A Synthesized Approach for Comparison and Enhancement of Clustering Algorithms in Data Mining for Improving Feature Quality

Heena Sharma, Navdeep Kaur Kaler

Abstract*-- *K-Means and Kohonen SOM clustering are two major analytical tools for unsupervised forest datasets. However, both have their innate disadvantages. Clustering is currently one of the most crucial techniques for dealing with massive amount of heterogeneous information on the databases, which is beyond human being's capacity to digest. Recent studies have shown that the most commonly used partitioning-based clustering algorithm, the K-means algorithm, is more suitable for large datasets. Also, as clusters grow in size, the actual expression patterns become less relevant. K-means clustering requires a specified number of clusters in advance and chooses initial centroids randomly; in addition, it is sensitive to outliers. SOM We present an improved approach to combined merits of the two and discard disadvantages.

Key-words*-- *Clustering, K-means, Kohonen SOM, Data Mining

I. NTRODUCTION

Data mining is the important step for discover the knowledge in knowledge discovery process in data set. Data mining provide an useful pattern or model to discovering important and useful data from whole database. Different algorithms are used to extract the valuable data. To mine the data important steps or tasks are: Clustering use to describe the data and categories into similar objects in groups. Find the dependencies between variables. Mine the data using tools. Classification is an important task in data mining. Classification is used to classify the data items into the predefined classes and find the model to analysis. Its purpose is to set up a classifier model and map all the samples to a certain class which can provide **[1]** much convenience for people to analyze data further more. Classification belongs to directed learning and the main methods include decision tree, Bayesian classification, neural network, genetic algorithm and rough set etc. Clustering and Classification are two of the mostly used methods of data mining which provide us much more convenience in our research .it is the extraction of hidden descriptive or predictive information from large databases.

A. CLUSTERING

Cluster analysis divides data into meaningful or useful groups (clusters). If meaningful clusters are the goal, then the resulting clusters should capture the "natural" structure of the data. For example, cluster analysis has been used to group related documents for browsing, to find genes and proteins that have similar functionality, and to provide a grouping of spatial locations prone to earthquakes. However, in other cases, cluster analysis is only a useful starting point for other purposes, e.g., data compression or efficiently finding the nearest neighbors of points. Whether for understanding or utility, cluster analysis has long been used in a wide variety of fields: psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining.

Component of Clustering:

Typical clustering activity involves the following steps: (i) Pattern representation. (ii) Definition of a pattern proximity measure appropriate to the data domain. (iii)Clustering or grouping. (iv) Data abstraction. (v) Assessment of output. Figure 1.1 depicts a typical sequencing of the first three of these steps including a feedback path where the grouping process output could affect subsequent feature extraction and similarity computations.

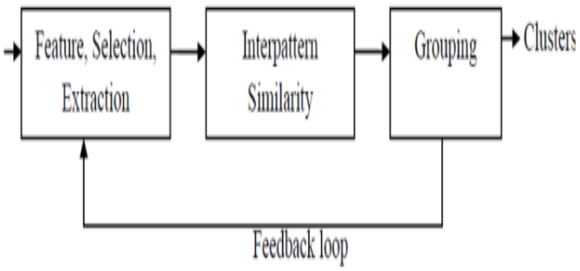

Fig 1.1 components of clustering

B. K-MEANS METHOD

It is a partitioned clustering algorithm. It partitions the given data into k clusters. the no of clusters are fixed .Let the set of data points (or instances) D be

$\{x_1, x_2, \ldots, x_n\}$,

Where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ is a vector in a real-valued space $X \subseteq R^r$, and r is the number of attributes (dimensions) in the dataset. The main objective is to minimize the sum of squared Euclidean distance between objects and cluster centroid.

Steps:

1. $E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$ Select k points at random as the initial centroids.

Manuscript Received on April 2014.

Heena Sharma, Research Scholar, Done B.Tech. (CSE) from L.L.R.I.E.T Moga (P.T.U), now doing M.Tech (CSE) from L.L.R.I.E.T Moga (P.T.U), Punjab, India

Navdeep Kaur Kaler, Assistant Professor in Department Of CSE, L.L.R.I.E.T, Moga, Punjab, India

2. Assign each object to the cluster with the closest centroid.
3. Recalculate the Centroid of each cluster as mean of the objects assigned to it.
4. Repeat steps 2 and 3 until no change.
5. Pass the solution to the next stage.

