International Journal of Mathematical Archive-2(7), July - 2011, Page: 1153-1158
Available online through www.ijma.info ISSN 2229 – 5046

MESSAGE PASSING BETWEEN DATA POINT ON GENE EXPRESSIONS FOR CANCER USING CLUSTERING ALGORITHM

D. NAPOLEON

Assistant Professor Department of Computer Science Bharathiar University
Coimbatore, Tamil Nadu, INDIA
E-mail: mekaranapoleon@yahoo.co.in

G. BASKAR*

Research Scholar Department of Computer Science Bharathiar University
Coimbatore, Tamil Nadu, INDIA
E-mail: baskarb@yahoo.com

M. PRANEESH

Research Scholar Department of Computer Science Bharathiar University
Coimbatore, Tamil Nadu, INDIA
E-mail: raja.praneesh@gmail.com

(Received on: 02-07-11; Accepted on: 14-07-11)

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. 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 Several years' different approaches have been proposed to improve global search properties of k-means algorithm and its performance on large data sets. The algorithm is tested on both colon and leukemia data set. The experimental results show that affinity propagation outperforms the k-means algorithm in terms of running time as well as the quality of the clustering.

Keywords: *Data Mining, k-means, global k-means, Affinity propagation.*

---

1. INTRODUCTION

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure In a collection of unlabeled data An important component of a clustering algorithm is the distance measure between data points. If the components of the data instance vectors are all in the same physical units then it is possible that the simple distance metric is sufficient to successfully group similar data instances. the problem arises from the mathematical formula used to combine the distances between the single components of the data feature vectors into a unique distance measure that can be To different clustering. Data mining is the process of discovering useful information that is patterns underlying the data Powerful techniques are needed to extract patterns from large data because traditional statistical tools are not efficient enough anymore[19].

Affinity Propagation has several advantages over alternative clustering and topic modeling approaches. K-means clustering algorithms assign each object to the best cluster. AP, on the other hand, is a clustering algorithm that finds the best assignment of all objects to clusters at the same time. Moreover, AP produces an exemplar that can best "summarize" the cluster. In colon and leukemia data, can effectively compress the stream of data,. Affinity Propagation make hard decisions on the cluster centers at each iteration. Affinity propagation is a low error, high speed, flexible, and remarkably simple clustering algorithm.

To classify the various types of cancer into its different subcategories, different data mining techniques have been used over gene expression data. A common aim is to use the gene expression profiles to identify groups of genes or samples

---

**Corresponding author: G. BASKAR*, *E-mail: baskarb@yahoo.com*

in which the members behave in similar ways. One might want to partition the data set to find naturally occurring groups of genes with similar expression patterns. Golub et al. (Golub, 1999), 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, Global k-means with affinity propagation, over colon and leukemia dataset Comparison is made in respect of accuracy and convergence rate.

2. K-MEANS ALGORITHM

The k-means algorithm (Mac Queen, 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.

$$S = \{xi: 1 ¡Ü i ¡Ü n\}.$$

Step: 1 Initialize a k-partition randomly or based on some prior knowledge.
i.e. {C1, C2, C3… Ck }.

Step: 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 .

Step: 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.

Step: 4 Calculate the average of each cluster and change the k-cluster centers by their averages.

Step: 5 Again calculate the cluster prototype matrix M.

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

3. GLOBAL K-MEANS

Introduced by A. Likas, N. Vlasis and J.J. Verbeek in the paper entitled "The Global *k*-means clustering algorithm" in 2003, the concept of clustering with Global*k*-means is partitioning the given dataset into *M* clusters so that a clustering criterion is optimized. The common clustering criterion is the sum of squared Euclide and distances between each data point and the cluster centroid.[17]

$$b_{n=}\sum_{j=I}^{n} \max(d^{j} - \| x_n - x_j \|^2, 0)$$

Global *k*-means deploys the *k*-means algorithm tofind locally optimal solutions by trying to keep the clustering error to a minimum. The *k*-means algorithm starts by placing the cluster center arbitrarily and at each step moves the cluster center with the aim to minimize the clustering error. The down side to this algorithm is that it is sensitive to the initial position of the cluster centers. To overcome this, *k*-means can be scheduled to run several times and each time with a different starting point. The gist of Global *k*-means is that instead of trying to find all cluster centers at once, it proceeds in an incremental fashion. Incremental in the sense that one cluster center is found at a time. Assume a *K*-clustering problem is to be solved; the algorithm starts by solving for a 1-clustering problem and the placement of the cluster center in this instance would equal the centroid of the given dataset. The next step would be to add another cluster center at its optimal position, given, the first cluster center has already been found. To do this, *N*-executions of *k*-means algorithm will be executed.

