Analysis of Simple K-Means with Multiple Dimensions using WEKA

Rupali Patil
Student, ME Computer
Pune University
PCCOE, Pune

Shyam Deshmukh
Student, ME Computer
Pune University
PCCOE, Pune

K Rajeswari
Assistant Professor
Pune University
PCCOE, Pune

ABSTRACT

Clustering techniques have more importance in data mining especially when the data size is very large. It is widely used in the fields including pattern recognition system, machine learning algorithms, analysis of images, information retrieval and bio-informatics. Different clustering algorithms are available such as Expectation Maximization (EM), Cobweb, FarthestFirst, OPTICS, SimpleKMeans etc. SimpleKMeans clustering is a simple clustering algorithm. It partitions n data tuples into k groups such that each entity in the cluster has nearest mean. This paper is about the implementation of the clustering techniques using WEKA interface. This paper includes a detailed analysis of various clustering techniques with the different standard online data sets. Analysis is based on the multiple dimensions which include time to build the model, number of attributes, number of iterations, number of clusters and error rate.

General Terms

Data mining, Clustering, WEKA interface.

Keywords

Data mining, SimpleKMeans Clustering, WEKA.

1. INTRODUCTION

KMeans [1] is the simple and widely used technique of clustering. It is based on partitioning methodology. It partitions n data items into k-groups where k indicates the number of clusters specified by a user. Clusters are formed such that each item in the cluster has minimum distance from the centroid. For calculating distance between an item and the centroid, KMeans algorithm uses the Euclidean distance measurement. Generally KMeans clustering forms the clusters in spacial extend as compared to EM clustering technique. EM clustering technique allows the clusters of different shapes**.** We are widely using the computers in different fields like banking, agriculture, and medical domains. Huge amount of data gets generated. So it is very difficult and tedious to process such huge amount of data manually without the use of computers. The paper focuses on the analysis of the bank data set [2] and the Irrigation census data of water lifts in all villages of country [3] and formation of the clusters using the SimpleKMeans clustering technique.

2. KMEANS CLUSTERING

Suppose a data set, D contains n objects in Euclidian space. Partitioning method distributes the objects in D into k partitions such that objects within a partition are similar to one another but dissimilar to objects in other partitions, where each partition represents a cluster.

2.1 Characteristics of K-Means [4]

1. The quality of the cluster can be measured by the variation within cluster, which is in terms of sum of squared error among all objects in that cluster.

2. The KMeans method is not guaranteed to converge to the global optimum and often terminates at a local optimum. The results may depend on the initial random selection of cluster centers.

3. The time complexity of the KMeans algorithm is O(nkt), where

n is the total number of elements or data points in the dataset D, k is the number of clusters, and t is the number of iterations.

4. Generally, k << n and t << n so the method is relatively scalable and efficient in processing large data sets.

5. To obtain the good results, it is common to run the KMeans algorithm multiple times with the different initial cluster centers.

