Volume 1, No. 4, Nov-Dec 2010

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Message Passing between Data Point on Clustering Algorithm for Gene Leukemia Dataset

D.Napoleon
Assistant Professor
Department of Computer Science
School of Computer Science and Engineering
Bharathiar University,Coimbatore,
Tamil Nadu, India
. mekaranapoelon@yahoo.co.in

G.Baskar
Research scholar
Department of Computer Science
School of Computer Science and Engineering
Bharathiar University,Coimbatore,
Tamil Nadu, India
baskarb@yahoo.com

Abstract: Clustering (or cluster analysis) aim to organize a collection of data item in to clusters, such that items within a cluster are more "similar" to each other than they are to item in the other clusters. Affinity propagation (AP) is a clustering algorithm which has much better performance than traditional clustering approach such as k-means algorithm. AP clustering handles large datasets by merging the exemplars learned from subsets. The algorithm is tested on leukemia data set. The experimental results show that affinity propagation outperforms clustering execution time and convergence rate.

Keywords: Data Mining, Clustering, k-means, x-means, Affinity propagation

I. INTRODUCTION

Clustering is to reduce the amount of data by categorizing or grouping similar data items together. Such grouping is pervasive in the way human's process information, and one of the motivations for using clustering algorithms is to provide automated tools to help in constructing categories or taxonomies [Jardine and Sibson, 1971, Sneath and Sokal, 1973]. The methods may also be used to minimize the effects of human factors in the process.

Clustering methods [Anderberg, 1973, Hartigan, 1975, Jain and Dubes, 1988, Jardine and Sibson, 1971, Sneath and Sokal, 1973, Tryon and Bailey, 1973] can be divided into two basic types: hierarchical and partition clustering. Within each of the types there exists a wealth of subtypes and different algorithms for finding the clusters Often considered more as an art than a science, the field of clustering has been dominated by learning through examples and by techniques chosen almost through trial-and-error. Many fundamental advances in Clustering however have been proposed since the mid 2000s. Ding et al. have highlighted the relationship between *K*-means[1].

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. Gene expression, many genes can be studied. Microarray analysis is emerging as a powerful technique to study thousands of genes simultaneously in a single experiment. A common aim is to use the gene expression profiles to identify groups of genes or samples Alizadeh et al (Alizadeh, 2000), Bittner et al (Bittner,2000) and Nielsen et al (Nielsen,2002) have considered the classification of cancer types using gene expression datasets.

In this paper, we make a comparative analysis of k-means, x-means with affinity propagation, over leukemia dataset Comparison is made in respect of accuracy and convergence rate.

II. K-MEANS ALGORITHM

The k-means algorithm (MacQueen, 1967) is one of a group of algorithms called *partitioning methods*. The k -means algorithm is very simple and can be easily implemented in solving many practical problems. The k-means algorithm is the best-known squared error-based clustering algorithm. Consider the data set with 'n' objects, i.e.

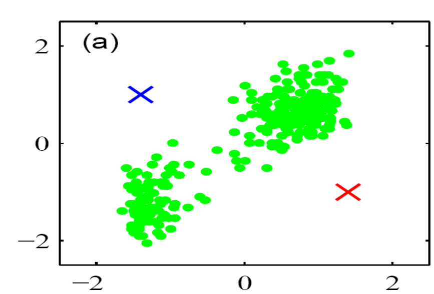


Figure. 1 Data and initial random

S = {xi : 1 ¡Ü i ¡Ü n}.

1) Initialize a k-partition randomly or based on some prior knowledge.

i.e. { C1 , C2 , C3 ,…….., Ck }.

2) Calculate the cluster prototype matrix M

(distance matrix of distances between k-clusters and

data objects) .

M = { m1 , m2 , m3, …….. , mk } where mi is a

column matrix 1× n .

3) Assign each object in the data set to the

nearest cluster - Cm i.e.

x j ¸Cm if || x j - Cm || ¡Ü || x j – Ci ||

Í 1 ¡Ü j ¡Ü k , j ,m where j=1,2,3,…….n.