© 2011, IJMA. All Rights Reserved

with the initial positions of the cluster centers being the first cluster which was found when solving for a 1-clustering problem and the second cluster's starting position will be at $n\ x$ where 1 $N\ n$ . . The final answer for a 2-clustering problem will be the best solution from the *N*-executions of *K*-Means algorithm. Let (c1(k),…,ck(k)) denote the final solution for the *k*-clustering problem. We will solve it iteratively which means solving a 1-clustering problem, which means solving a 1-clustering problem, then a 2-clustering problem, until a (*k*-1)-clustering problem and the solution of *k*-clustering problem can be solved by performing *N*-executions of *k*-means algorithm with starting positions of (c1(k-1),…,c(k-1)(k-1),Xn). A simple pseudo code of it will be Problem: to solve k-clustering problem for dataset, X

For i=1 to k
{
If i = 1 then
$C_{i=}$ centroid of dataset, X
Else
For j=1 to N
Run k-means with initial values of
{ j i i X c c , ,..., 1 . }

With the final solution, (c1(k),…,ck(k)), Global *k*-means has actually found solutions of all *k*-cluster problem where *k*=1,…,*K* without needing any further computations. This assumption seems very natural: we expect that the solution of a *k*-clustering problem to be reachable (through local search) from the solution of a (*k*-1)-clustering problem, once the additional center is placed at an appropriate position within the data set. [10] Alas, the downside is that the computational time of Global *k*-means can be rather long.

4. AFFINITY PROPAGATION

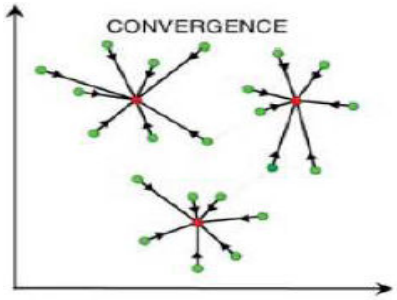


Fig 1. Iteration of message passing in Affinity Propagation

Affinity Propagation is a clustering algorithm that identifies a set of 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.[2]

Affinity Propagation (Frey and Dueck, 2006; 2007) takes as input a collection of real-valued similarities between data points, where the similarity $s(i, k)$ indicates how well the data point with index k is suited to be the class center for data point i. When the goal is to minimize the squared error, each similarity is set to a negative Euclidean distance: for points $\boldsymbol{x}i$ and $\boldsymbol{x}k$, $s(i, k) = -||\boldsymbol{x}i - \boldsymbol{x}k||2$.

The algorithm work on the following three steps:

Step1: Update responsibilities given availabilities. Initially this is a data driven update, and over time lets candidate exemplars competition for ownership of the data.

Step2: Update availabilities given the responsibilities. This gathers evidence from data points as to whether a candidate exemplar is a good exemplar.

Step3: Monitor exemplar decisions by combining availabilities and responsibilities. Terminate if reach a stopping point (e.g. insufficient change). The update rules require simple local computations and messages are exchanged between pairs of points with known similarities

© 2011, IJMA. All Rights Reserved

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)

Algorithm 1 Affinity Propagation

Step1: Initialization the availability a(i.k) to zero

$$a(i,k)=0 \tag{1}$$

Step2: update the responsibility using rule

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

Step3: update the availability using the rule

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

The self-availability is updated differently

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

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

5. DATA SET

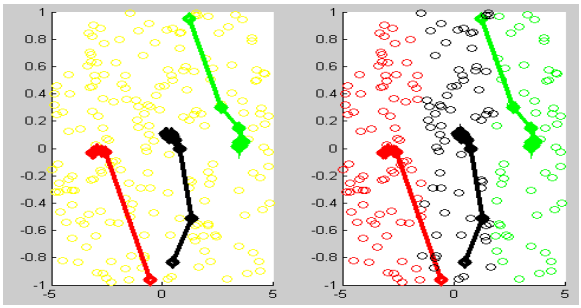


Fig 2.Cluster formation over leukemia

© 2011, IJMA. All Rights Reserved

Two data set has been used colon and leukemia, the colon dataset is a collection of gene expression measurements from 62 colon biopsy samples reported by Alon. It contains 22 normal and 40 colon cancer samples .the colon dataset consists of 2000 genes. The Leukemia data set is a collection of gene expression measurements from 72 leukemia (composed of 62 bone marrow and 10 peripheral bloods) samples reported by Golub.

