TECHMIND RESEARCH
ISSN: 2291-2118
International Journal of Research in Business and Technology
Volume 5 No.1 August 2014

Improved Clustering Algorithm based on Banking

Vijay Saini[1*], Arko Bagchi[2], Gurpreet Singh[3]
[1]Department of Comp. Sc. & Engg.
Lovely Professional University, Phagwara
ankit.dadwal@gmail.com
[2]Assistant Professor, Department of Comp. Sc. & Engg.
Lovely Professional University, Phagwara
[3]Assistant Professor, Department of Comp. Sc. & Engg,
SBBSIET, Padhiana.

Abstract- *This research can handle the large data set of banking related to customers. It can help us for reduce computation time and increase efficiency of data set related to banking. We can use clustering technique to distribute objects of similar type and extract knowledge from database. We can provide best services to deal with different customers. We checks out which algorithm produce better quality results as compare to previous algorithm. K mean algorithm is used in real applications. The proposed algorithm calculates minimum distance between objects using data set and compares it with previous k mean algorithm. Our algorithm improve accuracy of data set as compare to previous k mean algorithm.*

Keywords- *Data mining; k-mean; self-organizing map; clustering; cluster analysis; Types of data mining algorithms.*

1. INTRODUCTION

Cluster analysis divides data into meaningful or useful groups (clusters). Cluster analysis is process of grouping objects of similar class is called clustering. It contain collection of objects which are same size in one cluster and different in other. There are many tools are used for cluster analysis. We need to partition data in to groups. It is subset of objects of same size and different size. It can help us to reduce number of iterations by grouping them in to smaller set of clusters. It has different quality which depend on their size. Cluster analysis is used to identify earth climate, business to collect large amount of information related to customer. The information and knowledge used for various applications like banking services, classification, prediction, cluster analysis and outlier analysis.

Types of Data mining algorithms: It is collection of various methods which can perform task. Currently lot of data mining techniques used to handle large dataset.

Association rule algorithm: It mainly deals with search statistical relations between objects in dataset. It finds how events aggregate together.

Classification algorithm: It can describe or classify objects related to dataset into predefined set of classes. It is supervised learning approach. It includes objects in dataset used to understand existing objects and predict behaviour of new objects.

Clustering algorithm: It is collection of objects of similar type in one group. The cluster provides us better results.

Inductive and Deductive learning: Machine learning in mainly classify into two different types. In deductive learning, we learn something with existing knowledge and produce some new knowledge from existing knowledge. In inductive learning rules and patterns are extracted from large datasets. In clustering partition the dataset in to subsets for optimization.

Self-organizing map: It can describes all the points in high dimension space in to low dimension space. It can use neural network approach. It is more effective and reliable. The main goal of som is to transform all points from high dimension into low dimension space. In neural networks each neuron assign a weight vector with same dimension of input space. Each output neuron fully connected to all source nodes as input layer. It is impossible to assign network node to input class in advance. It can provides random weight vector initialization.

In order to elaborate the concept a little bit, let us take the example of the library system. In a library books concerning to a large variety of topics are available. They are always kept in form of clusters. The books that have some kind of similarities among them are placed in one k-means clustering.

The k-means algorithm (Lloyd, 1982) belongs to a family of algorithms known as optimization clustering algorithms. In these algorithms, clusters are obtained such that some principle of cluster quality is optimized. The examples are partitioned into clusters such that the clusters are optimal according to some measure. The improved K-means algorithm is used to handle large amount of data set related to banking. We need to optimize the large data set using clustering approach.

The proposed k-means algorithm is as follows:

1. Draw multiple divisions {DI,D2, . . . ,Dj } from the original dataset.
2. Repeat step 3 for n=l to i.
3. Apply combined approach for sub sample.

4. Compute centroids.
5. Choose minimum of minimum distance from cluster center.
6. Now apply new calculation again on dataset D for k1 clusters.
7. Combine two nearest clusters into one cluster and recalculate the new cluster center for the combined cluster until the number of clusters reduces into k.

