AN ANALYSIS OF LEUKEMIA DATA SET FOR VARIOUS CLUSTERING ALGORITHMS

S. Shylaja[1] *S. Ranjitha kumari*[2]

ABSTRACT

The data mining process is to extract information from large database, and it is non-trivial process of identifying valid, novel, potential useful and understandable pattern in data. The data mining task is using two major categories of predictive and descriptive tasks. Data mining involves the outlier detection, classification, clustering, regression and summarization. The clustering is the most important technique in data mining, which divides data into groups of similar object. Each groups (= cluster) consist of object that are similar among themselves. A wide range of clustering algorithms is available in literature and still an open area for researcher. In this paper Consider Affinity Propagation Clustering Algorithm and K-Means++ Clustering Algorithm. And I have tested two Kinds of Data set and the Experimental results shows Increase the Accuracy and Decrease the Execution Time

Keywords: Data mining, clustering, k-means++, Affinity propagation.

I. INTRODUCTION

The Data mining is the process of discovering useful information (i.e. patterns) underlying the data. Powerful techniques are needed to extract patterns from large data because traditional statistical tools are not efficient enough anymore. Clustering is an important data mining technique that puts together similar objects into a collection in which the objects exhibit certain degree of similarities. Clustering also separates dissimilar objects into different groups. This has made clustering an important research topic of diverse fields such as pattern recognition, bioinformatics and data mining. It has been applied in many fields of study, from ancient Greek astronomy to present-day insurance industry and medical.

astronomy to present-day insurance industry and medical.

Affinity Propagation is a clustering algorithm that identifies a set of exemplar points that are representative of all the points in the data set. The exemplars emerge as messages are passed between data points, with each point assigned to an exemplar. AP attempts to find the exemplar set which maximizes the net similarity, or the overall sum of similarities between all exemplars and their data points. k-means++ (David Arthur et. Al., 2007) is another variation of k-means, a new approach to select initial cluster centers by random starting centers with specific probabilities is used.

[1] Research Scholar, Department of computer Science, RVS College of Arts & Science, Coimbatore, Tamil Nadu, India shylurose.ss@gmail.com

[2] Assistant Professor, Department of computer Applications (MCA), RVS College of Arts & Science, Coimbatore, Tamil Nadu, India Ranjithakumari@rvsgroup.com

In this paper, Consider Affinity Propagation Clustering Algorithm and K-Means++ Clustering Algorithm. And I have tested two Kinds of Data set and the Experimental results shows Increase the Accuracy and Decrease the Execution Time.

II. Clustering

Clustering is the task of segmenting a heterogeneous population into a number of more homogeneous sub groups or clusters. Clustering and clusters are not synonymous. A clustering is an entire collection of clusters; a cluster on the other hand is just one part of the entire picture. Clustering is a division of data into group of similar objects. Each group, called cluster consist of objects that are similar amongst themselves and dissimilar compared to objects of other groups. [8]

III. Cluster Analysis

The process of grouping a set of physical or abstract object into classes of similar objects are called clustering. A cluster is a collection of data object that are similar to one another within the same cluster and are dissimilar to the object in other clusters.

Cluster analysis is an important human activity. One learns how to distinguish between cats and dogs, or between animals or plants, by continuously improving subconscious clustering schemes. Cluster analysis has been widely used in numerous applications, that including pattern recognition, data analysis, image processing, and market research by clustering.[6]

IV. Clustering Algorithms

There are many types of algorithms are used in clustering. That is given below.

4.1 K-Means ++ Algorithm :

k-means++ (David Arthur et. Al., 2007) is another variation of k-means, a new approach to select initial cluster centers by random starting centers with specific probabilities is used. The steps used in this algorithm are described below:

- Step 1: Choose first initial cluster center c_1 randomly from the given dataset X .

- Step 2: choose next cluster center $c_i = x_j$ " X with probability p_i where $p_j = \frac{D(X_j)^2}{\sum_{x \in X} D(X)^2}$.

D(X) denote the shortest distance from x to the closest center already chosen.

