www.ijecs.in
*International Journal Of Engineering And Computer Science ISSN:2319-7242*
*Volume 2 Issue 5 May, 2013 Page No. 1568-1571*

# MINIMIZATION OF INTRA-CLUSTERERROR CRITERION FUNCTION USING IMPROVED K-MEANS

* *Jain Sapna, Sharma Anu*

*M.Tech Student
Department of Computer Science and Engineering, College of Engineering, TeerthankerMahaveer University, India)
*E-mail*:rite2sapna@gmail.com
Assistant Professor
Department of Computer Science and Engineering, College of Engineering, TeerthankerMahaveer University, India)
*E-mail*:er.anusharma18@gmail.com

***Abstract*-Data mining is the process of sorting through large database or data warehouse and extracting knowledge interested by the people. The extracted knowledge may be represented as concept, rule, law and model.Clustering is one such technique in data mining that partitions the data in meaningful clusters so that the distances of objects in the same cluster is as small as possible.Among the wide range of clustering algorithms,k-means is one of the most popular clustering algorithms. This paper presents animproved k-means algorithm using Euclidean distance method. The intra-cluster error criterion function minimizes significantly using the improved k-means algorithm. Moreover, the distribution of the data objects also improved and the results are verified over two datasets namely- letter image recognition dataset and the seeds datasets using improved k-means algorithm. The effectiveness of the algorithm is shown by comparing the results over standard k-means algorithm and the improved k-means algorithm.**

***Keywords*- Data mining,Clustering,K-means algorithm, Euclidean distance, Criterion Function.**

## I. INTRODUCTION

Data Mining refers to the extraction of hidden information from a huge databasecalled as a data warehouse.Clustering is one such data mining technique.Clustering is a way that classifies the raw data reasonably and searches the hidden patterns that may exist in datasets [1].It is a process of grouping data objects into disjointed clusters so that the datas in the same cluster are similar, yet datas belonging to different cluster differ.The demand for organizing the sharp increasing datas and learning valuable information from data, which makes clustering techniques are widely applied in many application areas such as artificial intelligence, biology, customer relationship management, data compression, data mining, information retrieval, image processing, machine learning, marketing, medicine, pattern recognition,

psychology, statistics [2] and so on.An aggregation of objects could be referred to as a *cluster*.

Most of the conventional clustering methods assume that patterns having similar locations or constant density create a single cluster. Location or density becomes a characteristic property of a cluster. The properties of clusters have to be specified before clustering is performed. After the raw data is collected, it needs to be screened before it is ready for further analysis. It is desired to select a suitable clustering algorithm such as K-means algorithm to perform cluster analysis, if data has tendency to be clustered. K-means algorithm is one of the numerical, unsupervised, non-deterministic, iterative method[3]

The main advantages of k-means algorithm are its simplicity and speed which allows it to run on large datasets and its disadvantage is that it does not yield the same result with each run.

## 2. Procedure of K-Means Algorithm

1. Randomly distribute all objects to k number of different clusters.
2. For each cluster calculate the mean value and use it to represent the cluster.
3. According to the distance of an object to the cluster centre, re-distribute it to the closest cluster.
4. Calculate the mean value of the objects in each cluster.

5. Calculate the criterion function E, until it converges.
The k-means algorithm criterion function adopts square error criteria. The function of this criterion is to make the generated cluster be as compact and independent as possible .It could be defined as:
value ofcluster Ci [4]. This iterative process continues repeatedly until the criterion function becomes the minimum.

$$d = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2} \quad (2)$$

## 3. Improved K-Means Algorithm

In this section, we define the proposed methodology which is weighted Euclidean distance formula [5] for clustering approach. The improved Weighted Euclidean distance formula is given as:

$$d = ulp\left(6 * \left(var\right)^{-1} \sum_{j=1}^{n} \left(x_j - y_j\right)^2\right) \quad (3)$$

where, var is the variance of the attributes of instances. Variance is calculated using the given formula

$$\text{var} = \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2}{n - 1} \quad (4)$$

where, $\bar{x}$ is the mean of the attributes of instances.
The ulp () method returns the distance from a number to its nearest neighbors. This distance is called an *ULP*for unit of least precision or unit in the last place [6].