In the experiments reported here, the initial center vectors that were randomly selected from the dataset and the stopping criterion was based on the movement of the cluster center. when vectors no longer changed clusters between iterations (the clusters had stabilized), the algorithm terminated. The number of clusters are equal to the number of SOM output map neurons.

The disadvantage of k-means compared to SOM is that it is very costly technique. Each SOM is different from other. It is not very reliable.

2. OBJECTIVE

This research improves the traditional algorithms K-Means and SOM appropriately and provides algorithm based on K-means for mining large-scale high dimensional datasets. It evaluate the performance of clustering algorithm. It can analyze the banking data by applying clustering algorithms on it. It find best possible solution for handling large amount of data. It can help us to reduce complexity of banking data set and increase its accuracy. This research increases the efficiency and reduced the computation time of large scale banking data set. Clustering is the effective and efficient data mining technique. With the help of clustering we can distribute similar type of objects and mine the knowledge. K-means algorithm for develop a good clustering is to simply generate all possible partitions of n points into k clusters, evaluate some optimization results.

3. METHDOLOGY

Enhance the efficiency: In this we can increase the efficiency of data set using Improved K-means approach. It provides efficient results for handle large amount of data. Efficiency means the time, efforts and cost of performing task. Our major goal is to increase efficiency of k mean algorithm which helps to customers for handle large dataset.

Analysis between computation time: A problem is treated as comparatively difficult if its solution requires various resources, which ever the algorithm is used. The running time of previous algorithm as compare with proposed algorithm to deal with large data set.

Error rate: Our algorithm try to minimize the rate of errors made by a predictive model. In Improved k means algorithm we must try to reduce error rate to make our algorithm more better. In our data set contain lot of information which satisfy each customer.

Data set of banking.

Checking status	Duration	Credit History	Employment
100	below_1_year	critical/other existing	>=7
200	up_2_years	existing paid	1<=X<4
<200	lo_1_year	critical/other existing	4<=X<7
<300	up_2_years	existing paid	4<=X<7
<0	1_2_years	delayed previously	1<=X<4
no checking	up_2_years	existing paid	1<=X<4
no checking	1_2_years	existing paid	>=7
0<=X<200	up_2_years	existing paid	1<=X<4
no checking	lo_1_year	existing paid	4<=X<7
0<=X<200	up_2_years	critical/other	unemployed
0<=X<200	lo_1_year	existing paid	<1
<0	up_2_years	existing paid	<1
0<=X<200	lo_1_year	existing paid	1<=X<4
<0	1_2_years	critical/other existing	>=7
<0	1_2_years	existing paid	1<=X<4
<0	1_2_years	existing paid	1<=X<4
no checking	1_2_years	critical/other existing	>=7
<0	up_2_years	no credits/all paid	<1
0<=X<200	1_2_years	existing paid	>=7
no checking	1_2_years	existing paid	>=7
no checking	lo_1_year	critical/other existing	1<=X<4
<0	lo_1_year	existing paid	1<=X<4
0<=X<200	1_2_year	critical/other existing	1<=X<4
0<=X<200	1_2_years	existing paid	>=7
<0	1_2_years	existing paid	Unemployed
200	up_2_years	delayed previously	< =1
<0	up_2_years	existing paid	>=7

Weka tool used for analysis:

