International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 6, June - 2013

An Enhance Approach For Cluster Analysis For Large Datasets

Ms. Ayushi Laud
PG Scholar
Department of CSE
SIMS, Indore, M.P., India

Prof. Ritesh Shah
Professor and Head
Department of I.T.
SIMS, Indore, M.P., India

***Abstract*- data mining is an application for extracting hidden knowledge from the rough data. in this era of technology data mining and its applications are growing and promising to provide an accurate way for data analysis and prediction. in this paper we presents an new cluster analysis approach which is motivated by K-mean clustering approach. Cluster analysis is helpful for protein analysis, image analysis and a verity of applications. Here proposed approach is given for 2d hyper planes where each data is represented as a point in this space, the theoretical approach for estimating points and their distance is given here. Additionally the proposed system model is provided here.**

***Keywords*- centroid based algorithm, clustering, distance, computational complexity.**

I. INTRODUCTION

Machine learning focuses on prediction, based on *known* properties learned from the training data**[4]**. Data mining (which is the analysis step of Knowledge Discovery in Databases) focuses on the discovery of (previously) *unknown* properties on the data. Machine Learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience **[16]**. It a branch of artificial intelligence, is about the construction and study of systems that can learn from data.

Representation of data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the property that the system will perform well on unseen data instances; the conditions under which this can be guaranteed are a key object of study in the subfield of computational learning theory.

Clustering is such one of the functionality of data mining. Clustering is the task of segmenting a diverse group into a number of more similar subgroups. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other cluster. Clustering is an unsupervised learning technique **[3]**. The main advantage of clustering analysis is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes. Clustering algorithm can be applied in many domains: image segmentation, object and character recognition and document retrieval **[5].**

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

In machine learning, **unsupervised learning** refers to the problem of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. This distinguishes unsupervised learning from supervised learning and reinforcement learning. Unsupervised learning is closely related to the problem of density estimation in statistics.**[1]** However unsupervised learning also encompasses many other techniques that seek to summarize and explain key features of the data. Many methods employed in unsupervised learning are based on data mining methods used to preprocess data. Approaches to unsupervised learning include:

- clustering (e.g., k-means, mixture models, hierarchical clustering),
- blind signal separation using feature extraction techniques for dimensionality reduction (e.g., Principal component analysis, Independent component analysis, Non-negative matrix factorization, Singular value decomposition). **[2]**

II. BACKGROUND

in this section of this paper includes various efforts, methods and techniques that are previously proposed and implemented by the different authors and researchers.

A similar effort **[6]** attempts have been made to solve the problem of clustering categorical data via cluster ensembles, with the results being competitive to conventional algorithms, it is observed that these techniques unfortunately generate a final data partition based on incomplete information. The underlying ensemble-information matrix presents only cluster-data point relations, with many entries being left unknown. The paper presents an analysis that suggests this problem degrades the quality of the clustering result, and it presents a new link-based approach, which improves the conventional matrix by discovering unknown entries through similarity between clusters in an ensemble. In particular, an efficient

link-based algorithm is proposed for the underlying similarity assessment. Afterward, to obtain the final clustering result, a graph partitioning technique is applied to a weighted bipartite graph that is formulated from the refined matrix. Experimental results on multiple real data sets suggest that the proposed link-based method almost always outperforms both conventional clustering algorithms for categorical data and well-known cluster ensemble techniques.

The empirical study, with different ensemble types, validity measures, and data sets, suggests that the proposed link-based method usually achieves superior clustering results compared to those of the traditional categorical data algorithms and benchmark cluster ensemble techniques. The prominent future work includes an extensive study regarding the behavior of other link-based similarity measures within this problem context. Also, the new method will be applied to specific domains, including tourism and medical data sets. **[6]**

Due to the very large size of the databases, it is highly desirable to perform these updates incrementally. In this paper **[7]**, present the new algorithm based on Genetic algorithm. Our algorithm is applicable to any database containing data from a metric space, e.g., to a spatial database. Based on the formal definition of clusters, it can be proven that the incremental algorithm yields the same result as any other algorithm. A performance evaluation of algorithm Incremental Clustering using Genetic Algorithm (ICGA) on a spatial database is presented, demonstrating the efficiency of the proposed algorithm. ICGA yields significant speed-up factors over other clustering algorithms. ICGA requires distance function and, therefore, it is applicable to any database containing data from a metric space. In the future, deletions will be considered to further improve the efficiency of ICGA.

