Implementation of Clustering Through Machine Learning Tool

SREE RAM NIMMAGADDA[1], PHANEENDRA KANAKAMEDALA[2] and VIJAY BASHKARREDDY YARAMALA[3]

[1] LAKIREDDY BALIREDDY COLLEGE OF ENGINEERING
Mylavaram, Krishna Dist ,AP, India

[2] LAKIREDDY BALIREDDY COLLEGE OF ENGINEERING
Mylavaram, Krishna Dist ,AP, India

[3] LAKIREDDY BALIREDDY COLLEGE OF ENGINEERING
Mylavaram, Krishna Dist ,AP, India

ABSTRACT

Clustering is the process of gathering or acquiring similar objects into a group known as cluster. All the objects in a cluster or group are similar to each other. The object in one cluster is dissimilar to the object in another cluster. The process of clustering is also known as un supervisory or machine learning. Weka is a popular tool for machine learning which was written in java. The Weka provides a collection of visualization tools and algorithms for data analysis and predictive modeling through a graphical user interface.

Key words: - clustering, Weka, machine learning and data analysis.

1. INTRODUCTION

Today there are mountains of stored data-containing terabytes of data. New mountains are forming daily as the transactions are performing. To deal with this terabytes of data many data mining techniques can be used. Data mining sometimes referred to as Knowledge Discovery in Databases (KDD). “Data Mining” may be defined as the process of searching, and analyzing data in order to find implicit, but potentially useful information. The advantages associated with data mining are (a) the result of analysis is objective (b) the accuracy of data is constant (c) analysis work is done routinely and (d) large quantities of data can be processed rapidly. Machine learning techniques can be used for data mining because of their ability to extract patterns relating to the concepts to be learned. Clustering is one of the Machine learning and data mining techniques which can identify the similar patterns in the data.

Clustering is process of grouping similar type of objects into one group or cluster. Section 2 describes the clustering process in more detail, and section 3 provides discussion of k-means with an example problem, section 4 provides the detailed discussion of Weka, and section 5 describes about the ARFF files and section 6 describes the process of using Weka tool to simulate clustering process.

2. CLUSTERING

Cluster analysis or clustering is the process of grouping the objects into subsets so that the objects in subset are similar in some sense. Clustering is a method of un supervisory learning and a common technique for stastical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. The following diagram represents the clustering process

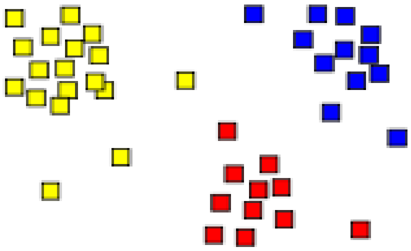


Fig 1: Result of the cluster analysis

The above diagram figure 1 represents that the result produced by the clustering process. The first step in this process is to adopt a mathematical description of similarity such as defining a proximity function. There are number of methods available to measure the similarity between the observations. The most popular distance measure is Euclidian Distance, which is defined as

$$\mathbf{d(i,j)=((x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+.....+(x_{in}-x_{jn})^{\,2})^{1/2}}$$

Euclidian Distance has to satisfy the

a. d(i,j)≥0: Distance is a non negative number
b. d(i,i)=0: The distance of object to itself is 0
c. d(i,j)=d(j,i): Distance is symmetric function
d. d(i,j)≤d(i,h)+d(h,j): going directly from object i to object j in space is no more than making a detour over any object h(triangular inequality).

The typical requirements of clustering are scalability, ability to deal with different types of attributes, Discovery of clusters with arbitrary shapes, Minimum requirements for domain knowledge to determine input parameters, Ability to deal with noisy data, High dimensionality and interpretability and usability. Many clustering algorithms are available. In general, the major clustering methods can be classified into following categories.

Partitioning method: given a database of n objects or data tuples a partioning method constructs k partitions of data, where each partition represents a cluster and≤nk. Each group must contain at least one object and each object must belonging to exactly one group.

