K-Means Clustering in Spatial Data Mining using Weka Interface

Ritu Sharma(Sachdeva)
M.Tech Student
Department of Computer Science Jamia Hamdard University
New Delhi-110062,India

M. Afshar Alam
Professor,Head
Department of Computer Science
Jamia Hamdard University
New Delhi-110062,India

Anita Rani
Lecturer
DAV Centenary College
Faridabad-121002, India

ABSTRACT
Clustering techniques have a wide use and importance nowadays and this importance tends to increase as the amount of data grows. K-means is a simple technique for clustering analysis. Its aim is to find the best division of n entities into k groups (called clusters), so that total distance between the group's members and corresponding centroid, irrespective of the group is minimized. Each entity belongs to the cluster with the nearest mean. It results into a partitioning of the data space into Voronoi Cells. This paper is about the implementation of k-means clustering using crop yield records by Weka Interface. The data has been taken from the website "Agricultural Statistics of India". This papers also includes detailed result analysis of rice data after demonstration via Weka Interface.

Keywords
K-means Clustering, Euclidean Distance, Spatial data mining, Weka Interface.

1.INTRODUCTION
k-means clustering is a partitioning based clustering technique of classifying/grouping items into k groups (where k is user specified number of clusters). The grouping is done by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid. A centroid (also called mean vector) is "the center of mass of a geometric object of uniform density".

Although K-means is simple and can be used for a wide variety of data types, it is quite sensitive to initial positions of cluster centers. There are two simple approaches to cluster center initialization i.e. either to select the initial values randomly, or to choose the first k samples of the data points. As an alternative, different sets of initial values are chosen (out of the data points) and the set, which is closest to optimal, is chosen. However, testing different initial sets is considered impracticable criteria, especially for large number of clusters, Ismail et al (1989). Therefore, different methods have been proposed in literature by Pena et al. (1999). Also, the computational complexity of original K-means algorithm is very high, especially for large data sets.

Computer science has been widely adopted in different fields like agriculture. One reason is that an enormous amount of data has to be gathered and analyzed which is very hard or even impossible without making use of computer systems. . The research of spatial data is in its infancy stage and there is a need for an accurate method for rule mining. Association rule mining searches for interesting relationships among items in a given data set. This paper enables us to extract pattern from spatial database using k-means algorithm which refers to patterns not explicitly stored in spatial databases. Since spatial association mining needs to evaluate multiple spatial relationships among a large number of spatial objects, the process could be quite costly.

1.1 Definition of K-Means Clustering
This algorithm randomly selects K number of objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the distance between the object and cluster mean. Then it computes new mean for each cluster. This process iterates until the criterion function converges.

1.2 Characteristics of K-Means
- This algorithm attempts to determine K partitions that minimize the squared error functions
- It is relatively scalable and efficient in processing large data sets because the computational complexity of the algorithm is O(nkt) where

  N=total number of objects

  K=number of clusters

  T=number of iterations

- This method often terminates at the local optimum.

It is well-known centroid based technique that takes the parameter, k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is but intercluster similarity is low

Clustering similarity is measured in regard to the mean value of the objects in cluster which can be viewed cluster's centroid or center of gravity.

1.3 K-Means Clustering - Algorithms
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 i.e. calculate the mean value of the objects for each cluster.
5. Until no change

1.4 Diagrammatic Representation

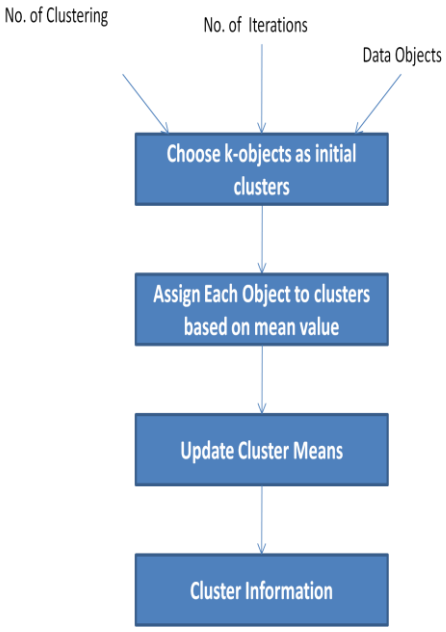

Fig 1: Step-by-Step K-means clustering approach

1.5 Formula to compute Euclidean distance

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

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

2. Weka

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License

2.1 Weka Interface

Fig 2: Weka Interface

Weka Interface has four components:

- *Simple CLI :* Simple CLI provides a commandline interface to weka's routines

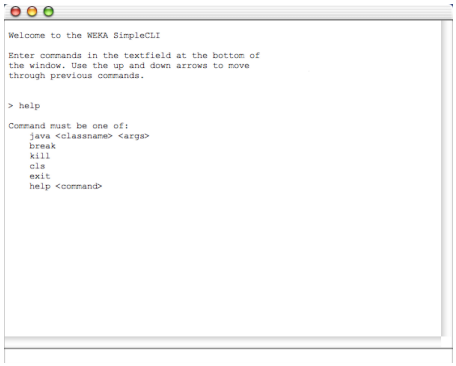

Fig 3: Simple CLI Component

- *Experimenter:* Experimenter allows you to build classification experiments

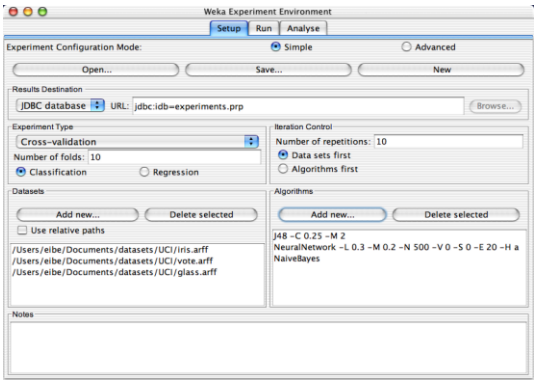

Fig 4 : Experimenter Component

Explorer: Explorer interface provides a graphical front end to weka's routines and components

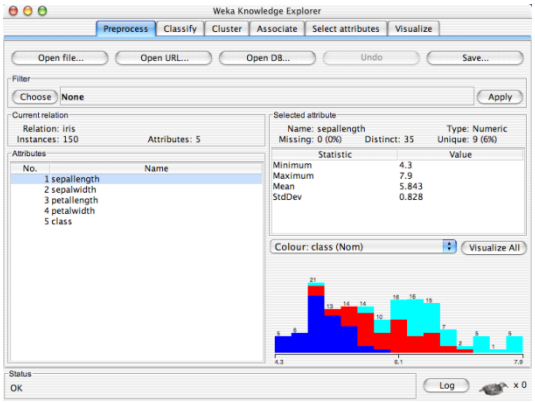

Fig 5 : Explorer Component

Knowledge Flow: Knowledge Flow provides an alternative to the Explorer as a graphical front end to Weka's core algorithms

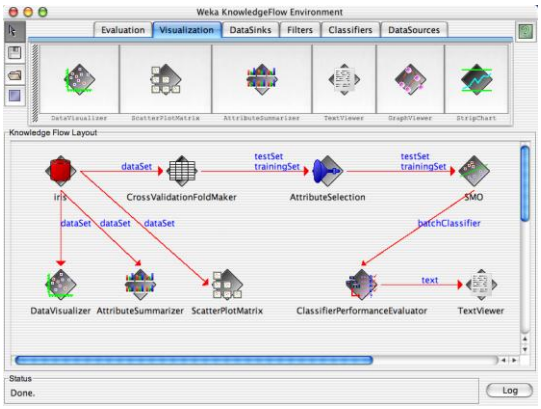

Fig 6 : Knowledge Flow Component

2.2 ARFF File

Attribute Relationship File Format (ARFF) is the text format file used by Weka to store data in a database.

3. DATA DESCRIPTION

This example illustrates the use of *k-means* clustering with WEKA The sample data set used for this example is based on the "Rice data" available in comma-separated format rice-data.csv. The statistics of rice crop has been taken from Agricultural Statistics of India having website name "agricoop.nic.in/Agristatistics.htm". This paper assumes that appropriate data preprocessing has been performed. The resulting data file is "rice-kmeans.arff". K-means algorithm will cluster the production in this rice data set. By having a close look at the production, we can found out that rate of production of rice crop in consecutive years. Also, we can found out the reasons of high and low level of production.