- Step 3: Repeat step2 until k cluster centers are chosen.

- Step 4: After initial selection of k cluster centers, Apply k-means algorithm to get final k clusters.

4.2 AFFINITY PROPOGATION ALGORITHM :

Affinity propagation (AP) can be viewed as a method that searches for minima of an energy function

$$E(C)= -"^{N} S(I,cj)\, s(I,cj)d" 0$$

$$I=1$$

Each label ci indicates the exemplar of the data point i, while s(i,ci) is the similarity between data point i and its exemplar ci.

For ci = i, s(i, ci) is the input preference for data point i indicating how suitable data point i can be the exemplar. In most cases, the statistical and geometrical structure of a data set is unknown so that it is reasonable to set all the preference value the same. The bigger this shared value is, the larger the number of clusters is. Throughout the following of this paper, the preferences are set to the same value if not mentioned. The process of AP can be viewed as a message communication process with two kinds of messages exchanged among data points, named responsibility and availability. The algorithmic is stated below:[8]

Input: s(i, k): the similarity of point i to point k.

p(j): the preferences array which indicates the preference that data point j is chosen as a cluster center.

Output:

idx(j): the index of the cluster center for data point j.

dpsim: the sum of the similarities of the data points to theircluster centers.

netsim: the net similarity (sum of the data point similarities and preferences).

expref: the sum of the preferences of the identified cluster centers

netsim: the net similarity (sum of the data point data point similarities and preference)

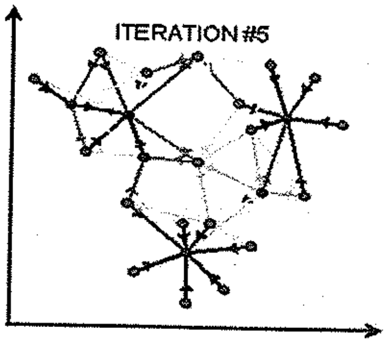

Figure 1 : Iteration affinity propagation

step1: Initialization the availability a(i.k) to zero

$$a(i,k)=0 \quad (1)$$

step2: update the responsibility using rule

$$r(i,k) \leftarrow s(i,k) - \max_{k' \text{s.t. } k' \neq k} \{a(i,k'). s(i,k')\}. \quad (2)$$

step3: update the availability using the rule

$$a(i,k) \leftarrow \min\{0, r(k,k) \sum_{i' \text{ s.t. } i' \neq i,k} \max\{0, r(i',k)\}\} \quad (3)$$

The self-availability is updated differently

$$a(k,k) \rightarrow \sum_{i' \text{ s.t. } i' \neq k} \max\{0, r(i',k)\}. \quad (4)$$

Step 4: The message-passing procedure may be terminated after a fixed number of iterations, after changes in the messages fall below a threshold or after the local decisions stay constant for some number of iterations.

Availabilities and responsibilities can be combined to make the exemplar decisions. For point i, the value of k that maximizes a(i, k)+r(i, k) either identifies point i as an exemplar if k=i or identifies the data point that is the exemplar for point i. When updating the messages, numerical Oscillations must be taken into consideration. As a result, each message is set to ë times its value from the previous iteration plus 1- ë times its prescribed updated value. The ë should be larger than or equal to 0.5 and less than 1. If ë is very large, numerical oscillation may be avoided, but this is not guaranteed. Hence a maximal number of iterations are set to avoid infinite iteration in AP clustering.

V. Collecting The Data:

In this paper, there are two data sets are used to compare the proposed algorithm with the existing algorithm. Iris Data set and Wine Data set This is the data sets which are taken from the http://www.ics.uci.edu/~mlearn/MLRepository.html UCI Repository.

5.1 Iris data set:

Iris is well known and studied dataset. It was first introduced by Sir Ronald Aylmer Fisher and describes collected data that quantifies the geographic variation of Iris flowers in the Gaspe Peninsula. It consists of 50 samples from each of three species of Iris flowers: Iris Setosa, Iris Versicolor and Virginica. Each sample has four features, describing some measurements, in cm, of the flowers: Sepal Length, Sepal Width, Petal Length, and Petal Width. A particular characteristic of this dataset is that the class of Iris-Setosa flowers is linearly separable from the other two. Also, the attributes Petal Length and Petal Width are highly correlated.