Hierarchical methods: A hierarchical method creates a hierarchical decomposition of the given set of data objects. The following figure 2 and figure 3 represents the hierarchical clustering.

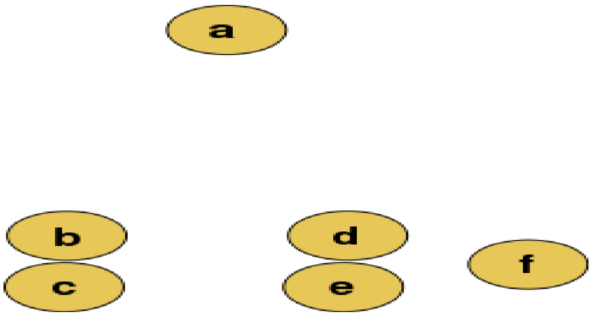


Figure 2: Raw Data

Density based methods: The general idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold; that is for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

Grid based methods: Grid based method quantizes the object space into a finite number of cells that form a grid structure. All of the clustering operations are performed on grid structure.

Model based methods: Model based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model. A model based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points.

The following diagram illustrates the hierarchical clustering.

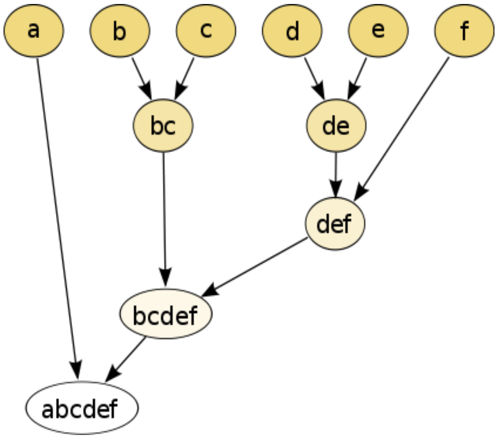


Figure 3: Hierarchical clustering

3. K-MEANS CLUSTERING

K-means is a prototype based clustering technique which performs one level partitions of the data objects. In this we first choose k initial centeroids, where k represents the number of clusters desired. Each point is then assigned to the closest centeroid, and each collection of points assigned to a centeroid is a cluster. The centeroids of each cluster is then updated based on the points assigned to the cluster. This assignment and update steps will continue until no point changes in cluster, equivalently, or centeroids remain the same.

The basic steps of k-means clustering are

1. Determine the centeroid coordinates.
2. Determine the distance of each object to the centeroids.
3. Group the objects based on minimum distance.