2.2 K-Means Clustering Algorithm [4]

The KMeans algorithm is used for partitioning, where each 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. Based on mean value of the elements in the cluster, (re)assign each element to the cluster to which the element is the most similar;
4. Update the cluster means, that is, calculate the mean value of the object for each cluster;
5. Until no change;
```

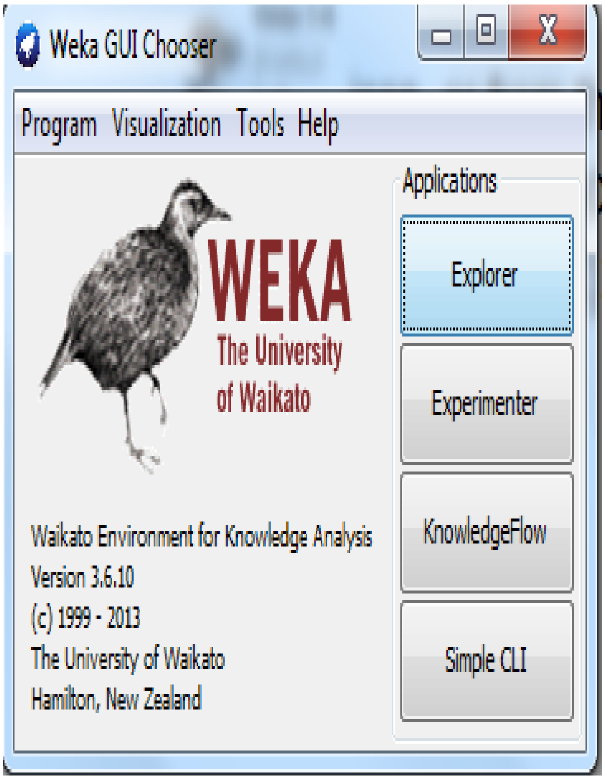
Distance between two elements is calculated using the Euclidian distance measure.

3. WEKA

WEKA [5] is a collection of machine learning algorithms and data preprocessing tools. It provides the extensive support for the whole process of experimental data mining, including the preprocessing of the data for the input, classification, clustering, association rules, evaluating learning schemes statistically and visualizing the input data and the result of learning. The algorithms can either be applied directly to the dataset or called from our own java code. The system is distributed under the terms of the GNU General Public License.

WEKA interface has four main components as shown in Figure 1 below [10, 11].

1. **Simple CLI** provides command line interface and allows the direct execution of WEKA commands.
2. **Explorer** is an environment for exploring the data.
3. **Experimenter** is an environment for conducting experiments and to perform statistical analysis between different learning schemes.
4. **Knowledge Flow** is the Java Beans based interface for setting up and running machine learning experiments.

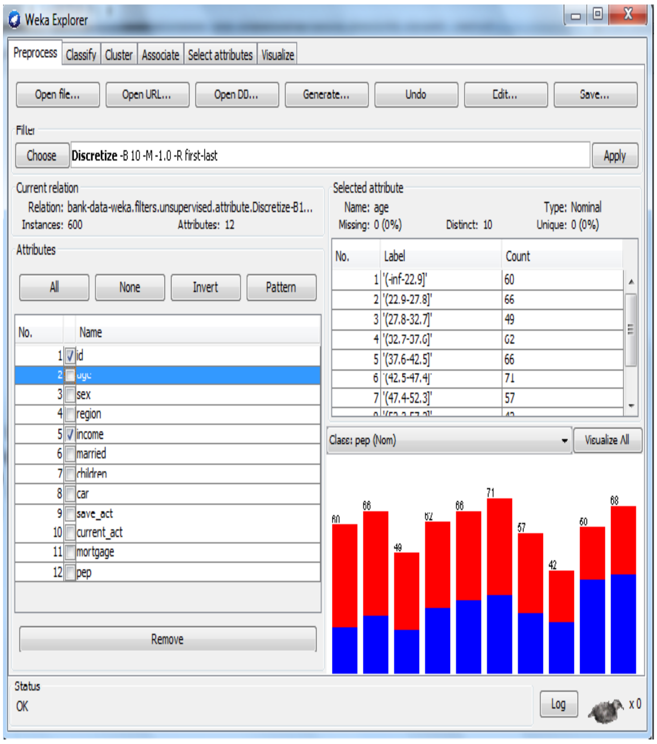


Waikato Environment for Knowledge Analysis
Version 3.6.10
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

Figure 1. WEKA Interface

4. DATA DESCRIPTION

This paper illustrates the use of KMeans clustering algorithm with WEKA. Sample data set used for the analysis is based on the Irrigation census data of water lifts in all villages of country-“srfcliftvilltab5.10.csv” and bank data – “bank-data.csv” in the comma separated format. Irrigation census data set contains attributes such as STATENAME, DISTNAME, BLOCK, VILLAGE, 0-2HP, 2-4HP, 4-6HP, 6-8HP, 8-10HP, 10+HP, unspecified, total. Out of these, 8 are numerical attributes. Bank