C. SELF ORGANISATION MAP (SOM)

The SOM means self-organizing maps introduced by Teuvo Kohonen is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map. Self-organizing maps are different from other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space. This makes SOMs useful for visualizing low-dimensional views of high-dimensional data; one of the most interesting aspects of SOMs is that they learn to classify data without supervision.

II. PREVIOUS WORK

Tipawan Silwattananusarn *et al.* (2012) they [1] explores the applications of data mining techniques which have been developed to support knowledge management process from 2007 to 2012 are analyzed and classified. They discussed on the findings is divided into four topics: knowledge resource, knowledge types or knowledge datasets, data mining tasks and data mining techniques and applications used in knowledge management.

Y. Ramamohan *et al.* (2012) explore that [2] data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends.

S.Balaji*et al.* (2012) explore that segmentation [3] of customer utilizes decision tree technique for customer preferences towards products.

Yong Shi *et al* (2011) presented research [4] on selecting proper dimensions for noisy data. They select good dimension candidates for further data analysis based on the observation of the alteration of the big difference of each dimension through the data mining processes.

III. IMPROVED ALGORITHM APPROACH

This research improves the performance of traditional algorithms K-Means and presents a synthesized algorithm SOM and K-means for mining large-scale high dimensional datasets. The mostly used algorithm is K- means which can deal with small convex datasets preferably. But it also exist some shortcomings. For example, it can only deal with numeric data, find convex or spherical sharps be sensitive to the input and noise and can't deal with large datasets.

To increase the degree of association between the members of the same cluster, increase cluster quality, PCA used for add more features and to overcome the limitations such as optimal searching samples, reduce the sensitivity to outliers or noises, outfit problems for classifying samples perfectly.

Objectives are:

1. To combine features of clustering algorithms such as K-means and SOM.
2. To find clusters in large high dimensional spaces efficiently.
3. To improve the results with clusters quality and performance.
4. To reduce the error rate and achieve accuracy.
5. To reduce the computational time of execution.

PROPOSED ALGORITHM

1. Draw multiple sub-samples {S_1, S_2, . . . ,Sj } from the original dataset.

2. Repeat step 3 for m=l to i

3. Apply combined approach for sub sample.

4. Compute the centroid.

5. Choose minimum of minimum distance from cluster center criteria

6. Now apply new calculation again on dataset S 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.

IV. IMPLEMENTATION & RESULTS

A. NETBEANS IDE:

It is an integrated development environment for developing primarily with java but also with other languages. It is also an application platform framework for java desktop applications and others. The NetBeans Platform allows applications to be developed from a set of modular Software components called modules. **NetBeans IDE** supports all Java application types. All the functions of the IDE are provided by modules. Each module provides a well defined function, such as support for the Java language, editing, or support for the CVS versioning system, and SVN. NetBeans contains all the modules needed for Java development in a single download, allowing the user to start working immediately. Modules also allow NetBeans to be extended.

B. WEKA :

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in java..Weka is free software available under the GNU General Public License. The Weka (pronounced Weh-Kuh)work bench. It contains a collection of visualization tools and algorithms for data analysis and predictive modeling. Together with graphical user interfaces for easy access to this functionality. The front-end to (mostly third-party) modeling algorithms implemented in other programming languages, this original version was primarily designed as a tool for analyzing data from agricultural domains.