## 4. Dataset Description

This paper selects two different data sets downloaded from the UCI [7] repository of machine learning databases to test the efficiency of the improved k-means algorithm and the standard k-means. This
paper uses image recognition and seeds [7] as the test datasets used in experiment evaluation. Table 1[1] shows some characteristics of the datasets and gives a brief description of the datasets

**TABLE I. CHARACTERISTIC OF THE DATASETS**

| Dataset | Number of attributes | Number of Records |
|---|---|---|
| IMAGE RECOGNITION | 17 | 20,000 |
| SEEDS | 7 | 210 |

The dataset was however trimmed for the test purposes. In the image recognition dataset the first attribute was dropped and next five(x-box, y-box, width, high, on-pix) attributes were considered,while for the seeds dataset all attributes were

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2 \quad (1)$$

In which, E is total square error of all the objects in the data cluster, p is given data object, $m_i$ is mean
The Euclidean distance between one vector x=(x1 ,x2 ,…xn) and another vector y=(y1 ,y2 ,…yn ), The
Euclidean distance d(xi, yi) can be obtained as follow:
taken. The above said distance formula is implemented in WEKA using NetBeans IDE.6.0.1

## 5. Result

The analysis of the result findings could be elaborated under the two headings:the intra cluster error criterion function (E) and the cluster distribution of data objects.

*A. Intra Cluster Criterion Function (E)*

With the help of improved k-means, the intra-cluster error criterion function reduced significantly. This sum of squared errors (E) became minimum. This is shown in Table II[2]

**TABLE II. INTRA-CLUSTER ERROR CRITERION FUNCTION (E)**

| Dataset Name | K-Means | Improved K-Means |
|---|---|---|
| IMAGE DATASET | 1215.709 | 6.36 |
| SEEDS | 29.402 | 2.16 |

The graphical representation of the same is given in the figure1[1] given below:

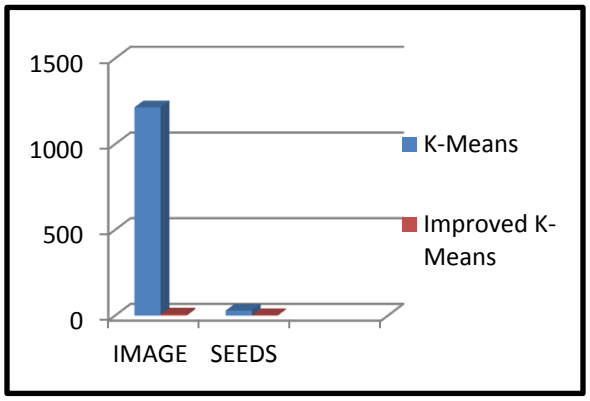


**Figure 1: Intra Cluster Criterion Function (E)**

*B. Cluster Distribution of Objects*
Using the improved k-means algorithm, there was significant improvement in the distribution of objects into the clusters. The more closely the clusters are spaced the efficient is the clustering algorithm.

This is illustrated in the Table III[3] as given below:

**TABLE III .PERCENTAGEDISTRIBUTION OF DATA for S=10**

| | K-Means | | Improved K-Means | |
|---|---|---|---|---|
| | Cluster 0 | Cluster 1 | Cluster 0 | Cluster 1 |
| IMAGE | 42% | 58% | 54% | 46% |
| | Cluster 0 | Cluster 1 | Cluster 0 | Cluster 1 |
| SEEDS | 40% | 60% | 54% | 46% |

As shown below in Figure2[2] and Figure 3[3] the improved k-Means algorithm enhanced the subsequent intra-cluster distribution of the data elements along two clusters for the image recognition dataset and seeds dataset.