WEKA is data mining software which is used for

university and research purpose. It proposes number of data mining methods from examining data analysis, statistical learning, machine learning and databases. The data mining tools like mat lab, WEKA, SIPINA etc available.Weka is more powerful tool, it contains some supervised learning and other paradigms include clustering, factorial analysis, and parametric and non parametric selection algorithm. It can include various operations which operations to be performed. Initially it can select the data set and after that it checks the performance of data set.

```
m_ReplaceMissingFilter = new ReplaceMissingValues();
Instances instances = new Instances(data);
instances.setClassIndex(-1);
m_ReplaceMissingFilter.setInputFormat(instances);
instances = Filter.useFilter(instances, m_ReplaceMissingFilter);

m_Min = new double [instances.numAttributes()];
m_Max = new double [instances.numAttributes()];
for (int i = 0; i < instances.numAttributes(); i++) {
  m_Min[i] = m_Max[i] = Double.NaN;
}

m_ClusterCentroids = new Instances(instances, m_NumClusters);
int[] clusterAssignments = new int [instances.numInstances()];

for (int i = 0; i < instances.numInstances(); i++) {
  updateMinMax(instances.instance(i));
}

Random RandomO = new Random(m_Seed);
int instIndex;
HashMap initC = new HashMap();
DecisionTable.hashKey hk = null;
```

Fig 1: Net bean

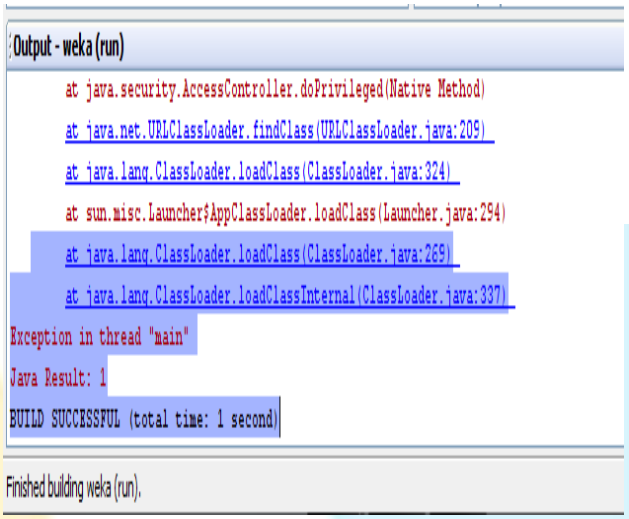

Fig 2: Net beans result of computation time.

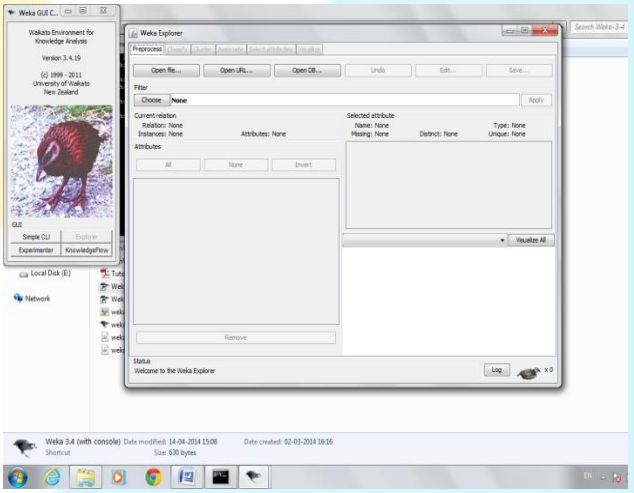

Fig 3: Weka tool graphical user interface.