data set contains attributes as id, age, sex, region, income, married, children, car, save-act, current act, mortgage, pep. Out of these only 3 are numerical attributes. Appropriate preprocessing is performed on the data set and then it is used for the further analysis. In the agricultural (Irrigation related data) field, analyzing the data manually is very difficult and tedious. It becomes very easy to analyze such data using WEKA. Figure 2 below shows bank data set after preprocessing.



Discretize -B 10 -M -1.0 -R first-last

Relation: bank-data-weka.filters.unsupervised.attribute.Discretize-B1...
Instances: 600 Attributes: 12

Name: age Type: Nominal
Missing: 0 (0%) Distinct: 10 Unique: 0 (0%)

No.	Label	Count
1	'(-inf-22.9]'	60
2	'(22.9-27.8]'	66
3	'(27.8-32.7]'	49
4	'(32.7-37.6]'	62
5	'(37.6-42.5]'	66
6	'(42.5-47.4]'	71
7	'(47.4-52.3]'	57

No.		Name
1	☑	id
2	□	age
3	□	sex
4	□	region
5	☑	income
6	□	married
7	□	children
8	□	car
9	□	save_act
10	□	current_act
11	□	mortgage
12	□	pep

Class: pep (Nom)

OK

Figure 2. Data preprocessing

5. EXPERIMENT RESULTS AND COMPARISON

Using WEKA tool we have performed experiments on two data sets containing 12 attributes each. Bank data set contain 600 instances and Irrigation census data set contains 52132 instances. In this paper we have done the analysis based on the parameters such as number of iteration, execution time, squared errors and number of clusters. It was concluded through thorough testing and analysis that these three proved to be the basic measures of our analysis process. The main goal through our proposed work is to identify how efficient simple KMeans work.

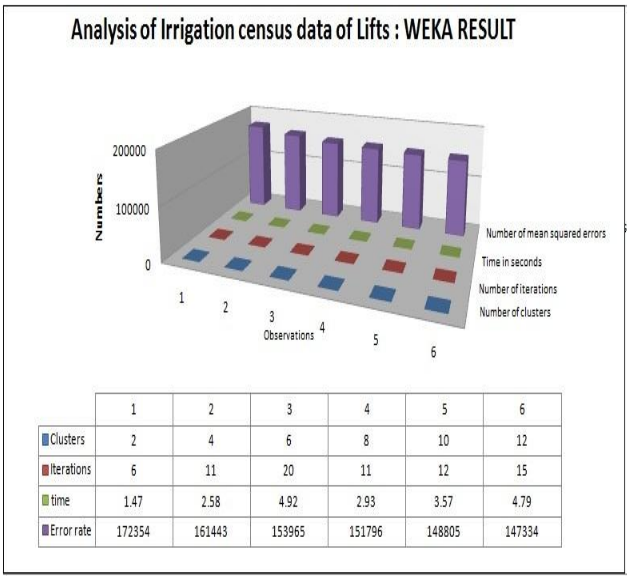
5.1 Experiment No 1:

Figure 3 below shows the graph and table of observations recorded of first experiment for analysis. Experiment was performed with 52132 instances of Irrigation census data of lifts and 12 categorical as well as numerical attributes. Out of which 8 are numerical and 4 are categorical attributes. Total 6 observations were noted by taking number of clusters as 2, 4, 6, 8, 10, and 12. Numbers of iteration, Error rate (mean squared errors) and time to build the model were recorded. From the graph shown in above Figure 3, it is clear that, as number of clusters gets increased, error rate is decreasing. Time required to build the model is dependent on the number of clusters and it is varying in between 1.47 second to 4.92 seconds.