5.2 Wine data set:

Here, two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult: [Web Link] or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

VI. Result Over IRIS And Wine Data Set :

The analysis of two different types of clustering algorithm (Affinity propagation and K-Means++) is done with the help of the Iris Data set and Wine Data set. The Experimental Result shows the Comparable performance of Affinity Propagation and K-Means++ clustering Algorithm. Both Algorithms are performing well than other Clustering algorithms. In My proposition of K-Means++ Clustering Algorithm is given the better result than Affinity propagation. The K-Means++ is Increase the Average Accuracy and Decrease the Execution Time while we comparing with the Affinity Propagation Clustering Algorithm. The Average Accuracy and Execution Time of these two types of Algorithms are shown below in the Table.

Table 1 : AccuracyValues of Affinity Propagation and K-Means++ Algorithm

S.No	Algorithm	Accuracy
1	Affinity Propagation	88%
2	K-Means++	94.93%

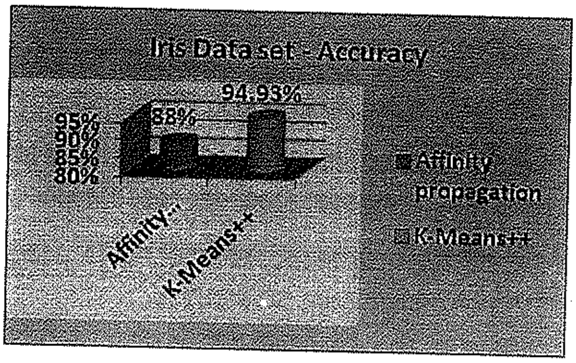

Figure 2 : Accuracy Values of Affinity Propagation and K-Means++ Algorithm

This graph shows the Accuracy or measurement of the system between the Affinity Propagation and K-Means++ algorithm for most relevant to the Iris Data set. The algorithm in X-axis and the Accuracy value percentage in the Y-axis Algorithms are measured. The Accuracy value of K-Means++ is higher than the Affinity Propagation algorithm.

Table 2 : AccuracyValues of Affinity Propagation and K-Means++Algorithm

S.No	Algorithm	Accuracy
1	Affinity Propagation	91%
2.	K-Means++	94.61%

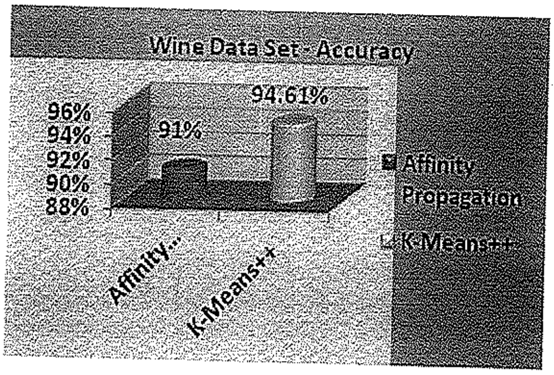

Figure 3 : Accuracy Values of Affinity Propagation and K-Means++ Algorithm

Table 3 : Execution Time For Iris Data Set

S.No	Algorithm	Execution Time
1	Affinity Propagation	0.23 Sec
2	K-Means++	0.12 Sec

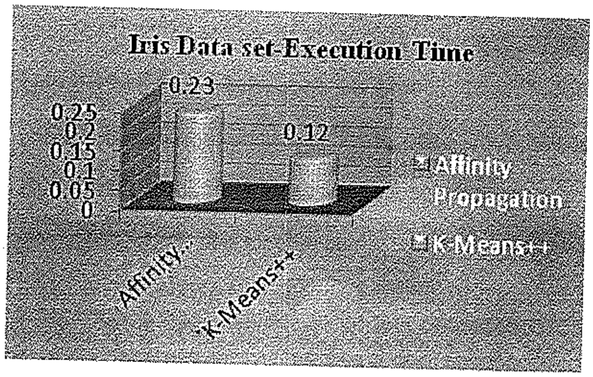

Figure 4 : Execution time of Affinity Propagation and K-Means++ Algorithm

This graph shows the Execution time comparison between the Affinity Propagation and K-Means++ Algorithm. X-axis defines the algorithm and Running time in milliseconds in the Y-axis. The Running time of K-Means++ algorithm is lower than the Affinity Propagation algorithm.

