International Journal of Engineering Research and Science & Technology

ISSN : 2319-5991
Vol. 2, No. 2
May 2013

www.ijerst.com

Email: editorijerst@gmail.com or editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 2, No. 2, May 2013
© 2013 IJERST. All Rights Reserved

Research Paper

PERFORMANCE ANALYSIS OF DATA CLUSTERING IN RAPID MEDICAL DEVELOPMENT

S Kavi Priya[1]* and M Lingaraj[1]

*Corresponding Author: **S Kavi Priya,** ✉ kavipriya.vignesh@gmail.com

The explosive growth of the amount of publicly available genomic data, a new field of computer science, i.e., bioinformatics has been emerged, focusing on the use of computing systems for efficiently deriving, storing, and analyzing the character strings of genome to help to solve problems in molecular biology. The flood of data from biology, mainly in the form of DNA, RNA and Protein sequences, puts heavy demand on computers and computational scientists. Data mining techniques can be used efficiently to explore hidden pattern underlying in biological data. Un-supervised classification, also known as clustering; which is one of the branches of Data mining can be applied to biological data and this can result in a better era of rapid medical development and drug discovery.

Keywords: Bioinformatics, Genome, DNA, RNA, Clustering Techniques

INTRODUCTION

Comprehensive studies have been carried out on application of clustering techniques to machine learning data and bioinformatics data. Two sets of dataset have been considered in this paper for the simulation study. The first set contains 15 datasets with class-labeled information and second set contains three datasets without class-labeled information. To validate clustering algorithm, for first set of data (i.e., data with class-labeled information), clustering accuracy was used as cluster validation metric while for second set of data (i.e., data without class-labeled information), HS ratio was used as cluster validation metric. This paper explores and evaluates Hard C-means (HCM) and Soft C-means (SCM) algorithm for machine learning data as well as bioinformatics data. Unlike HCM, SCM algorithm provides a presence of a gene in more than one cluster at a time with different degree of membership.

APPLICATION OF DATA MINING

Data mining has become an important area of research since last decade. Important area where Data mining can be effectively applied are as follows: Health sector (Biology/ Bioinformatics), Image processing (Image segmentation), Ad-hoc

[1] Department of Computer Science, VLB Janakiammal College of Arts and Science, Coimbatore-42.

wireless Network (clustering of nodes), Intrusion detection system, Finance sector, etc. In this paper focus has been given on clustering techniques and their application to machine learning and bioinformatics data.

BIOINFORMATICS

Bioinformatics is the application of information technology to the field of molecular biology. The term bioinformatics was coined by Paulien Hogeweg in 1978 for the study of informatics processes in biotic systems. The National Center for Biotechnology Information (NCBI, 2001) defines bioinformatics as "the field of science in which biology, computer science, and information technology merges into a single discipline".

There are three important areas in bioinformatics:

1. The development of new algorithms and statistics with which to assess relationships among members of large data sets;
2. The analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures;
3. And the development and implementation of tools that enable efficient access and management of different types of information. The explosive growth in the amount of biological data demands the use of computing systems for the organization, the maintenance and the analysis of biological data.

GENE EXPRESSION DATA

A microarray experiment typically assesses a large number of DNA sequences (genes, cDNA clones, or expressed sequence tags) under multiple conditions. These conditions may be a time series during a biological process or a collection of different tissue samples.

Application of Data Mining Techniques to Bioinformatics

There are several sub areas in bioinformatics where Data mining can be effectively used for finding useful information from the biological data. Some of the areas are described:

- Data mining in Gene Expression: Gene expression analysis is the use of quantitative mRNA-level measurements of gene expression (the process by which a gene's coded information is converted into the structural and functional units of a cell) in order to characterize biological processes and elucidate the mechanisms of gene transcription.
- Data mining in genomics: Genomics is the study of an organism's genome and deals with the systematic use of genome information to provide new biological knowledge.
- Data Mining in Proteomics: Proteomics is the large-scale study of proteins. Data mining can be used particularly for prediction of protein's structures and functions. As far as this paper is concerned, the main area of focus is application of data mining techniques to gene expression analysis. In next section, microarray technology has been described for gene expression data.

MICROARRAY DATA

DNA microarrays are high-throughput methods for analyzing complex nucleic acid samples. It makes possible to measure rapidly, efficiently and accurately the levels of expression of all genes present in a biological sample. The application of such methods in diverse experimental conditions

generates lots of data. However, the main problem with these data occurs while analyzing it. Derivation of meaningful biological information from raw microarray data is impeded by the complexity and vastness of the data.

CLUSTERING

The clustering problem is defined as the problem of classifying *n* objects into *C* clusters without any apriori knowledge. Let the set of *n* points be represented by the set *S* and the *C* clusters be represented by *V*1, *V*2, ..., *VC*. Then

$Vi \neq 0$ for $i = 1, 2, ..., C;$

$Vi \cap Vj = /0$ for $i = 1, 2, ..., C$ and $i \neq j$

and $U_{i=1}^{C} Vi = S$

Clustering is the process of grouping data objects into a set of disjoint classes, called clusters, so that "objects within the same class have high similarity to each other, while objects in separate classes are more dissimilar". Clustering is an example of un-supervised classification. "Classification" refers to a procedure that assigns data objects to a set of predefined classes. "Un-supervised" means that clustering does not rely on predefined classes and training examples while classifying the data objects.