Table 1

Objectives	K-Means	SOM	Improved clustering algo
Error rate	0.6491	0.5585[1.2076]	221.66/1000=0.22166
Computation time	16 ms(.016 sec)	16ms(.016sec)[0.032sec]	0 sec
No of clusters	3	4	2
Accuracy(corresponding to error rate)	High	high	Higher
Distance Normalization	Variance	variance	Mean

Fig 1.2: Shows Error Rate

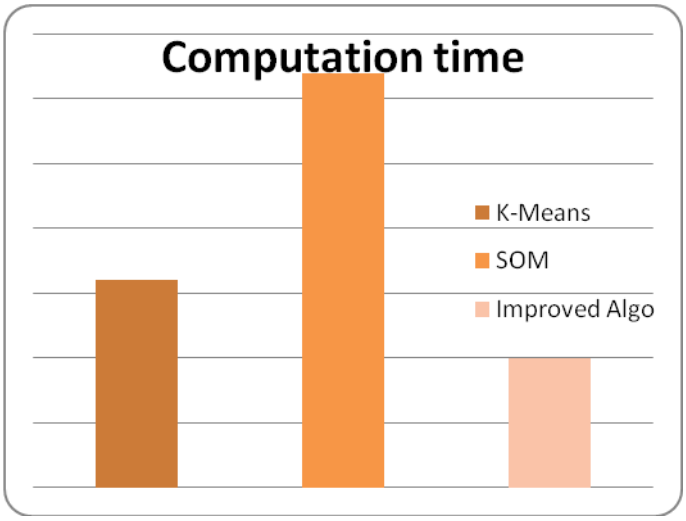

Fig 1.3: Shows Computation Time

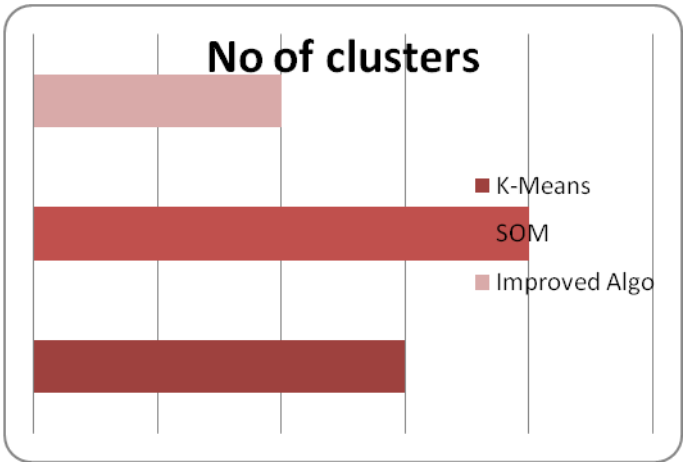

Fig 1.4: Shows No. of clusters

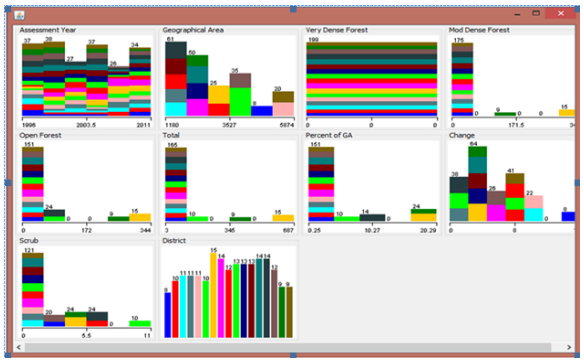

Fig 1.5: Visualize all attributes

```
Enhanced Clustering Algorithm Developed By Heena Sharma
======

Number of iterations used: 2
Within cluster error rate: 221.66503373359728

Cluster centroids:

Cluster 0
        Mean/Mode:  2005.3333 3386       0      343      344      687       20.29     4
        Std Devs:      4.1173    0       0        0        0        0           0     0
Cluster 1
        Mean/Mode:  2001.2222 2113       0      146      244      390       18.46     3
        Std Devs:      4.7376    0       0        0        0        0           0     0
Cluster 2
        Mean/Mode:  2003.4343 2861.8457  0       12.3429  28.8743  41.2171    1.7914   5.72
        Std Devs:      4.5239 1392.6902  0       13.1873  27.0069  38.3419    2.3239   3.871

Clustered Instances

0      15 (  8%)
1       9 (  5%)
2     175 ( 88%)
```

Fig 1.6 Final Output in WEKA

V. CONCLUSION

We considered the problem of finding a globally optimal partition, optimum with respect to improved criterion, of a given documents into a specified number of clusters. We proposed some algorithms for this problem. By modeling partitioning problem as an optimization problem, improved partitioning clustering algorithm is proposed. Then the improved algorithm was extended by K-means, Kohonen SOM Algorithms through improved partitioning methods. In the experiments, we have used forest dataset of which characteristics are quite different. We conducted some experiments to test the performance of the improved algorithm and compare with the other algorithms. The improved algorithm is better than the K-means, Kohonen SOM in terms of the quality of the clustering solutions. From experiments, our methods improve the performance of in respect of Error Rate and Execution Time as compare to other algorithms.