5.2 Experiment No 2:

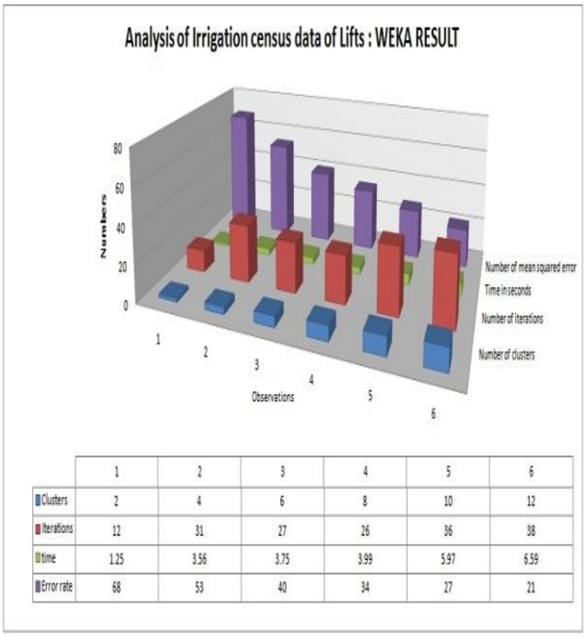
Figure 4 below shows the graph and table of observations recorded of second experiment for analysis.

Experiment was performed with 52132 instances of Irrigation census data of lifts. Experiment 2 is different than Experiment 1 only in terms of number of attributes taken for observations. i.e. only 6 numerical attributes were chosen for Experiment 2. From the graph shown in above Figure 4, it is clear that, as number of clusters gets increased, error rate is decreasing but the number of iterations is getting increased. As compared to Experiment 1, there is significant change in the error rate in Experiment 2. i.e. Error rate has fallen from 172354 to 21 due to the selection of only numerical attributes in Experiment 2.



	1	2	3	4	5	6
🟦 Clusters	2	4	6	8	10	12
🟥 Iterations	6	11	20	11	12	15
🟩 time	1.47	2.58	4.92	2.93	3.57	4.79
🟪 Error rate	172354	161443	153965	151796	148805	147334

Figure 3. Graph and Observations recorded for Analysis of Irrigation census data using WEKA tool for 12 attributes of dataset. (Numerical and Categorical).



	1	2	3	4	5	6
🟦 Clusters	2	4	6	8	10	12
🟥 Iterations	12	31	27	26	36	38
🟩 time	1.25	3.56	3.75	3.99	5.97	6.59
🟪 Error rate	68	53	40	34	27	21

Figure 4. Graph and Observations recorded for Analysis of Irrigation census data using WEKA tool for 6 attributes of dataset (only numerical attributes).

5.3 Experiment No 3:

Figure 5 below shows the graph and table of observations recorded of third experiment for analysis. Experiment was performed with 600 instances of Bank data set, with 9 categorical and 3 numerical attributes. Total 6 observations were noted by taking number of clusters as 2, 4, 6, 8, 10, and 12.

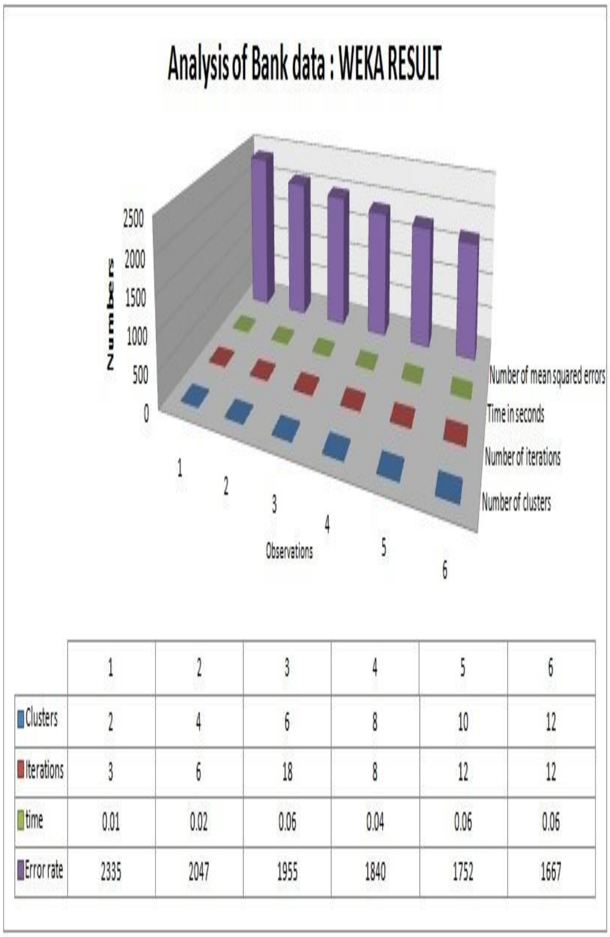
Numbers of iteration, Error rate (mean squared errors) and time to build the model were recorded.

From the graph shown in below Figure 5, it is clear that, as number of clusters gets increased, error rate is decreasing. Time required to build the model is dependent on the number of clusters and it is varying in between 0.01 second to 0.06 seconds.

