IASET: Journal of Computer Science and Engineering (IASET: JCSE)
ISSN(P): Applied; ISSN(E): Applied
Vol. 1, Issue 1, Jan - Jun 2016, 25-30

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

AN UNSUPERVISED CLUSTERING ALGORITHM APPLIED FOR GENE SELECTION MICRO ARRAY DATA

G. BASKAR[1] & P. PONMUTHURAMALINGAM[2]

[1]Research Scholar, Department of Computer Science, Government Arts College (Autonomous), Coimbatore, Tamil Nadu, India

[2]Associate Professor & Head, Department of Computer Science, Government Arts College, Coimbatore, Tamil Nadu, India

ABSTRACT

To explore data-mining techniques as an preset means of reducing the difficulty of data in the large Bioinformatics database and of discovering useful patterns and relationships in data. Data mining isn't an end point, but in one period is an overall data discovery process. It is an iterative process in which preceding processes are modified to support new hypotheses suggested by the data. The stability of feature selection has recently become a topic of strong in both the machine learning and in Bioinformatics community, Feature selection is a term frequently used in data mining for decreasing input to a manageable size for processing and analysis. Micro array data is a commonly used technique for choosing candidate gene in various cancer studies. In this paper, we proposed clustering algorithm on cancer data set, with time, accuracy and memory space.

KEYWORDS: Clustering, Feature Selection, Micro Array, Stability

INTRODUCTION

Data mining increases to clustering the problems of very large data sets with many attributes of different types. Clustering is a partition of data into groups of similar objects, and clustering is the subject of active research in many fields such as statistics, machine learning and bio Bioinformatics. Stability is considered by the scattering of pairwise similarities between clustering obtained from sub samples of the data. The pairwise similarities indicate a stable clustering pattern and a DNA micro array technology has now made it possible to at the same time as monitor the expression levels of thousands of genes throughout the importance of biological processes. Feature selection is a term normally used in data mining for falling inputs to manageable for processing and analysis and it has been a dynamic research area in pattern recognition and data mining communities. Feature selection has been improved the result, and in unsupervised learning, the data mining algorithms are designed to find natural grouping of the examples in the feature space. The feature selection in unsupervised learning has a good subset of features that forms high quality of clusters for a given number of clusters. In this paper, we have implemented the micro array dataset (cancer) with clustering algorithm. The following section will describe the clustering algorithmsused in this paper.

DATA MINING CLUSTERING ALGORITHM

K-Means

One of the simplest Unsupervised learning algorithms by (MacQueen, 1967), A well-known cluster problem is

solved using this algorithm. K-means clustering is a method of classifying/grouping items into k groups (where k is the number of clusters), the grouping is through by minimizing the sum of squared distances (Euclidean distances) between items and the equivalent centroid, and in this the algorithm is implemented using Mat lab. The MATLAB Toolbox has some good functions for performing and interprets k-means clustering analyses.

Algorithm Steps

Step 1: *k* initial "means " (inthis case *k*=2) are randomly generated within the data domain

Step 2: *k* clusters are created by correlating every observation with the nearest mean

Step 3: The centroid of each of the *c*lusters becomes the new mean.

Step 4: Steps 2 and 3 are repeated until convergence.

Fuzzy C-Means

Fuzzy c-means (FCM) is a method of clustering, which permits one piece of data fit to two or more clusters. This method (established by Dunn in 1973 and enhanced by Bezdek in 1981) This algorithm works by passing on membership to each data point and corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the specific cluster center. Clearly, summation of the membership of each data point should be equal.The updated has been done after each iteration membership and cluster centers.

Algorithm Steps

Step 1: Randomly select *'c'* cluster centers.

Step 2: Calculate the fuzzy: membership *'μij'* using

$$\mu ij = 1/\Sigma_{k=1}^{c} (dij/dik)^{(2/m-1)}$$

Step 3: Compute the fuzzy centers *'vj'* using:

$$Vj = \left(\Sigma_{i=1}^{N}(\mu ij)^m xi\right) / \left(\Sigma_{i=1}^{n}(\mu ij)^m\right), \quad j=1,2,....c$$

Step 4: Repeat step 2) and 3) until, the minimum *'J'* value is reached or $||U(k+1) - U(k)|| < \beta$.

where, *'k'* is the iteration step. *'β'* is the termination criterion between [0, 1].

*'U = (μij)n*c'* is the fuzzy membership matrix,*'J'* is the objective function.

Modify Fuzzy C-Means

This MFC has several problems to be implemented. One of the first problems is the number of clusters.