4. K-MEANS CLUSTERING IN WEKA

INTERFACE

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. WEKA provides filters to accomplish all preprocessing tasks. But these are **not necessary for clustering in WEKA** . This is because WEKA SimpleKMeans algorithm automatically handles a mixture of categorical and numerical attributes.

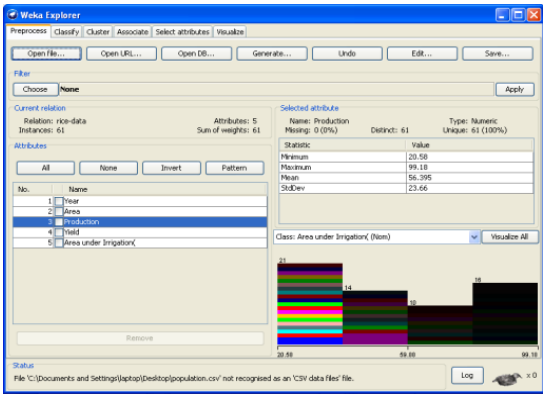

Fig 7: shows the main WEKA Explorer interface with the data file rice-data.csv loaded

Furthermore, the algorithm automatically normalizes numerical attributes when doing distance computations. The WEKA SimpleKMeans algorithm uses Euclidean distance measure to compute distances between instances and clusters.
To perform clustering, select the "Cluster" tab in the Explorer and click on the "Choose" button. This results in a drop down list of available clustering algorithms. In this case we select "SimpleKMeans". Next, click on the text box to the right of the "Choose" button to get the pop-up window shown in Fig 8 , for editing the clustering parameter.

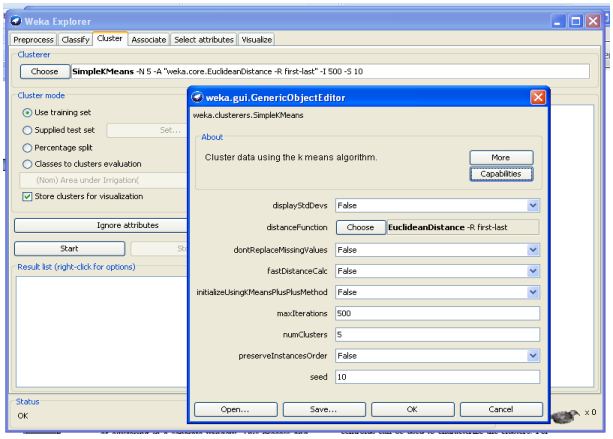

Fig 8: Selecting Clustering Parameters

In the pop-up window we enter **5** as the number of clusters (instead of the default values of 2) and we leave the value of "seed" as is. The seed value is used in generating a random number which is, in turn, used for making the initial assignment of instances to clusters. Note that, in general, K-means is quite sensitive to how clusters are initially assigned. Thus, it is often necessary to try different values and evaluate the results.
Once the options have been specified, we can run the clustering algorithm. We should make sure that in the "Cluster Mode" panel, the "Use training set" option is selected, then click on "Start". To view the results of clustering in a separate window, just right click the result set in the "Result list" panel.

This process and the resulting window are shown in Figures 9 and Figure10.