4. Update the centeroids.

The following flowchart represents the complete process of k-means

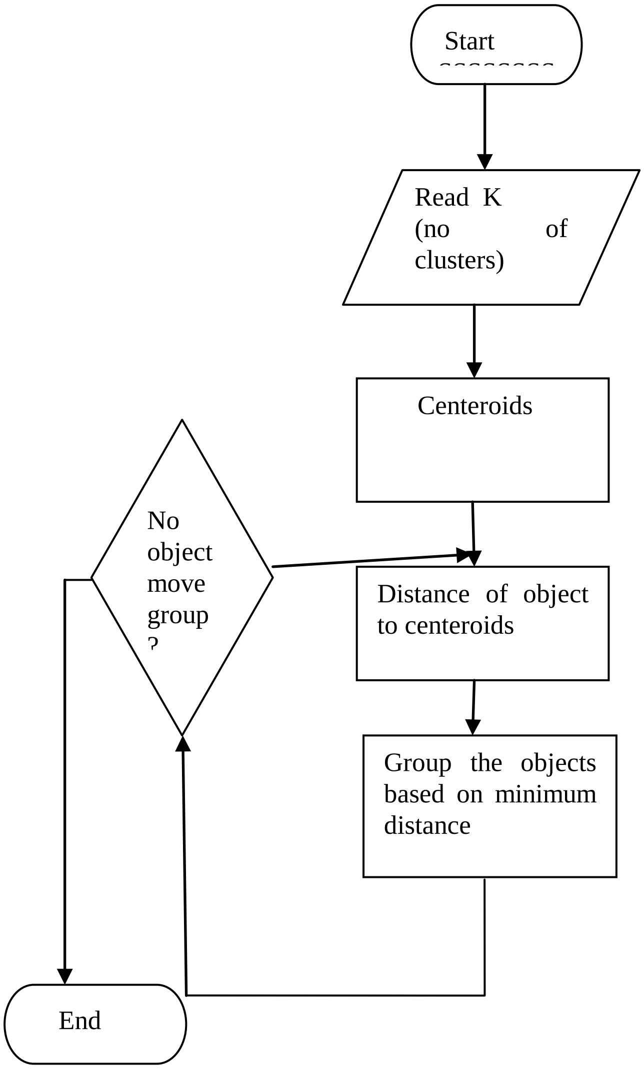

Start
Read K (no of clusters)
Centeroids
Distance of object to centeroids
Group the objects based on minimum distance
No object move group ?
End

Figure 4: flowchart representing K-means

Example:
Suppose we have several objects with two attributes as shown in the following table. Our goal is group these objects in to two clusters i.e. k=2

Object	Attribute 1	Attribute 2
A	2	5
B	4	6
C	5	9
D	9	12
E	11	14

Table 1: objects with two attributes

Each object represents one point with two attributes(x,y) that we can represent it as coordinate in an attribute space as shown in the following figure

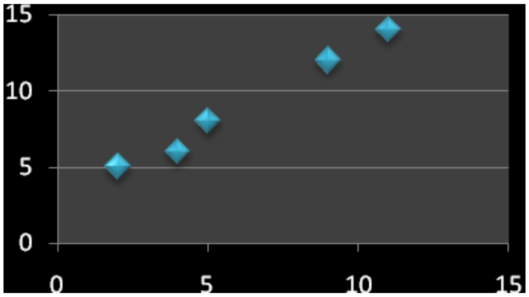


Figure 5: Attribute space of objects

Initial values of the centeroids: suppose we use object A and object B as the first centeroids that indicates k=2. Let c1 and c2 denote the coordinates of the centeroids, then c1=(2,5) and c2=(4,6)

Iteration 0:
Now calculate Euclidian Distance from c1 to all points and c2 to all points and construct a distance matrix D and group matrix G. Euclidian Distance between two objects can calculate as follows

$$ED=((x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+.....+(x_{in}-x_{jn})^{2})^{1/2}$$

Euclidian distance between two objects represents the geometric distance between two objects
ED between (2, 5) and (4, 6) =√ $(2-4)^2 + (5-6)^2$ =2.236

D^0=

0	2.236	5	9.9	12.73
2.236	0	3.16	7.48	10.63

Now depending on the above distance table we need to construct group tale by considering the minimum distance between centeroid and object

G^0=

1	0	0	0	0
0	1	1	1	1

The above table represents that there two cluster instances and object A in one cluster and objects B, C, D, and E in another cluster.

Iteration 1:

Now group1 has only one object so that the centeroid remains same i.e. (2,5)

Group 2 has 4 objects so that we need to recomputed the centeroid as

((4+5+9+11)/4,(6+9+12+14)/4)=7.25, 10.25

The new centeroid for group2 is (7.5, 10.25)

Now compute the distance from new centeroid and all other objects and construct distance table as

D^1=

0	2.236	5	9.9	12.73
7.42	5.35	2.57	2.47	5.3

Now group table is

G^1=

1	1	0	0	0
0	0	1	1	1

The above table represents that objects A, B belong to one cluster and objects C, D, E belong to another cluster.