The K-means algorithm is one of the most common techniques used for clustering. However, the results of K-means depend on the initial state and converge to local optima. In order to overcome local optima obstacles, a lot of studies have been done in clustering. This paper **[8]** presents an efficient hybrid evolutionary optimization algorithm based on combining Modify Imperialist Competitive Algorithm (MICA) and K-means (K), which is called K-MICA, for optimum clustering N objects into K clusters. The new Hybrid K-ICA algorithm is tested on several data sets and its performance is compared with those of MICA, ACO, PSO, Simulated Annealing (SA), Genetic Algorithm (GA), Tabu Search (TS), Honey Bee Mating Optimization (HBMO) and K-means. The simulation results show that the proposed evolutionary optimization algorithm is robust and suitable for handling data clustering.

The results illustrate that the proposed Hybrid K-MICA optimization algorithm can be considered as a viable and an efficient heuristic method to find optimal solutions for clustering problems of allocating N objects to K clusters.

Classification is an important step for automatic recognition. This paper present a new Classification method based on our Improved K-means clustering algorithm. As it is known that the performance of the traditional k-means algorithm largely depends on the choice of the initial centers, and the algorithm generally uses random procedures to get them. In order to improve the efficiency of the k-means algorithm, a good selection method of clustering starting centers is proposed in **[9]**. The proposed algorithm determines an initial scale for each cluster of patterns, and calculates initial clustering centers according to the norm of the points. Experiments results show that the proposed algorithm provides good performance of clustering.

The proposed initial clustering algorithm **[9]** has the following merits: (1) generating better initial cluster centers with little more calculation cost; (2) the number of iterations and the total calculation times of the k-means method employing the initial clustering algorithm are less than original one; (3) better final clusters are obtained by the method with the initial clustering algorithm. Therefore, proposed initial clustering algorithm can provide good performance of data clustering.

In **[10]** the most representative algorithms K-Means and K-Medoids were examined and analyzed based on their basic approach. The best algorithm in each category was found out based on their performance. The input data points are generated by two ways, one by using normal distribution and another by applying uniform distribution. The time taken for one execution of the program for the Uniform Distribution is less than the time taken for Normal Distribution. Usually the time complexity varies from one processor to another processor, which depends on the speed and the type of the system. The partition based algorithms work well for finding spherical-shaped clusters in small to medium-sized data points. The advantage of the K-Means algorithm is its favorable execution time. Its drawback is that the user has to know in advance how many clusters are searched for. From the experimental results, it is observed that K-Means algorithm is efficient for smaller data sets and K-Medoids algorithm seems to perform better for large data sets.

In this paper, author choose Page Ranking Method called Weighted Page Rank Algorithm **[11]** which is based on the popularity of the page by taking the importance of both the in links and out links of the pages. After that we use time rank algorithm for improving the weighted page rank score by using the visit time of the web page. So this concept is very useful to display most valuable pages on the top of the result list.

this paper **[11]** focused that by using Page Rank and Weighted Page Rank algorithms users may not get the required relevant documents easily, but in new algorithm we use two algorithms i.e. Weighted Page Rank and Time Rank in which user can get more relevant and important pages easily on the top list of the results as it employs web structure mining. In this firstly run Weighted Page Rank algorithm on k means Cluster then after that we apply time Rank on it. By doing this get more relevant data which satisfied the user query more accurately and give relevant data to the user.

a survey clustering algorithms for data sets appearing in statistics, computer science, and machine learning, and illustrate their applications in some benchmark data sets, the traveling salesman problem, and bioinformatics, a new field attracting intensive efforts. Several tightly related topics, proximity measure, and cluster validation, are also discussed in **[12]**.

Here, author place the focus on the clustering algorithms and review a wide variety of approaches appearing in the literature. These algorithms evolve from different research communities, aim to solve different problems, and have their own pros and cons. Though we have already seen many examples of successful applications of cluster analysis, there still remain many open problems due to the existence of many inherent uncertain factors. These problems have already attracted and will continue to attract intensive efforts from broad disciplines. We summarize and conclude the survey with listing some important issues and research trends for cluster algorithms.