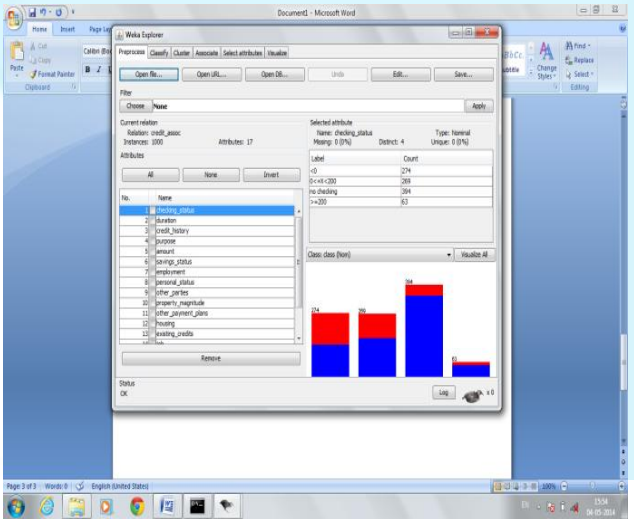

Fig 4: Classify various class attributes

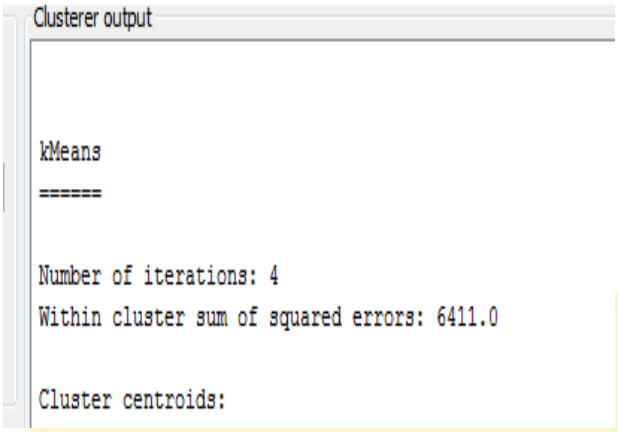

Fig 5: Output results of k means.

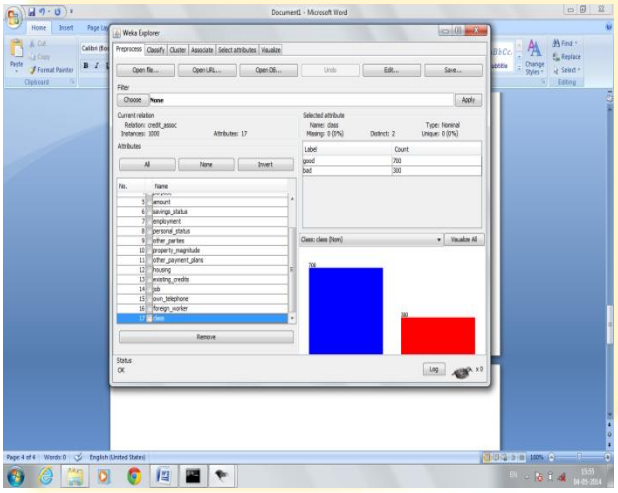

Fig 6: Classify the class attributes.

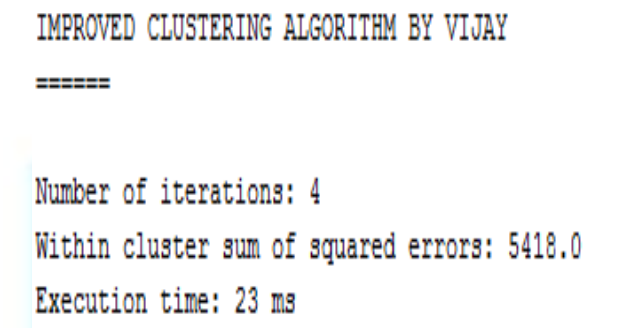

Fig 7: Results of improved k mean algorithm.

	K-Means	Improved Algorithm
No. of Iterations	4	4
Error Rate	64.11	54.18
Computation Time	93 ms	23 ms
No. of Clusters	2	6

Table: 2 Result & Discussion.

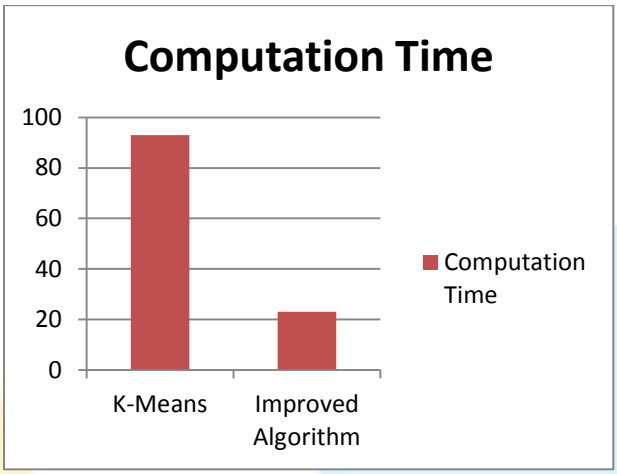

Fig: 8 Comparison & Results of computation time.