Iteration 2:

Group 1 has two elements so that there is need to compute centeroid as ((2+4)/2, (5+6)/2) i.e. (3, 5.5)

Group 2 has three elements so that there is need to compute centeroid as ((5+9+11)/3,(9+12+14)/3) i.e. (8.33, 11.67)

Now construct distance table

D^2=

1.12	1.12	4.03	8.85	11.67
9.2	7.2	4.3	0.75	3.54

Now group table is

G^2=

1	1	1	0	0
0	0	0	1	1

The above table represents that the objects A, B, C belong to first cluster and objects D, E belong to second cluster

Iteration 3:

The first cluster is having three objects as its members and second cluster is having two objects as its members. So that there is need to compute the centeroids of two clusters.

Centeroid of first cluster is ((2+4+5)/3, (5+6+9)/3)=(3.67, 6.67)

Centeroid of second cluster is ((9+11)/2, (12+14)/2)=(10,13)

Now construct distance table

D^3=

2.36	0.74	2.68	7.537	10.36
11.31	9.21	6.40	1.41	1.41

Now group table is

G^3=

1	1	1	0	0
0	0	0	1	1

The above table represents that the objects A, B, C belong to first cluster and the objects D, E belong to second cluster.

We obtain result that G^2=G^3. Comparing the grouping of last iteration and this iteration reveals that the objects do not move group any more. Thus the computation of the k-means clustering has reached its stability and no more iteration is required. We get the final grouping as results as

Object	Attribute X	Attribute Y	Cluster or group
A	2	5	1
B	4	6	1
C	5	9	1
D	9	12	2
E	11	14	2

This can be represented with the help of a graph

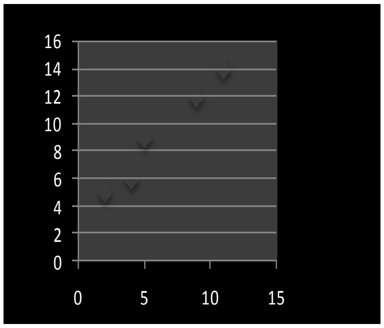
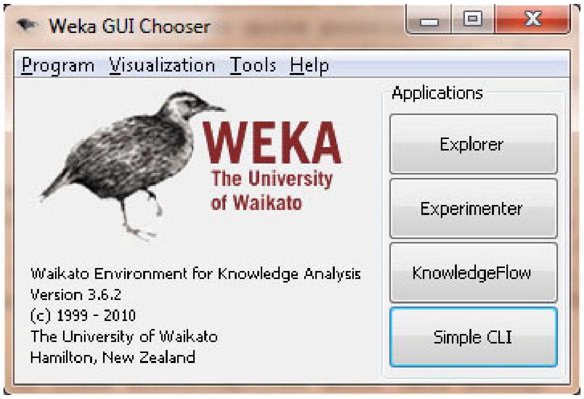


Figure 6: attribute space of clustered objects

4. WEKA

Weka is a popular open source machine learning software package implementing the many state of art machine learning algorithms. It is collection of machine learning algorithms. The algorithms can be either applied directly to data set or called from java code. Weka contains tools for data pre processing, classification, regression, clustering, association rules and visualization. Weka was written in java, developed at the University of Waikato, New Zealand. Weka is a free software available under General Public License(GNU). It provides Graphical User Interface. The fully java based version Weka 3 for which development is started in 1977, is now used in many application areas in particular for education and research.
Weka mainly consists of four interfaces

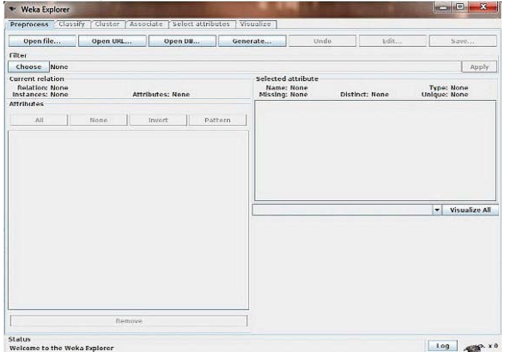
- Explorer
- Experimenter
- Knowledge Flow
- Simple CLI