4) Calculate the average of each cluster and

change the k-cluster centers by their

averages.

5) Again calculate the cluster prototype matrix M.

6) Repeat steps 3, 4 and 5 until there is no change for each cluster.

III X-MEANS ALGORITHM

X-means algorithm (Dan Pelleg and Andre Moore, 2000) searches the space of cluster locations and number of clusters efficiently to optimize theBayesian Information Criterion(BIC) or The Akaike Information Criterion(AIC) measure . The kd-tree technique is used to to improve the speed for the algorithm.In this algorithm , number of clusters are computed dynamically using lower and upper bound supplied by the user.

The algorithm consists of mainly two steps which are repeated until completion.

Step1: (Improve-Params) In this step , we apply k-means algorithm initially for k clusters till convergence. Where k is equal to lower bound supplied by the user.

Step2:(Improve -Structure) This structure improvement step begins by splitting the each cluster center into two children in opposite directions along a randomly chosen vector. After that we run k-means locally within each cluster for two clusters. The decision between the children of each center and itself is done comparing the BIC-values of the two structures.

Step 3: if k> =k_{max} (upper bound) stop and report to best scoring model found during search otherwise goto to step 1.

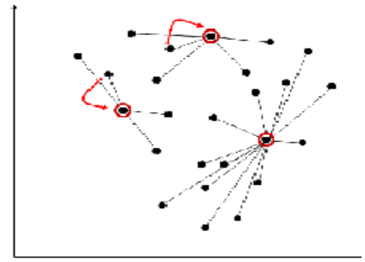


Figure 2 For each cluster pick best new center

IV AFFINITY PROPAGATION

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

$$E(C)=-\sum_{I=1}^{N} S(I,c_j)\ s(I,c_j)\leq 0$$

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 their cluster 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 similarities and preference)

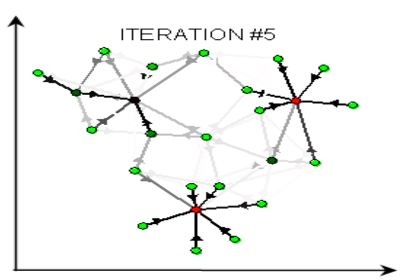


Figure. 3.Iteration affinity propagation

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

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

step2: update the responsibility using rule

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

step3: update the availability using the rule

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

The self-availability is updated differently

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

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 thatis the exemplar for point *i*. When updating the messages, numerical Oscillations must be taken into consideration. As a result, each message is set to λ times its value from the previous iteration plus $1-\lambda$ times its prescribed updated value. The λ should be larger than or equal to 0.5 and less than 1. If λ 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. RESULT OVER LEUKEMIA DATASET

The Leukemia data set is collections of gene expression measurements from 72 leukemia (composed of 62 bone marrow and 10 peripheral bloods) samples reported by Golub. It contains an initial initial training set composed of 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of acute myeloblastic leukemia (AML). Here we take two variants of leukemia dataset one with 50-genes.

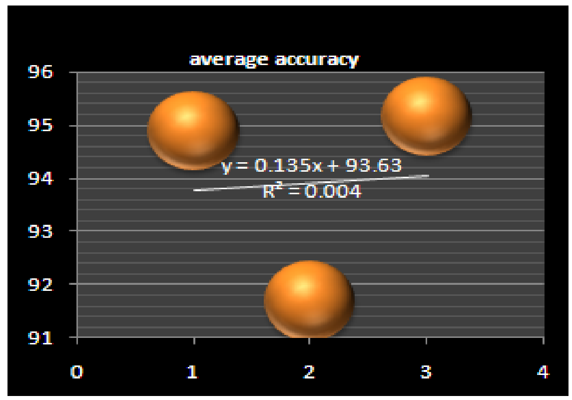
Table 1: Result over different variations of k-means algorithm using 50-gene leukemia (Total number of record present in dataset=72)

Clustering Algorithm	Correctly Classified	Average Accuracy
k-means	*68*	*94.88*
x-means	*67*	*91.67*
Affinity propagation	*69*	*95.15*

VI. CONCLUSION AND FUTURE WORK

The analyses of k-means, x-means algorithm are done with the help of leukemia dataset. The average accuracy is shown that the performance of affinity propagation algorithm is better in50 gene leukemia dataset,on clustering execution time and convergence rate and found much low error when compare with k-means.