5.4 Experiment No 4:

Figure 6 below shows the graph and table of observations recorded of forth experiment for analysis.

Experiment was performed with 600 instances of Bank data set, with 3 categorical and 3 numerical attributes. From the graph shown in above Figure 6, it is clear that, as number of clusters gets increased, error rate is decreasing but the number of iterations is varying in between 3 to 13. There is significant change in the error rate in Experiment 4 i.e. Error rate has fallen from 2335 to 311 in Experiment 4. Time required to build the model is comparatively very small in case of the bank data set than in the Irrigation census data set.



	1	2	3	4	5	6
🟦 Clusters	2	4	6	8	10	12
🟥 Iterations	3	6	18	8	12	12
🟩 time	0.01	0.02	0.06	0.04	0.06	0.06
🟪 Error rate	2335	2047	1955	1840	1752	1667

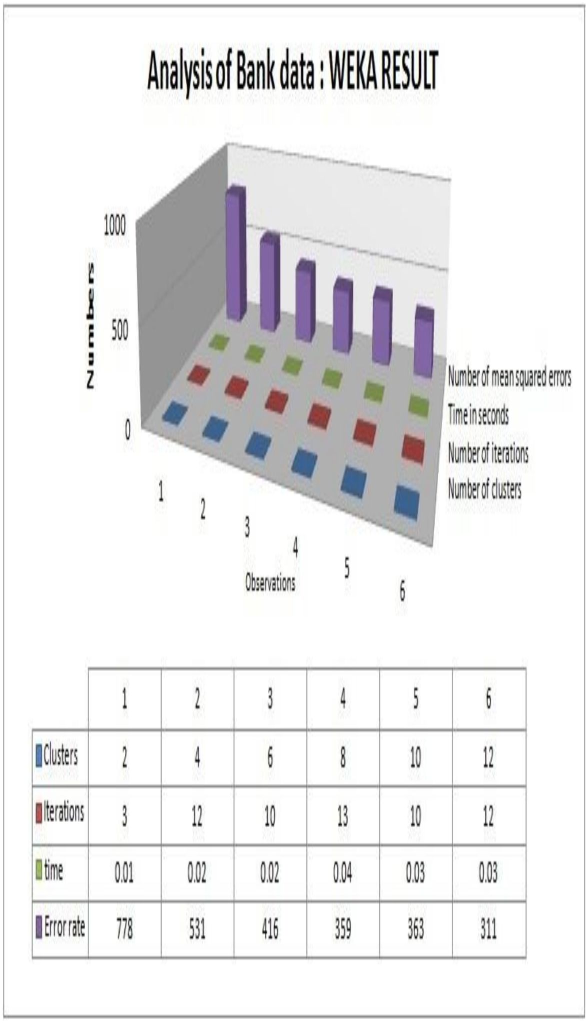
Figure 5. Graph and Observations recorded for Analysis of Bank data using WEKA tool for 12 attributes of dataset. (Numerical and Categorical).

6. CONCLUSION

Experiments 1 to 4 were conducted on the Irrigation census dataset with numerical attributes rather than taking the categorical attributes. From the Experiments 1 to 4, we conclude that, SimpleKMeans clustering is best suited for numerical attributes than categorical attributes to greatly reduce the error rate. Also it is observed that time required for bank dataset is less as compare to the Irrigation census dataset. Hence we conclude that time required to build the model depends on the size of data set. Therefore, SimpleKMeans clustering is efficient only if data set is smaller in size and with numerical attributes.

7. ACKNOWLEGEMENT

Many thanks and deep gratitude to the researchers who had given their valuable contribution in implementing SimpleKMeans clustering technique. Authors want to thank to the Pimpri Chinchawad College of Engineering for providing the required resources. Authors want to extend thanks to the WEKA team for developing such an interface which plays an important role in the data mining. Also thank for making the data set publicly available.



	1	2	3	4	5	6
🟦 Clusters	2	4	6	8	10	12
🟥 Iterations	3	12	10	13	10	12
🟩 time	0.01	0.02	0.02	0.04	0.03	0.03