Input: Pattern vector, target vector, K the number of the patterns and the partitions intervals of each attribute..k Pk,i

Output: Centers, membership values and the new projected partitions.

Fuzzy Possibilistic C-Means Clustering

An Fuzzy Possibility C-Mean algorithm could recover the outlier and noise in fuzzy c-mean, but it has been based on Euclidean distance measure, which can only to detect spherical structural clusters and the method be different from the existing clustering methods, the membership values can be taken as degrees of possibility points belonging to the classes and An suitableimpartial function whose minimum will describe a good possibilisticpartition of the data is created, then the membership for the update prototype equations are resulting from essential conditions for minimization of the measure function.

Modified Fuzzy Possibilistic C - Mean Algorithm

FPCM algorithm challenges to partition a finite collection of elements X={x1, x2, x3……xn} into a collection of c fuzzy Clusters with high opinion to some given measure. Given a limited set of data, the algorithm proceeds a list of c cluster centers V, such that V=vi,i=1,2,3……,c And a partition matrix "U" such that U=Uij, i=1,2,3,……c, j=1,2,……n, Where uij is a numerical value in [0, 1] that tells the degree to which the elements xj belongs to the i-th cluster. Defines a family of fuzzy sets {Ai, i=1,2,3……..c} as a fuzzy c partition on a universe of data points x

Algorithm Steps

Step 1: Fuzzy set allows for degree of membership

Step 2: A single points can have partial membership in more than one class.

Step 3: There can be no empty classes and no class that contains no data points

Kernal Based Fuzzy C-Means Clustering Algorithm

The KFCM algorithm adds information to the fuzzy C-means algorithm, it overcomes the drawback of FCM which can't handle the small difference of the cluster. The main aim of fuzzy kernel c-means algorithm is the kernel method which maps nonlinear input data space into a high dimensional feature space.

The following are the iterative steps

Step 1: Set values for C, m, and ε.

Step 2: Initialize the fuzzy algorithm partition matrix

Step 3: Set the loop counter

Step 4: Calculate the C cluster centers using the function

Step 5: Calculate the membership matrix U (b+1)by using function

Step 6: If {U(b)- U(b+1)}< ε then stop, otherwise, set b=b+1 and go to step 4.

Modify Kernel Fuzzy C-Means Algorithm

Although KFCM can be directly applied like FCM are propose to modify the algorithm of KFCM to (MKFCM)

Step 1: To initialize the cluster centers by expectation maximization for optimal choice of the centers

Step 2: To take into account the center of the cluster must consider and modify the centroid calculation

DESCRIPTION OF MICRO ARRAY DATA SETS

In this experiment four frequent micro array data set where taken, the colon cancer data set consist of gene expression profile of 2000 genes for 62 tissues sample which 40 are colon cancer tissues and 22 are normal tissues, leukemia data set it consist of gene expression profile of two classes (i) acute lymphoblastic leukemia (ALL) and acute myeloblastic leukemia(AML) the data set consist of 7,129 genes and 72 samples (47 ALL and 25 AML),in prostate 6034 genes and 102 samples and in lung 12533 genes and 181 tissues samples.