- There is no clustering algorithm that can be universally used to solve all problems. algorithms are designed with certain assumptions and favor some type of biases.
- New technology has generated more complex and challenging tasks, requiring more powerful clustering algorithms.
    1) generate arbitrary shapes of clusters rather than be confined to some particular shape;
    2) handle large volume of data as well as high-dimensional features with acceptable time and storage complexities;
    3) detect and remove possible outliers and noise;
    4) decrease the reliance of algorithms on users-dependent parameters;
    5) have the capability of dealing with newly occurring data without relearning from the scratch;
    6) be immune to the effects of order of input patterns;
    7) provide some insight for the number of potential clusters without prior knowledge;
    8) show good data visualization and provide users with results that can simplify further analysis;
    9) be capable of handling both numerical and nominal data or be easily adaptable to some other data type Of course, some more detailed requirements for specific applications will affect these properties.
- At the preprocessing and post-processing phase, feature selection/extraction and cluster validation are as important as the clustering algorithms.

Using the advantages of K-means clustering and overcoming its disadvantages, a new text clustering algorithm is presented **[13]**. Firstly, texts are preprocessed to satisfy succeed process. Then, the paper analyzes common K-means clustering algorithm and improves the algorithm principle K-means and corrects its cluster seed selection method of to overcome efficiency of low stability of K-means algorithm which is very sensitive to the initial cluster center and the isolated point text. The experimental results indicate that the improved algorithm has a higher accuracy and has a better stability, compared with the original algorithm.

This article describes a network clustering technique based on PAM or *k*-medoid algorithm with appropriate modification. This algorithm works faster than the classical *k*-medoid based algorithms designed for networks and provides better results **[14]**. A better final cluster structure is obtained as the sum of within cluster spreads. The result has compared with those obtained by a graph *k*-medoid and a geodesic distance based network clustering algorithms. the degree of a node is a significant contributor for better clustering. If any or more than one of these properties change, then same algorithm with same number of clusters may produce different results. There is no such algorithm till now to claim to produce the best clustering results, under all possible combinations of network properties and number of clusters. here tried to show the effect of degree of a median node over clustering performance. It is observed that considering nodes with a moderate degree should also be considered as candidate medians. Here, used only very small network and simple distance measures. In future work, use other distance measures, large scale networks and will try to minimize the computation time as much as possible.

the number of clusters K needs to be initialized, the initial cluster centers are arbitrarily selected, and the algorithm is influenced by the noise points. In view of the shortcomings of the traditional K-Means clustering algorithm, this paper **[15]** presents an improved K-means algorithm using noise data filter. The algorithm developed density-based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By preprocessing the data to exclude these noise data before clustering data set the cluster cohesion of the clustering results is improved significantly and the impact of noise data on K-means algorithm is decreased effectively and the clustering results are more accurate. Proposed clustering algorithm based K-Means algorithm by excluding the interference of outliers firstly which can reach a result of higher accuracy but cost more time when dealing with large data sets.

III. CHALLENGES

in the previous section we study various research papers and articles that are provides us guidelines and demonstrated various challenges on data cluster analysis. we found a major issue, which is defined as our problem domain there are no generalized algorithm by which the all kinds of data are classify with high performance parameters. each and every algorithm which is used for cluster analysis are not complete satisfy the QoS parameters to be adoptable.

in the next section we propose a generalized algorithm that promises to achieve high accurate classification over 2d space.

IV. CONTRIBUTION

Proposed Algorithm- due to study we found that there are various algorithms and methods are previously proposed and implemented for finding the better optimum approach for cluster analysis. in the proposed algorithm we consider the dataset D with the n number of instances. Each instance contains a set of attributes (a1,a2,….an). This is processed using the proposed algorithm given as:

Input: dataset with n number of instances.

Output: dense and clear classification over 2D hyper plane.

1. Read all data sequentially.

2. Select an instance Xi

   For each Xi in data set Dn

   Find D(p,q)=√∑ (Qi-Pi)

   Create List Dlist (D(p,q))

   End for

3. Find a normalize value for each instance using

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \mu^2$$

4. The above value provide a point x for an instance.

5. Transpose the created Dn Repeat step 2 on transpose (Dn) that values generate y component for the instance.

6. plot all (Xi,Yi) in 2d vector space.

