchical Clustering

It is a agglomerative (bottom up) clustering method.

1) Start with 1 point (singleton).

2) Recursively adds two or more appropriate clusters.

3) Stop when k number of clusters is achieved.

Divisive (top down)

1) Start with a big cluster.

2) Recursively divides into smaller clusters.

3) Stop when k number of clusters is achieved.

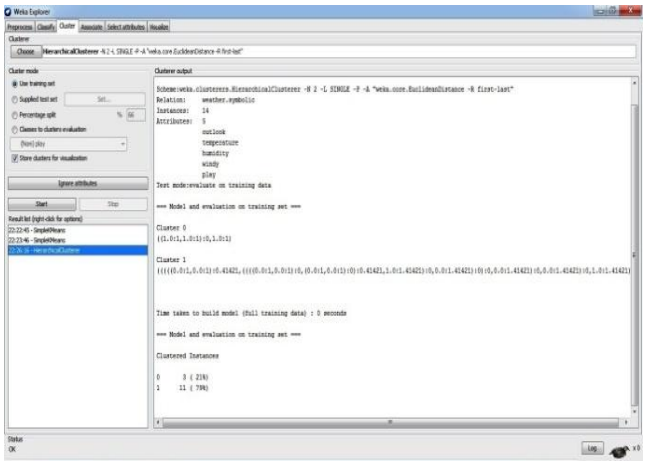

Fig 7: Result of Hierarchical Clustering.

Figure [7] shows the results of hierarchical clustering obtained using the dataset weather. Nominal.

Advantages

1) No apriority information about the number of clusters required.

2) Easy to implement and gives best result in some cases.

Disadvantages

1) It may not scale well

2) No automatic discovering of optimum clusters.

3.3 Density based Clustering

One of the most well known density-based clustering algorithms is the DBSCAN [7].

DBSCAN separates data points into three classes:

1) Core points: These are points that are at the interior of a cluster.

2) Border points: These points' falls within the neighborhood of a core point.

3) Noise points: A noise point is any point that is not a core point or a border point.

To find a cluster, DBSCAN starts with an arbitrary instance (p) in data set (D) and retrieves all instances of D with respect to Eps and Min Pts. The algorithm makes use of a spatial data structure(R*tree) to locate points within Eps distance from the core points of the clusters[8].Another density based algorithm OPTICS is introduced[9],which is an interactive clustering algorithm, works by creating an ordering of the data set representing its density-based clustering structure.

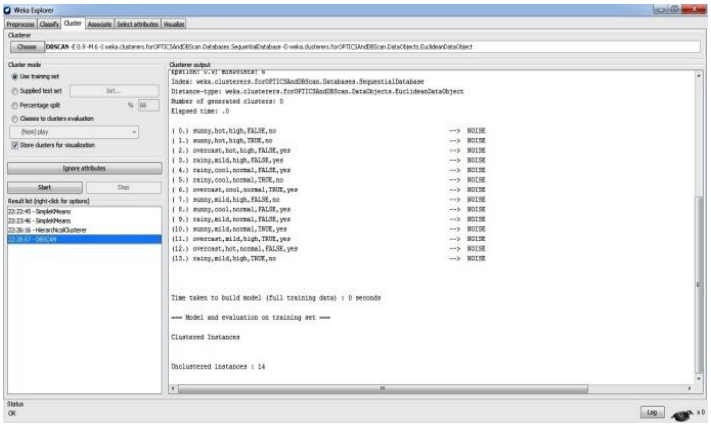

Fig 8: Result of Density based Clustering.

Advantages

1) DBSCAN does not require the knowledge of the number of clusters in the data a priori, as opposed to k-means.

2) DBSCAN can find arbitrarily shaped clusters.

3) DBSCAN has a notion of noise.

Disadvantages

1) DBSCAN can only provide good result in the case of well formed clusters.

2) DBSCAN cannot cluster data sets well with large differences in densities.

4. RESULT AND CONCLUSION

Weka is an open source platform-independent data mining tool. This paper provides a detailed introduction to weka clustering algorithms (K-means algorithms, Hierarchical clustering, and Density based clustering algorithm).

The advantages and disadvantages of each algorithm are analyzed in detail. The pros and cons of each algorithm are identified.

The following conclusions can be observed:

1) K-means clustering algorithm is the simplest algorithm.

2) All the algorithms have some defect in certain aspects.

3) Density based clustering algorithm is not suitable for data with high variance in density.

4) K-Means algorithm is more suitable for large dataset.

5) Hierarchical clustering algorithm is more sensitive for noisy data.

5. REFERNCES

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth "From Data Mining to Knowledge Discovery in Databases".

[2] K-Means clustering using Weka Interface- By Sapna Jain, M Afshar Aalam and M. N Doja, Jamia Hamdard University, New Delhi, Proceedings of the 4th National Conference, INDIA Com-2010 Computing for Nation Development, February 25-26, 2010 Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi.

[3] MacQueen J. B, University of California-Los Angeles "Some Methods for classification and Analysis of Multivariate Observations".

[4] Lloyd, S. P. "Least square quantization in PCM". IEEE Transactions on Information Theory 28, 1982,pp. 129–137.

[5] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, second Edition, (2006).

[6] E.B Fawlkes and C.L. Mallows,‖ *A method for comparing two hierarchical clustering"*, Journal of the American Statistical Association, 78:553–584, 1983.

[7] Timonthy C. Havens. "Clustering in relational data and ontologies" July 2010.

[8] Xu R. Survey of clustering algorithms .IEEE Trans. Neural Networks 2005.

[9] BOHM, C., KAILING, K., KRIEGEL, H.-P., AND KR¨OGER, P. 2004. Density connected clustering with local Subspace preferences. In Proceedings of the 4th International Conference on Data Mining (ICDM).