Waikato Environment for Knowledge Analysis
Version 3.6.2
(c) 1999 - 2010
The University of Waikato
Hamilton, New Zealand

Figure 7 : Weka Interface

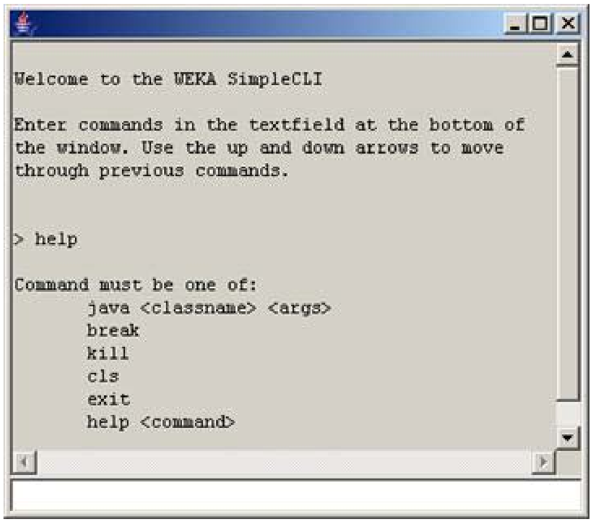
Explorer: it is the main user interface, but the sane functionality can be accessed through the component based knowledge flow interface and from the command line.



Welcome to the Weka Explorer

Figure 8: Explorer window

Experimenter: it allows systematic comparison of the predictive performance of the Weka's machine learning algorithms on a collection of data sets.
Knowledge Flow: In knowledge flow the user select weka components from a tool bar place them on layout canvas, and connect them into a directed graph that process and analyzes the data.
Simple CLI: It provides a command line mode to access a weka. The CLI is a text based interface to the weka environment. All weka commands are similar to the java commands. To execute a command enter it in white box below the window



```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.


> help

Command must be one of:
	java <classname> <args>
	break
	kill
	cls
	exit
	help <command>
```

Figure 9 : Simple CLI window

The machine learning tool weka support several data mining tasks more specially data pre processing, clustering, classification, and regression. Weka can perform data mining tasks directly either on data base by using JDBC or on data sets in ARFF file

5. ARFF FILES

An ARFF (Attribute-Relation File Format) is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine learning project at the department of computer science of the University of Waikato to use with Weka machine learning software. All ARFF files two distinct sections. The first section is the header information followed by the data information. The relation and attributes are declared in the header section. The relation name is defined as the first line in the ARFF file as follows
@relation <relation-name>
Relation-name is a string.

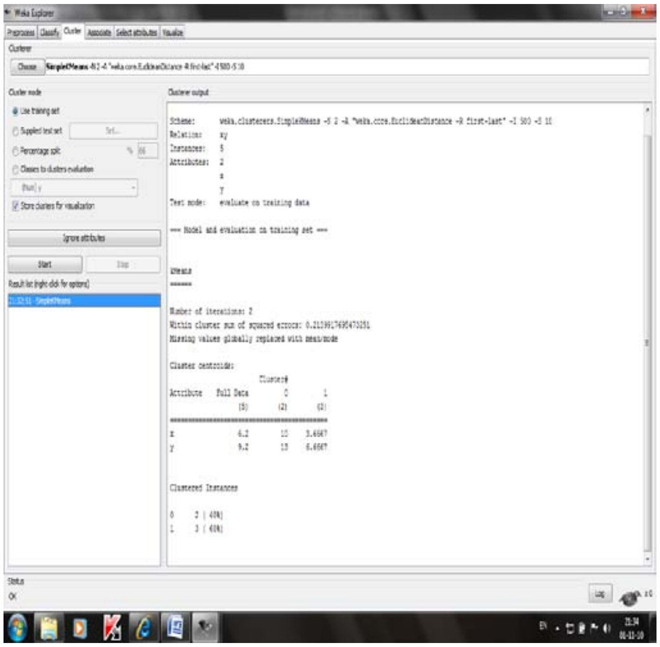
The attribute declaration is
@attribute <attribute-name> <data type>
The data type can be any one of the four types.