Above given algorithms first generate the entire X component for a point and then using the column wise generates Y component. The overall points are plotted over 2D vector spaces. In the next step the centroids are mounted in the 2D space. The proposed algorithms flow diagram is given using the fig 1.

to justify the proposed algorithm we compare the proposed algorithm with the K mean clustering algorithm over various performance parameters i.e. accuracy, memory uses, classification time and others the proposed system model for algorithm selection and comparison is given using fig 2.

Fig 1 shows the flow diagram

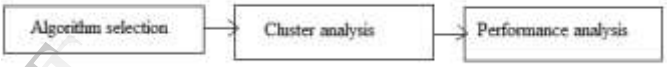

Fig 2 shows the system

Algorithm selection: here both algorithms is placed and evaluated as user selected.

Cluster analysis: using this module data model is generated using the training data pattern.

Performance evaluation: in this phase the model is used for prediction and cluster analysis additionally the performance of the system is evaluated using n cross validation technique.

V. CONCLUSION

in this paper we review the various clustering methods and techniques and find our problem to work, to solve the identified problem here we propose an generalized algorithm using column and row wise data analysis, and for cluster generation uses the scaling techniques.

in near future we implement the proposed data model using visual studio IDE and provide the performance and merits and limitation of our proposed work.

VI. REFERENCES

[1] Acharyya, Ranjan (2008); A New Approach for Blind Source Separation of Convolutive Sources, ISBN 978-3-639-07797-1. (this book focuses on unsupervised learning with Blind Source Separation)

[2] Jordan, Michael I.; Bishop, Christopher M. (2004). "Neural Networks". In Allen B. Tucker. Computer Science Handbook, Second Edition (Section VII: Intelligent Systems). Boca Raton, FL: Chapman & Hall/CRC Press LLC. ISBN 1-58488-360-X.

[3] T.Velmurugan and T.Santhanam "Computational Complexity between KMean and K-Medoids Clustering Algorithms for Normal and Uniform Distribution of Data Points",Journal of Computer Science 6(3):363-368,2010.

[4] Phil Simon (March18,2013). Too big to ignore: The Business Case for Big Data. Willey.p.89.ISBN 978-1118638170.

[5] A.M. Fahim,G.Saake,A.M. Salem,F.A.Torkeyand M.A.Ramadan " K-Means for Spherical Clusters with Large Variance in Sizes",world Academy of Science ,Engineering and Technology 45,2008.

[6] Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price "A Link-Based Cluster Ensemble Approach for Categorical Data Clustering" VOL. 24, NO. 3, MARCH 2012.

[7] Atul Kamble "Incremental Clustering in Data Mining using Genetic Algorithm" International Journal of Computer Theory and Engineering, Vol. 2, No. 3, June, 2010 1793-8201.

[8]Taher Niknam ,ElaheTaherianFard , NargesPourjafarian , AlirezaRousta a "An efficient hybrid algorithm based on modified imperialist competitive algorithm and K-means for data clustering " Engineering Applications of Artificial Intelligence 24 (2011) 306–317.

[9] Zhili Zhao, Bo Liu and Wei Li " Image Clustering Based on Extreme K-means Algorithm" IEIT Journal of Adaptive & Dynamic Computing, 2012(1), 12-16, 2012 © (2011) Institute of Electronic and Information Technology doi: 10.5813/www.ieit-web.org/IJADC/2012.1.3.

[10] T. Velmurugan and T. Santhanam "Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points" Journal of Computer Science 6 (3): 363-368, 2010
ISSN 1549-3636 © 2010 Science Publications.

[11] Amar Singh, Navjot Kaur "A Survey on k-Means Clustering Algorithm Using Different Ranking Methods in Data Mining" A Monthly Journal of Computer Science and Information Technology ISSN 2320–088X IJCSMC, Vol. 2, Issue. 4, April 2013, pg.111 – 115.

[12] Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE " Survey of Clustering Algorithms" IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005.

[13] Li Xinwu " Research on Text Clustering Algorithm Based on Improved K-means" 2010 International Conference On Computer Design And Appliations (ICCDA 2010).

[14] Samik Ray and Malay K. Pakhira "Clustering of Scale Free Networks Using a k-medoid Framework" International Conference on Computer & Communication Technology (ICCCT)-2011.

[15] Juntao Wan g and Xiaolong Su " An improved K-Means clustering algorithm" 978-1-61284-486-2/111$26.00 ©2011 IEEE.

[16] Pat Langley, Stanford and Herbert A. Simon, Pittsburgh. "Application of Machine Learning and Rule Induction".