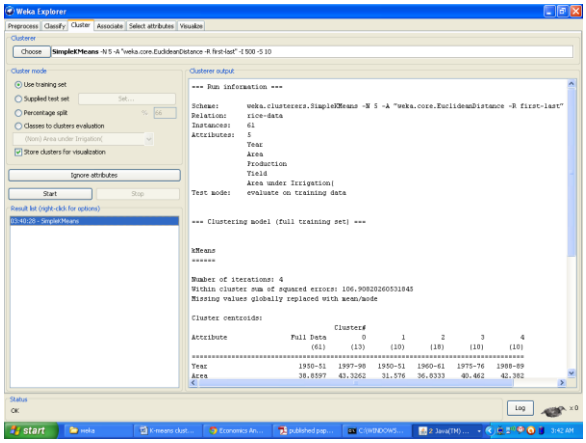

Fig 9 : Progressive stage of k-Means Clustering

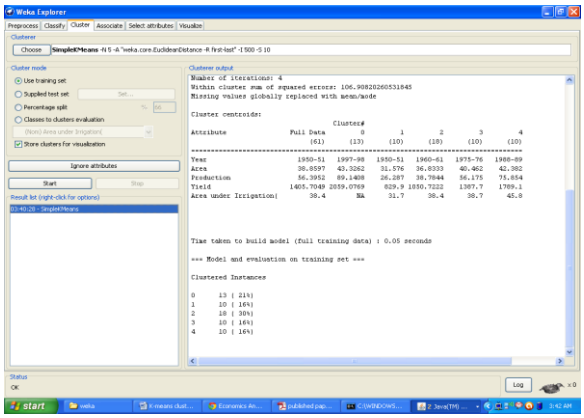

Fig 10 : Resultant Data Output

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. To understand the characteristics of each cluster is through visualization. We can do this by right-clicking the result set on the left "Result list" panel and selecting "Visualize cluster assignments". This pops up the visualization window as shown in Figure 11

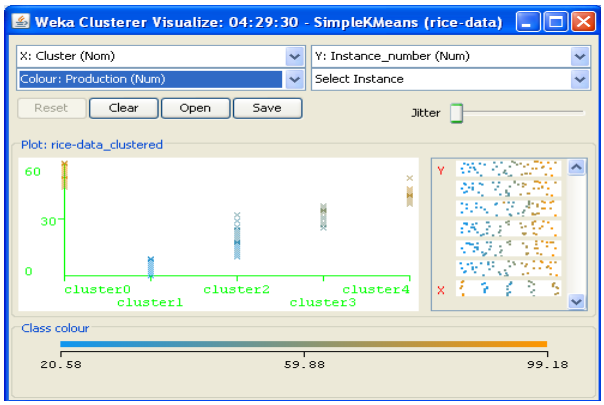

Fig 11: Visualization Clusters

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 color). Different combinations of choices will result in a visual rendering of different relationships within each cluster. In the above example, we have chosen the cluster number as the x-axis, the instance number (assigned by WEKA) as the y-axis, and the "Production" attribute as the color dimension. This will result in a visualization of the level of production in each cluster. For instance, you can note that clusters 1 and 2 are dominated by low level of production, while clusters 0 and 4 are dominated by high level of production. In this case, by changing the color dimension to other attributes, we can see their distribution within each of the clusters. Finally, we have the facility in saving the resulting data set which included each instance along with its assigned cluster. To do so, we click the "Save" button in the visualization window and save the result as the file "rice-kmeans.arff ". When the file is opened in any text-editor, following results/output will be displayed.

@relation rice-data_clustered

@attribute Instance_number numeric
@attribute 'Year ' {1950-51,1951-52,1952-53,1953-54,1954-55,1955-56,1956-57,1957-58,1958-59,1959-60,1960-61,1961-62,1962-63,1963-64,1964-65,1965-66,1966-67,1967-68,1968-69,1969-70,1970-71,1971-72,1972-73,1973-74,1974-75,1975-76,1976-77,1977-78,1978-79,1979-80,1980-81,1981-82,1982-83,1983-84,1984-85,1985-86,1986-87,1987-88,1988-89,1989-90,1990-91,1991-92,1992-93,1993-94,1994-95,1995-96,1996-97,1997-98,1998-99,1999-00,2000-01,2001-02,2002-03,2003-04,2004-05,2005-06,2006-07,2007-08,2008-09,2009-10*,2010-11**}
@attribute Area numeric
@attribute Production numeric
@attribute Yield numeric
@attribute 'Area under Irrigation(' {31.7,32.3,33.6,34.4,34.9,35.4,36.4,36.3,35.8,36.8,37.5,37.4,37.1,37.3,36.5,37.9,38.6,38.4,38.2,37.2,39.1,38.8,38.7,40.2,41.6,42.8,40.7,41.5,42.0,42.7,43.7,42.9,44.1,43.6,45.8,46.1,45.5,47.3,48.0,48.6,49.8,49.9,51.0,50.8,52.3,53.9,53.6,53.2,50.2,52.6,54.7,56.0,56.7,56.9,NA}
@attribute Cluster {cluster0,cluster1,cluster2,cluster3,cluster4}