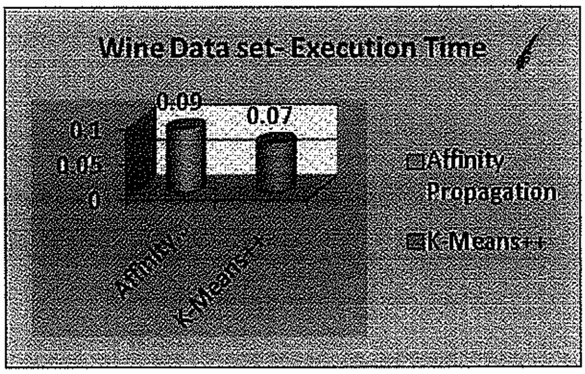

Figure 5 : Execution time of Affinity Propagation and K-Means++ Algorithm

This graph shows the Running time comparison between the Affinity Propagation and K-Means++ Algorithm. X-axis defines the algorithm and Running time in milliseconds in the Y-axis. The Running time of K-Means++ algorithm is lower than the Affinity Propagation algorithm.

VII. Conclusion And Future Work :

The Analysis of Affinity Propagation and K-Means++ Clustering Algorithms are done with the help of Iris data set and Wine Data set. Then the Experimental Results shows the Average Accuracy and Execution Time which is applied in the Two Kinds of Data sets. The K-Means++ Algorithm is Performing well than the Affinity Propagation Clustering Algorithm. And K-Means++ Algorithm is Advanced version of K-Means, this is Increase the Average Accuracy and decrease the Execution Time While we are Comparing to the Affinity Propagation Clustering Algorithm. Performance of this K-Means++ algorithm can be improved with the help of variants of other Data sets in other papers. In Future, Efficient k-means, fuzzy logic to get better quality of cluster. So these algorithm help to get Good Result.

Referemces :

1. Arun. K. Pujari, "*Data Mining Techniques*", Universities press (India) Limited 2001, ISBN81-7371-3804.

2. Alizadeh A., Eisen M.B, Davis R.E, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000; 403 (6769): 503–511.

3. Bingham E, Mannila H: Random projection in dimensionality reduction: applications to image and text data. Knowledge Discovery and Data Mining 2001:245-250

4. David Arthur and Sergei Vassilvitskii: k-means++:The advantages of careful seeding, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027—1035, 2007.

5. Gibbons F.D, Roth F.P. Judging the quality of gene expression-based clustering methods using gene annotation. Genome Res. 2002;12(10):1574–1581.

6. L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, 1990.

7. MacQueen, J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In Proc. of 5th Berkley Symposium on Mathematical Statistics and Probability, Volume I: Statistics, pp. 281–297 ..

8. Parvesh Kumar, Siri Krishan Wasan., Comparative Analysis of k-mean Based Algorithms, IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.4, April 2010 314.

9. Greg Hamerly *"Making k-means evenfaster"* 2010 academic.research.microsoft.

10. Federico Ambrogi, Elena Raimondi, Daniele Soria, Patrizia Boracchi and Elia Biganzoli1 Cancer profiles by Affinity Propagation University of Nottingham, School of Computer Science, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB,

11. Jinze Liu, Jiong Yang and Wei Wang, *"Biclustering in Gene Expression Data by Tendency"*. Paul Bunn, Rafail Ostrovsky *"Secure Two-Party k-Means Clustering"* 2007.