- Numeric
- Nominal
- String
- Date

The comments can be specified with %
Now we define an ARFF file for our example problem in section 3
%this is header section
%it contains relation declaration and attribute declarations.
@relation xy
@attribute x numeric
@attribute y numeric
% data section started.
@data
2,5
4,6
5,9
9,12
11,14
Now save this file with .arff as its extension and name this file as sample.arff.The missing values in data section can specify with commas.

6. K-MEANS CLUSTERING WITH WEKA

This example illustrates the use of k-means with Weka. The sample data set used for this example is taken from sample.arff which contains five instances. Some implementations of K-means only allow numerical values for attributes. In that case, it may be necessary to convert the data set into the standard spreadsheet format and convert categorical attributes to binary. It may also be necessary to normalize the values of attributes that are measured on substantially different scales (e.g., "age" and "income"). While Weka provides filters to accomplish all of these preprocessing tasks, they are not necessary for clustering in Weka**.** This is because Weka Simple K Means algorithm automatically handles a mixture of categorical and numerical attributes. Furthermore, the algorithm automatically normalizes numerical attributes when doing distance computations. The Weka Simple K Means algorithm uses Euclidean distance measure to compute distances between instances and clusters.
First go to the Explorer interface. The explorer interface contains the various panels such as pre process, classify, cluster, associate, select attribute, and visualization. Now go to the preprocess panel. Now click on open file and browse for ARFF file which contains objects. After loading the ARFF file we can see elements in object space by using visualization. Visualization panel represents instances in ARFF file in object space by taking their attributes. Now click on cluster and click on choose and select simple k-means and start. The result is as follows



SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) y
☑ Store clusters for visualization

21:32:51 - SimpleKMeans

```
Scheme:       weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     xy
Instances:    5
Attributes:   2
              x
              y
Test mode:    evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 0.2139917695473251
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
                   (5)        (2)        (3)
============================================
x                  6.2         10     3.6667
y                  9.2         13     6.6667



Clustered Instances

0      2 ( 40%)
1      3 ( 60%)
```

Figure 9: output screen

The output is two clusters one with two objects and another with three objects with sum of squared errors 0.21399.

7. CONCLUSION

Weka is an efficient and flexible machine learning tool. By using this we can implement all standard data mining algorithms. Here we are having one drawback for simple k-means that is we can not specify the k value i.e. number of clusters, so that for any number of objects Weka generates only two clusters. If we compare the result of weka clustering with manual procedure as we saw in section 3 we may not ensure the accuracy of the result. Thus if there is an ensure for accuracy of results and make provision to choose number of clusters it can be more and more efficient

8. REFERENCES

1. Beibei Zou, Xuesong Ma, Gen Newton, Donia precup Data Mining using relational database Management systems Supported by NSERC, CFI, NRC
2. Tapas Kanungo, senior member IEEE, David M Mount memer IEEE "An Efficient K-Means Clustering Algorithm: Analysis and implementation." IEEE Transactions on Pattern Analysis and Machine Intelligence vol 24 No 7, July 2009.

3. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Pearson, Introduction to Data Mining.
4. Dmitry Lizokin, Pavel Velikhov, Maxim Grine, Denis Turdakov "Accuracy estimate and optimization techniques for SimRank Computation" Spinger-Verlag 2009.
5. Remi Lehn, Viviane Lambert, Marie-Pierre Nachouki "Data Warehousing tool's architecture: From multidimensional analysis to data mining" IEEE
6. Arun K Pujari "Data Mining Techniques" Universities pres.
7. Jiawei Han and Micheline Kamber "Data Mining Concepts and Techniques" Elesevier.