@data
0,1950-51,30.81,20.58,668,31.7,cluster1
1,1951-52,29.83,21.3,714,31.7,cluster1
2,1952-53,29.97,22.9,764,32.3,cluster1
3,1953-54,31.29,28.21,902,33.6,cluster1
4,1954-55,30.77,25.22,820,34.4,cluster1
5,1955-56,31.52,27.56,874,34.9,cluster1

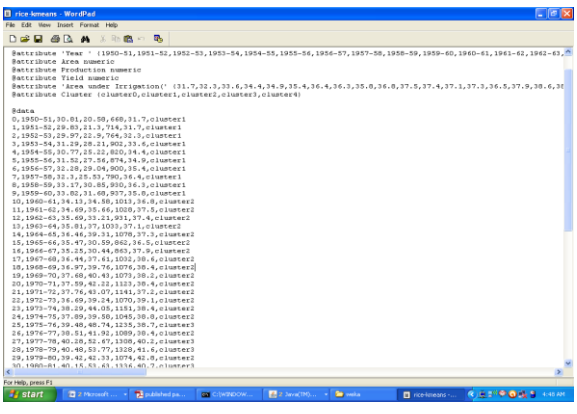

Fig 12: File rice-kmeans.arff

Note that in addition to the "instance_number" attribute, WEKA has also added "Cluster" attribute to the original data set. In the data portion, each instance now has its assigned cluster as the last attribute value. By doing some simple manipulation to this data set, we can easily convert it to a more usable form for additional analysis or processing.

5. K-MEANS CLUSTERING WEAKNESSES

- With fewer samples of data, initial grouping will determine the cluster significantly.
- The number of clusters, k, must be determined before hand.
- With fewer samples of data, inaccurate clustering can occur.
- We never know which variable contributes more to the clustering process since we assume that each has the same weight.
- The accuracy of mathematical averaging weakens because of outliers, which may pull the centroid away from its true position.
- The results are clusters with circular or spherical shapes because of the use of distance.

6. POSSIBLE SOLUTIONS TO THE WEAKNESSES OF K-MEANS CLUSTERING

- Include as many samples of data as possible (the more data, the more accurate the results).
- To avoid distortions caused by excessive outliers, the median can be used instead of the mode.

7. ACKNOWLEDGEMENT

Many thanks and deep gratitude to the researchers who had given their valuable time and contribution in finding and implementing k-means clustering technique. Also I extend my thankx to the weka team for developing such an interface which plays an important role in the field of data mining.

8. REFERENCES

[1] Privacy-Preserving K-Means clustering over vertically Partitioned Data- By Jaideep Vaidya and Chris Clifton, Deptt. Of Computer Sciences, Purdue University, 250 N University St, West Lafayette, IN 47907-2066

[2] K-means clustering via Principal component analysis – By Chris Ding and Xinofeng He, Computational research division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, Preceeding of 21st International conference on Machine Learning,Banff, Canada, 2004.

[3] K-means clustering Tutorial- By Kardi Teknomo, Ph.D

[4] Application of spatial data mining for Agriculture- By D.Rajesh, AP-SITE, VIT University, Vellore-14, International Journal of computer applications(0975-8887), Vloume 15-No. 2, February 2011

[5] A hybridized k-means clustering approach for high dimensional dataset- By Rajashree Dash, Debahuti Mishra, Amiya Kumar Rath, Milu Acharya, Orissa, International Journal of Engineering, Science and technology, Vol 2, No. 2, 2010, pp. 59-66

[6] K-means encyclopedia

[7] 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; INDIACom-2010 Computing For Nation Development, February 25 – 26, 2010 Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi

[8] S. Celis and D. R. Musicant. Weka-parallel: machine learning in parallel. Technical report, Carleton College,CS TR, 2002.

[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang and C.-J. Lin. LIBLINEAR: A library for large linear classification. Journal of Machine Learning. Research, 9:1871–1874, 2008

[10] J. E. Gewehr, M. Szugat, and R. Zimmer. BioWeka — extending the weka framework for bioinformatics. Bioinformatics, 23(5):651–653, 2007.

[11] K. Hornik, A. Zeileis, T. Hothorn, and C. Buchta. RWeka: An R Interface to Weka, 2009. R package version 0.3-16.

[12] S. Celis and D. R. Musicant. Weka-parallel: machine learning in parallel. Technical report, Carleton College,CS TR, 2002