REFERENCES

1. Balaji,S., .Srivatsa, S.K. (2012)" *Decision Tree induction based classification for mining Life Insurance Databases*" IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 2, No.3, June 2012 pp 699-704.
2. Dan,Ji, Jianlin,Qiu, Xiang, Gu,Li,Chen, Peng, He (2010)" *A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree*" 2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010) 978-0-7695-4108-2/10 © 2010 IEEE ,pp 2722-2728
3. Fayyad, Piatetsky-Shapiro, Smyth, "*From Data Mining to Knowledge Discovery: An Overview*" in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, Advances in Knowledge Discovery and Data Mining ,AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34.
4. Hatamlo, Abdolreza, Abdullah, Salwani (2011) "*Two stage algorithm for clustering*" in Data Mining and Optimisation Research Group, Center for Artificial Intelligence Technology Int' Conf. Data Mining DMIN 2011, pp 135-139.
5. Huo, Jianbing, Wang, Xizhao, Lu, Mingzhu, Chen, Junfen (2006) " *Induction of Multi-stage decision tree*" 2006 IEEE International Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan pp 835-839
6. .Kanellopoulos,Y. , Antonellis, P., Tjortjis,C., Makris,C., Tsirakis, N. (2011) " *k-attactors a partitional clustering algorithm for numeric data analysis*"Applied Artificial Intelligence, 25:97–115, 2011 Copyright 2011 Taylor & Francis Group, LLC pp 97-115.
7. Kristensen, Terje, Jakobsen, Vemund (2011)*" Three Different Paradigms for Interactive Data Clustering*" Int' Conf. Data Mining DMIN 2011 pp-3-9.
8. Li, Xiangyang, Ye, Nong (2006) "*A Supervised Clustering and Classification Algorithm for Mining Data With Mixed Variables*" IEEE Tranactionon systems, man and Cybernetics-part systems and humans, VOL. 36, NO. 2, MARCH 2006 pp 396-406
9. Lin, Zetao, Ge, Yaozheng, Tao, Guoliang (2005) "*Algorithm for Clustering Analysis of ECG Data"* Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, September 1-4, 2005 pp-3857-3860
10. Mao,Guojun ,Yang, Yi (2011)" *A micro-Cluster based Ensemble Approach for Classifying Distributed Data Streams*" 2011 23rd IEEE International Conference on Tools with Artificial Intelligence pp-753-759.
11. Shi, Yong, Meisner, Jerry (2011) "*An Approach to Selecting Proper Dimensions for Noisy Data"* Int' Conf. Data Mining DMIN 2011 pp 172-175.
12. Silwattananusarn, Tipawan , Tuamsuk, Kulthida (2012) " *Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012*" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012 pp 13-24.
13. Ramamohan, Y., Vasantharao, K., Chakravarti, C. Kalyana, Ratnam, A.S.K. (2012)" *A Study of Data Mining Tools in Knowledge Discovery Process*" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012 pp 191-194.

AUTHORS PROFILE

Heena Sharma, Research Scholar, Done B.Tech. (CSE) from L.L.R.I.E.T Moga (P.T.U), now doing M.Tech (CSE) from L.L.R.I.E.T Moga (P.T.U), Punjab, India, Research area is Data Mining.

Navdeep Kaur Kaler, Assistant Professor in Department Of CSE, L.L.R.I.E.T, Moga, Punjab, India, have done B.Tech. (CSE) from Guru Nanak Dev Engineering College, Ludhiana and have done M.Tech. (CSE) from Punjab Agriculture University (PAU), Ludhiana, Research area is Software Engineering, Data Mining