TABLE- 1

Result over different variation of k-means algorithm using 3859-gene leukemia (total number of record present in dataset=72)

Clustering algorithm	Correctly classified	Average accuracy
K-Means	61	84.72
Global K-Means	65	91.67
Affinity Propagation	67	92.15

TABLE- 2

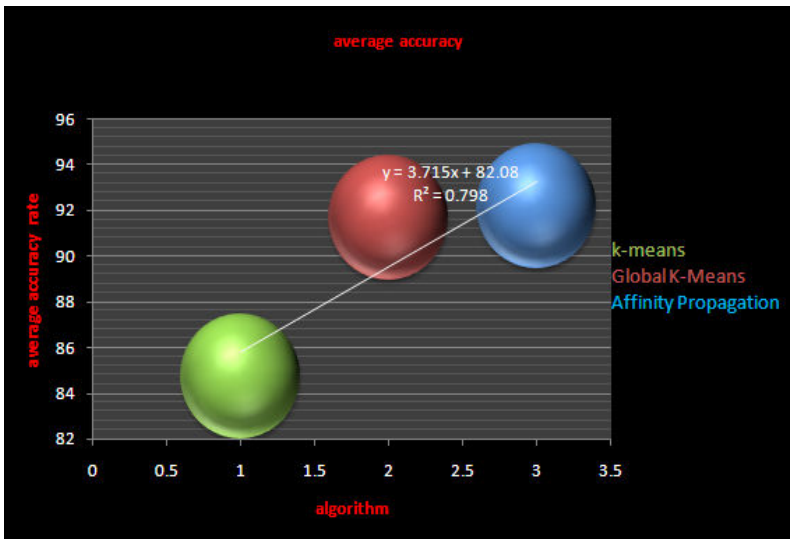
Result over clustering algorithm using 2000 gene colon dataset (Total number of records present in data set =62)

Clustering algorithm	Correctly classified	Average accuracy
K-Means	33	53.23
Global K-Means	37	59.68
Affinity propagation	35	60.41

6. CONCLUSION AND FUTURE WORK

The analysis of k-means, global k-means algorithm is done with the help of colon and leukemia dataset. The average accuracy is shown that the performance of affinity propagation algorithm is better in leukemia dataset. As far convergence rate is also higher and speed of execution time is good and found much low error when compare with k-means, global k-means. Performance of this algorithm can be improved with the help of variants K-means++, fuzzy logic to get better quality of cluster. So these algorithm help to get good result.

GRAPH-1



© 2011, IJMA. All Rights Reserved

7. REFERENCES

[1] 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.

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

[3] Golub T. R, Slonim D.K, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286(5439):531–53.

[4] likas, A., Vlassis,M.& Verbeek,j.(2003), the global k-means clustering algorithm, pattern recognition,36,451-461.

[5] Nielsen T.O, West R.B, Linn S.C, et al. Molecular characterization of soft tissue tumours: a gene expression study. Lancet2002.

[6] Yeung K. Y, Haynor D.R, Ruzzo W. L. Validating clustering for gene expression data. Bioinformatics. 2001.

[7] Frank R Kschischang, Brendan J. Frey, and hans Andrea Loeliger, Factor graphs and the sum-product algorithm IEEE Transactions on Information Theory 47(2):498{519} 2001.

[8] Gibbons F.D, Roth F.P. Judging the quality of gene expression-based clustering methods using gene annotation. Genome Res. 2002; 12(10):1574–158.

[9] Michele Leone, Sumedha, and MartinWeigt. Clustering by soft-constraint affinity propagation: Applications to gene-expression data. Bioinformatics, 23:2708, 2007.

[10] M. S Aldenderfer and R. K blashfield, cluster analysis. Beverly hills. CA: sage, 1984.

[11] Bell, R. M., Koren, Y., Volinsky, C., 2007. Modeling Relationships a Multiple Scales to Improve Accuracy of Large Recommender Systems. Proc. 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Jose, California USA, p.95-104.

[12] Frey, B. J., Dueck, D., 2006. Mixture Modeling by Affinity Propagation. Neural Information Processing neural information processing system.

[13] Guha, S., Rastogi, R., Shim, K., 2001. CURE: an efficient clustering algorithm for large databases. Inf. Syst., 26(1): 35-58.

[14] J.A. Lozano, J. M. Pena, P. Larranaga, An empirical comparison of four initialization methods for the k-means algorithm Lett. 20 (1999) 1027–1040.

[15] G. W. Milligan, M. C. Cooper, An examination of procedures for determining the number of clusters in a data set, Psychometrika 50 (1985) 159–179.

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

[17] Bagirov, A. M. [Adil M.], Modified global k-means algorithm for minimum sum-of-squares clustering problems, October 2008.

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

[19] Jaiwei Han, Michelinne Kamber, "Data Mining: Concepts and Techniques ", 2001, II Edition

[20] Jason T. L. Wang, Mohammed J. Zaki, Hannu T. T Toivonen and Dennis Shasha (EDS), "Data mining in bioinformatics" Pg. no: 654, Springer International Edition.

****

© 2011, IJMA. All Rights Reserved