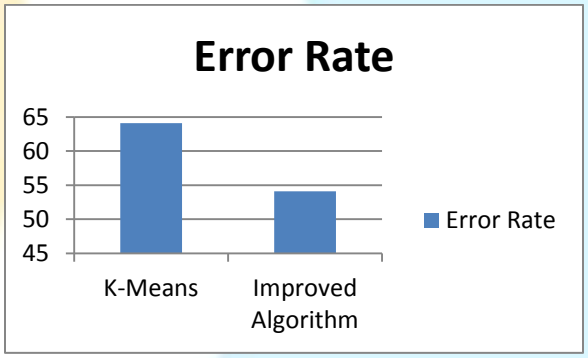

Fig: 9 comparison and Results of error rate.

4. CONCLUSION & FUTURE WORK

This paper is representing improved k-mean algorithm on banking data set. Where, k-mean depends upon initial clusters and initial clusters based on randomly selected centroid. The proposed algorithm is improving the computation time and error rate. After implementation of these algorithms on Banking data set, the results has shown better performance and efficiency with improved algorithm as compare to K-Means. The future work is to reduce the error rate up to minimum value to achieve the maximum performance of the system.

REFERENCES

[1] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, "An Efficient enhanced k-means clustering algorithm," *Journal of Zhejiang University*, 10(7):1626–1633, 2006.

[2] Grigorios F. Tzortzis and Aristidis C. Likas(2009) "Global Kernal K-mean algorithm for clustering in feature space*". IEEE transaction on neural network* volume 20,no.7

[3] Kiri Wagstaff, Claire Cardie(2010) "Constrained K-means Clustering with Background Knowledge" .*Proceedings of eighteenth international conference on machine learning*. Pp. 577-584.

[4] K.A Abdul Nazeer, M.P Singh(2009) "Improving the accuracy and efficiency of k means, kohenon self organizing map and hierarchical agglomerative clustering". *Proceedings of world congress on engineering*. Volume 1, London u.k.

[5] Shi Na ,Liu Xumin and Guan Young(2010) "Research on k means clustering algorithm". *Third international symposium of intelligent information technology and security information*. ISSN: 978_0_7695_4020_7.

[6] Souptik Dutta (2009) "Distributed K-Means Clustering over a Peer-to-Peer Network", *IEEE transactions on information and data engineering*, vol. 21, no. 10 pp 257-345.

[7] Ran Vijay Singh, M.P.S Bhatia(2011) "Data Clustering with Modified k-mean". *IEEE conference on recent trends in Information technology"* ISSN: 978_1_4577_0590_8.

[8] Abdul Nazeer and Sebastian M.P "Improving the accuracy and efficiency of k-means clustering algorithm" *Procedings of world congress on engineering volume* 1,wce 2009.

[9] Vora Pritesh and Oza Bhavesh(2013) "a survey on k-means clustering and particle swarm optimization". IJISME ,ISSSN:2319-6386,Volume 1,Issue 3.

[10] Grigorious F. Tzortzis and Aristidis C.likas."Global k-mean clustering algorithm for clustering in feature space" *IEEE transaction on neural network,* volume 20,no 7,july 2009.

[11] Wagstaff kiri et al (2001) "Constrained k-means clustering with background knowledge" *Proceedings of eighteenth international conference on machine learning*. PP 557-584.

[12] Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, (2):283–304, 1998.

[13] Narendra Sharma ,Aman Bajpai ,Mr.Ratnesh Litoriya*," "* Comparison the various clustering algorithms of weka tools," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, pp.73-80, May 2012.

[14] Zhexu Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", *Data Mining and Knowledge Discovery* 2, 283–304 (1998).