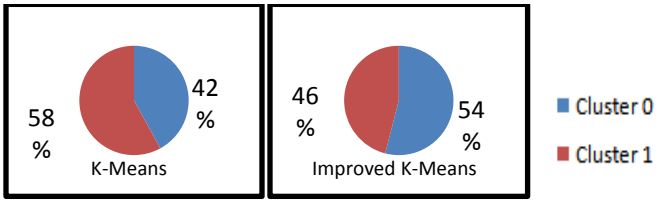


**Figure 2: Image Recognition Dataset**

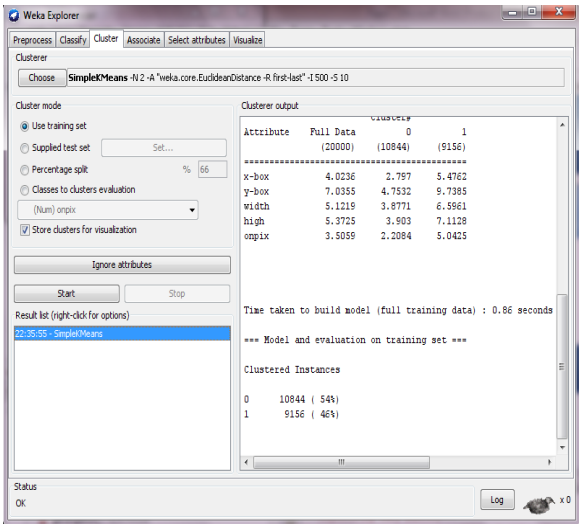


**Figure 4: Snapshot for the Image Recognition Dataset**

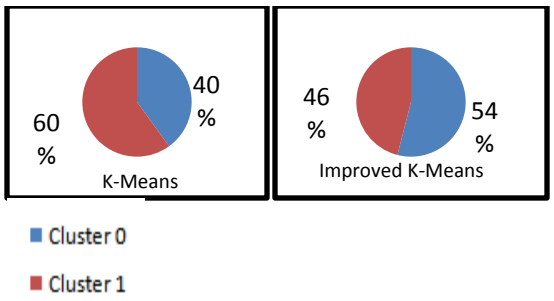


**Figure 3: Seeds Dataset**

The snapshots of the above percentage distribution of data objects using improved k-means algorithm are given as in Figure 4[4] and Figure 5[5] respectively.

## 6. Conclusion

The paper proposed an improved k-means algorithm that uses the Euclidean distance method for clustering. The results were tested and verified over two real time datasets form UCI repository. The results showed a significant reduction in the intra-cluster error criterion function .Also, the percentage distribution of data objects also improved subsequently using the improved k-means algorithm.

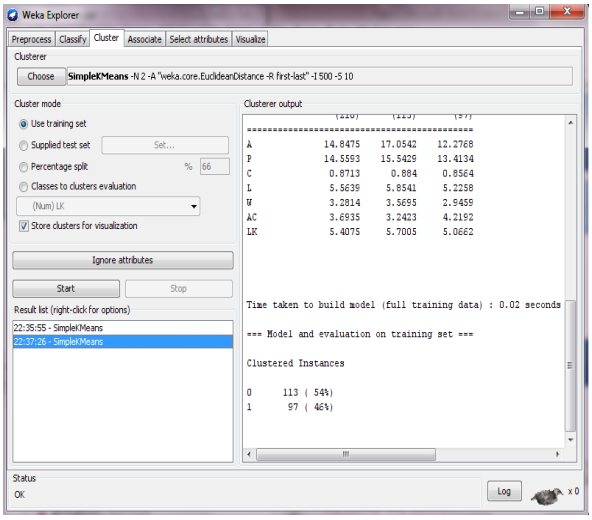


**Figure 5: Snapshot for the Seeds Datas**

## 7. References

[1] Huang Z, "Extensions to the k-means algorithm for clustering large data
sets with categorical values," Data Mining and Knowledge Discovery,
Vol.2, pp:283–304, 1998.
[2] Sun Shibao, Qin Keyun,"Research on Modified k-means Data Cluster

Algorithm"I. S. Jacobs and C. P. Bean, "Fine particles, thin films and
exchange anisotropy," Computer Engineering, vol.33, No.13, pp.200–
201,July 2007
[3]Na Shi, Xumin L., Yong G., "*Research on k-means Clustering Algorithm:An Improved k-means Clustering Algorithm*", Third International Symposium on Intelligent Information Technology and Security Informatics.
[4] Wang J., Su X.; "An Improved K-means Clustering Algorithm", Communication Software and Networks (ICCSN), (2011), IEEE 3rd International conference,2011,China.
[5] ChimY.C. , Kassim A. A., Ibrahim Y., "Character Recognition Using Statistical Moments", Image and Vision Computing 17 (1999) 299–307.
[6]http://www.ibm.com/developerworks/java/library/j-math2/index.html.
[7] archive.ics.uci.edu/ml/datasets.html.