RESULTS OVER THE CLUSTERING ALGORITHM

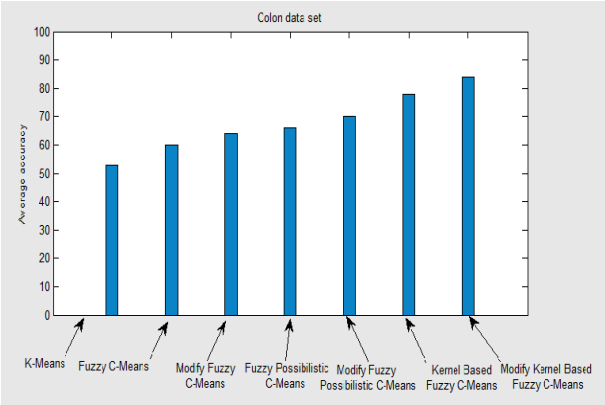

Figure 1: Colon Dataset

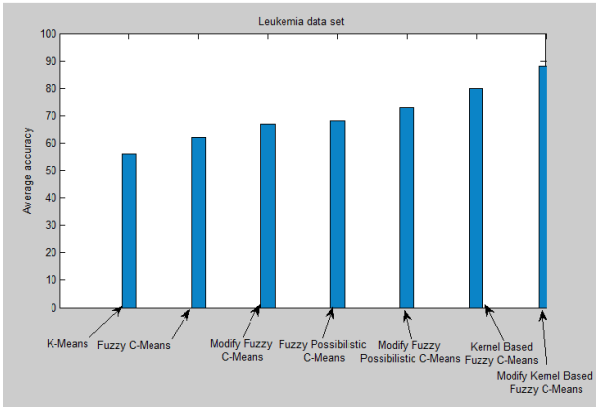

Figure 2: Leukemia Dataset

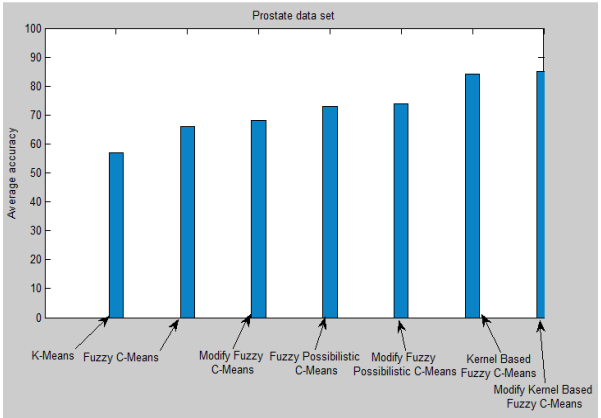

Figure 3: Prostate Dataset

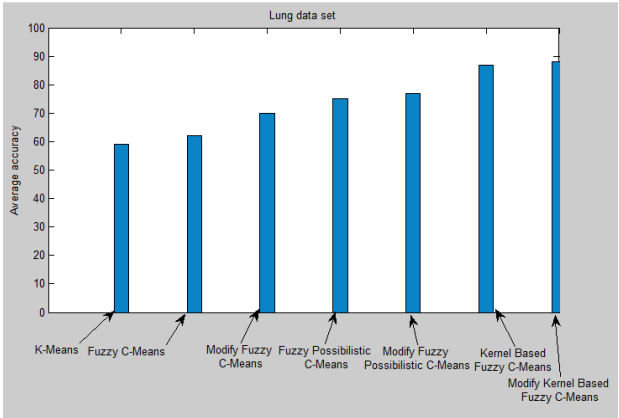

Figure 4: Lung Dataset

CONCLUSIONS AND FUTUR WORK

The analysis of clustering algorithm with micro array dataset are shown in the graph, modify kernel fuzzy c-means algorithm has better accuracy than other clustering algorithm on without feature selection method. In future, it can be improved by the classification algorithm with feature selection method.

REFERENCES

1. Ding, C., Peng, H.: Minimum Redundancy FeatureSelection from Microarray Gene Expression Data. In:Proceedings of the Computational Systems Bioinformatics conference (CSB 2003), pp. 523–529 (2003)
2. Pepe, M.S., Etzioni, R., Feng, Z., et al.: Phases of Biomarker Development for Early Detection of Cancer.J. Natl. Cancer Inst. 93, 1054–1060 (2001)
3. C. A. Davis, F. Gerick, V. Hintermair, et al.,"Reliablegene signatures for microarray classification: assessment ofstability and performance," *Bioinformatics*, vol. 22, pp. 2356– 2363, 2006.
4. 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.
5. Nielsen T.O, West R.B, Linn S.C, et al. Molecular characterization of soft tissue tumours: a gene expression study. Lancet 2002
6. Arun. K. Pujari, "Data Mining Techniques", Universities press (India) Limited 2001, ISBN81- 7371-3804.
7. Bagirov, A.M.[Adil M.], Modified global k-means algorithm for minimum sum-of-squares clustering problems, October 2008, pp. 3192-3199.
8. Bloisi, D.D.[Domenico Daniele], Iocchi, L.[Luca], Rek-Means: A k-Means Based Clustering Algorithm,Springer.
9. Cheung, Y.M.[Yiu-Ming], K*-Means: A new generalized k-means clustering algorithm, PRL(24), No.15, November, 2003.
10. D. Jiang, C. Tang, and A. Zhang, Cluster analysis for gene expression data: a survey, IEEE Transactions on Knowledge and Data Engineering 16 (2004).

AUTHOR'S DEATAILS

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, and India in 2010, and He is currently working towards the PhD degree in Department of Computer Science, Government Arts College, Coimbatore, Tamil Nadu, INDIA in 2011. His area of interest includes Data mining, bioinformatics.

P. Ponmuthuramalingam received his Master's Degree in Computer Science from Alagappa University, Karaikudi in 1988 and the Ph.D. in Computer Science from Bharathiar University, Coimbatore. He is working as Associate Professor and Head in Department of Computer Science, Government Arts College Coimbatore.