🟪 Error rate	778	531	416	359	363	311

Figure 6. Graph and Observations recorded for Analysis of Bank data using WEKA tool for 6 attributes of dataset (3 attributes are numerical and 3 are categorical).

8. REFERENCES

[1] 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.

[2] Bamshad Mobasher, School of CTI, DePaul University, Bank data set, http://facweb.cs.depaul.edu/mobasher/classes/ect584/weka/k-means.html.

[3] National Informatics Centre (NIC), Irrigation census data of water lifts in all villages of country, http://data.gov.in/.

[4] K-Means Clustering in Spacial Data Mining using Weka Interface- By Ritu Sharma (Sachdeva), M. Afshar Alam, Anita Rani, Department of Computer Science Jamia Hamdard University New Delhi, International Conference on Advances in Communication and Computing Technologies (ICACACT) 2012, Proceedings published by International Journal of Computer Applications (IJCA).

[5] The University of Waikato. Weka 3 – Machine Learning software in Java, http://www.cs.waikato.ac.nz/ml/weka.

[6] S. Celis and D. R. Musicant, Weka-parallel: Machine Learning in parallel, Technical report, Carleton College, CS TR, 2002.

[7] K-Means clustering Tutorial- By Kardi Teknomo, Ph. D.

[8] Privacy-Preserving K-Means clustering over vertically Partitioned Data-By Jaideep Vaidya and Chris Clifton, Dept. of Computer Sciences, Purdue University, 2050 N University St, West Lafayette, IN 47907-2066.

[9] Application of special data mining for Agriculture- By D. Rajesh, AP-SITE, VIT University, Vellore-14, International Journal of Computer Applications (0975-8887), volume 15 – No.2, February 2011.

[10] The university of WAIKATO, WEKA Manual for version 3-6-8, http://www.nilc.icmc.usp.br/elc-ebralc2012/minicursos/WekaManual-3-6-8.pdf.

[11] Eibe Frank, Mark Hall, Geoffey Holmes, Richard Kirkby, Bernhard Pfahringer, lan H. Witten, Len Trigg, “Weka-A Machine Learning Workbench for Data Mining”, Data mining and Knowledge Discovery Handbook, pp. 1269-1277, 2010.

[12] Analysis of Different Data Mining Tools using Classification, Clustering and Association Rule Mining – By Pritam Patil, Suvarna Thube, Bhakti Ratnaparkhi, K. Rajesweri, International Journal of Computer Applications (0975-8887), volume 93- No. 8, pp. 35-39, May 2014.