12. Margaret H. Dunham and S. Sridharz *"Data Mining Introductory and Advanced Topics"* Dorling Kindersley (India) Pvt. Ltd., 2006.

13. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd edition, Morgan Kaufmann, 2011. p. 444.

14. T.W. Liao, *"Clustering of Time Series Data: A Survey,"* Pattern Recognition, vol. 38, no. 11, pp. 1857-1874, Nov. 2005.

15. A.K. Jain, *"Data Clustering: 50 Years Beyond K-means," Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, June 2009.

16. S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. OCallaghan, *"Clustering Data Streams: Theory and Practice," IEEE Trans. Knowledge and Data Eng.*, vol. 15, no. 3, pp. 515-528, May 2003.

17. J. Beringer and E. Hullermeier, *"Online Clustering of Parallel Data Streams," Data and Knowledge Engineering*, vol. 58, no. 2, pp. 180-204, Aug. 2006.

18. A. Likas, N. Vlassis, and J.J. Verbeek, *"The Global k-means Clustering Algorithm," Pattern Recognition*, vol. 36, no. 2, pp. 451-461, Feb. 2003.

19. A.M. Alonso, J.R. Berrendero, A. Hernandez, A. Justel, *"Time Series Clustering based on Forecast Densities," Computational Statistics and Data Analysis*, vol. 51, no. 2, pp. 762-776, Nov. 2006.

20. B.J. Frey and D. Dueck, *"Response to Comment on 'Clustering by Passing Messages Between Data Points'," Science*, vol. 319, no. 5864, pp. 726a-726d, Feb. 2008.

21. Hui Li, Sourav S. Bhowmick, Aixin Sun, Blog Cascade Affinity: Analysis and Prediction portal.acm.org/ft_gateway.

22. Ambrogi, Elena Raimondi, Daniele Soria, Patrizia Boracchi and Elia Biganzoli1 Cancer profiles by Affinity Propagation University of Nottingham, School of Computer Science, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB.

23. Wu Jiang, Fei Dingy, Qiao-Liang Xiang An Affinity Propagation Based Method for Vector Quantization Codebook Design College of Computer and Information Science, North-eastern University.

24. C. Yang, L. Bruzzone, R.C. Guan, L. Lu, and Y.C. Liang, *"Incremental and Decremental Affinity Propagation for Semisupervised Clustering in Multispectral Images," IEEE Trans. Geosci. and Remote Sens.*, vol. 51, no. 3, pp. 1666-1679, Mar. 2013.

25. C.C. Aggarwal, J. Han, J. Wang, and P.S. Yu, A Framework for Clustering Evolving Data Streams, *Proc. 29th Int'l Conf. Very Large Data Bases* (VLDB '03), pp. 81-92, 2003.

26. D. Chakrabarti, R. Kumar, and A. Tomkins, *"Evolutionary Clustering," Proc. Knowledge Discovery and Data Mining* (KDD '06), pp. 554-560, Aug. 2006.

27. Brendan J. Frey and Delbert Dueck. Clustering by Passing Messages between Data Points. Science 315, 972 (2007).

28. D. Jiang, C. Tang, and A. Zhang, Cluster analysis for gene expression data: a survey, IEEE Transactions on Knowledge and Data Engineering 16 (2004).

Authors Biography

S.Shylaja is now a M.Phil. Research Scholar in Rathnavel Subramaniam College of Arts and Science, at Affiliated to Bharathiar University. She is Received B.Sc (CS) degree in Angappa College of Arts and Science. Malumichampatti, Coimbatore in 2011, and Completed M.Sc (CS) in Government Arts and Science College, at Affiliated to Bharathiar University. Coimbatore in 2013. And her specialization is Data Mining.

Ms.S.Ranjitha kumari is a Assistant Professor in Rathnavel Subramaniam College of Arts and Science, at Affiliated to Bharathiyar University. She has more than nine years of teaching experience. Her areas of interest are Network Security and Machine Learning. She has successfully produced six M.Phil Scholars and guiding three scholars under Bharathiar University