Performance of this algorithm can be improved with the help of variants 3859-gene-leukemia using efficient k-means, fuzzy logic to get better quality of cluster. So these algorithm help to get good result.



Graph 1.1

VII. REFERENCES

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

[2] *Bagirov, A.M.[Adil M.]*, Modified global k-means algorithm for minimum sum-of-squares clustering problems, October 2008, pp. 3192-3199.

[3] Bloisi, D.D.[Domenico Daniele], Iocchi, L.[Luca], Rek-Means: A k-Means Based Clustering Algorithm,Springer DOI Link

[4] Cheung, Y.M.[Yiu-Ming], K*-Means: A new generalized k-means clustering algorithm, *PRL(24)*,No.15,November,2003.

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

[6] Dr. Padma R. Chavan, "Application of Bioinformatics in the Field of Cancer Research".

[7] E. Papageorgiou, I. Kotsioni, A. Linos, "Data Mining: A New Technique In Medical Research", Hormones 2005, 4(4):189-191.

[8] Brendan j.Frey and Delbert Duec clustering by passing message between data point science, 315(5814):972{976}

[9] Anjan Goswami. Department of Computer Science and Engineering" Fast and Exact Out of-Core and Distributed K-Means Clustering 2001

[10] A. Ben-Dor, R. Shamir, and Z. Yakhini, “Clustering gene expression patterns” in *J Comput Biol* 6(3-4):281-97.

[11] 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,

[12] Greg Hamerly “Making k-means even faster”2010academic.research.microsoft

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

[14] Jinze Liu, Jiong Yang and Wei Wang, “Biclustering in Gene Expression Data by Tendency”.

[15] Lai, J.Z.C.[Jim Z.C.], Liaw, Y.C.[Yi-Ching], Improvement of the k-means clustering filtering algorithm, PR(41),No.12, December2008.

[16] Paul Bunn, Rafail Ostrovsky “Secure Two-Party k-Means Clustering” 2007

[17] Margaret H. Dunham and S. Sridharz“Data Mining Introductory and Advanced Topics” Dorling Kindersley (India) Pvt. Ltd., 2006.

[18] Wu Jiang, Fei Dingy, Qiao-Liang Xiang An Affinity Propagation Based Method for Vector Quantization Codebook Design College of Computer and Information Science, Northeastern University

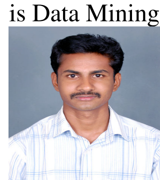
[19] Xuqing Zhang, Fei Wu, Dingyin Xia, and Yueting Zhuang Partition Affinity Propagation for Clustering Large Scale of Data in Digital Library College of Computer Science, Zhejiang University**,** Hangzhou,

[20] Zhang Y. , Mao J. and Xiong Z.: An efficient Clustering algorithm, In Proceedings of Second International Conference on Machine Learning and Cyber netics, November 2003.

[21] Zhenjie Zhang, Bing Tian Dai, Anthony K.H. Tung On the Lower Bound of Local Optimums in K-Means Algorithm2003.

AUTHOR:

D. Napoleon received the Bachelor’s Degree in B.Sc Physics from MaduraiKamaraj University in 99,Master’s Degree in Computer Applications from Madurai Kamaraj University in 2002, and M.Phil degree in Computer Science from Periyar University in 2007.He is working as Assistant Professor in the Department of Computer Science ,School of Computer Science Engineering, Bharathiar University, Coimbatore, Tamilnadu, India. He has published articles in National and International Journals. He has presented papers in National and International Conferences. His research area is Data Mining.



Mr. G.Baskar received his Master’s Degree in Information Technology in K.S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu India in 2008 and M.Phil Degree in Computer Science from Bharathiar University, Coimbatore, Tamil Nadu, India in 2010. His area of interest includes Data Mining**.**