Hard C-Means Clustering Algorithm

The Hard C-means clustering (HCM) algorithm is one of the best-known squared error based clustering algorithm. It can work very well for compact and hyper-spherical clusters. The time complexity of Hard C-means is '*O*(*n*£*C*£*N*)' and space complexity is '*O*(*n*+*C*)', where *n* is number of data points, *N* is number of feature and *C* is number of cluster in consideration. *C* and *N* are usually much less than *n*, Hard C-means can be used to cluster large data sets in least time.

NOTATIONS

X: Datasets;

X_i : i^{th} Data point of *X*;

x: any data point in *X*;

n: Total Number of data point in *X*;

N: Dimension of each data;

$V_j = j^{th}$ Cluster;

C: Number of cluster

$V_j = j^{th}$ Cluster Center;

$V_j^* = j^{th}$ Cluster Center after updation;

I,j, and *p*: index variable

Input:

Given datasets X; total number of data is *n*; each data is of dimension *N*. Therefore dataset *X*={*X*1, *X*2,.....*Xn*};

Total Number of Cluster is :*C*.

Output:

$V_1, V_2, ..., V_C$

Objective Function:

$J =$

$$\sum_{j-1}^{c} \sum_{xi \varepsilon vj} \left\| x_i - v_j \right\|^2 \quad j \varepsilon 1,2,...,C \text{ and } i = \{1, 2,...,n\}$$

Hard C-means Clustering Algorithm:

Step 1: Chose "C" initial cluster centers(i.e. Prototype vector) *v*1, *v*2,....*vc* either randomly or using any intelligent techniques from the given 'n' data points *X*1, *X*2,....*X*n.

Step 2: Now compute

$\left\| x_i - v_j \right\| < \left\| x_{i-v_p} \right\|$ for $p \in 1,2,...,C$ but $p \neq j$ where

X_i, $i = 1, 2, ..., n$.

If

Then put data X in cluster .

If any ties occurred, then resolved it arbitrarily.

Step 3: Compute new cluster centers $\{v_1^*, v_2^*, ..., v_c^*\}$ as follows:

$$v_j^* = \frac{1}{|v_j|} \sum_{xi \in Vj} xi,\ j = 1, 2, ..., C \text{ where } |v_j|$$

j=1,2,.....C, Where =is the number of elements belonging to cluster .

Step 4: If $v_j^* = v_j$; then terminate. Otherwise repeat from Step 2.

Step 5: If the process does not terminate at Step 4.

The HCM clustering algorithm is one of the most widely used clustering techniques, attempts to solve the clustering problem by optimizing an objective function *J*, which is Mean Square Error (MSE) of formed cluster. The objective function *J* is given as follows:

$$Minimize(J) = \sum_{j=1}^{c} \sum_{x_i \in V_j} \left\| x_i - v_j \right\|^2$$

$$j \in 1, 2, ..., C$$

To validate the feasibility and performance of the HCM clustering algorithm, HCM clustering has been implemented in MATLAB 7.0 (Intel C2D processor, 2.0 GHz, 2 GB RAM) and applied it to Machine learning data as well as to bioinformatics data. Since the datasets used for simulation studies possess class label information, clustering accuracy has been used as cluster validation metric to judge quality of the cluster formation algorithm. Clustering Accuracy is defined as follows:

Clustering Accuracy = (Number *o f Correct Count/ Total number o f instances genes/sample*) × 100

Soft C-means (SCM) Clustering Algorithm

The Soft C-means Algorithm (SCM), which is also known as Fuzzy C-Means algorithm (FCM), generalizes the Hard C-means algorithm, to allow data points to partially belonging to multiple clusters at same time. Therefore, it produces a soft partition for a given data set. In fact, it generates a constrained soft partition.

To achieve this, the objective function of HCM clustering algorithm has to be extended in two ways:

1. Incorporation of degree of fuzzy membership in clusters *fV*1*, V2……VCg* and
2. An introduction of fuzziness parameter *m*, a weight exponent in the fuzzy membership.

The extended objective function *J* (MSE, Mean square error) is defined as follows

$$Minimize(J) = \sum_{j=1}^{c} \sum_{x_i \in V_j} \left(\mu V_j(x_i)\right)^m \left\| x_i - v_j \right\|^2$$

$$\text{for } j \in 1, 2, ..., C$$

where *V* is a fuzzy partition of the data set X formed by clusters *fV*1*, V2……..VCg*. The parameter '*m*' is a weight that determines the degree to which partial members of a cluster affect the clustering result.

Genetic Algorithm Based Clustering

Clustering can be formally formulated as a NP-hard grouping problem in optimization perspective. This research finding has stimulated the search for efficient approximation algorithms,

including not only the use of ad-hoc heuristics for particular classes or instances of problems, but also the use of general-purpose metaheuristics. Particularly, evolutionary algorithms are metaheuristics widely believed to be effective on NP-hard problems, being able to provide near-optimal solutions to such problems in reasonable time. Under this assumption, a large number of evolutionary algorithms for solving clustering problems have been proposed in the literature. These algorithms are based on the optimization of some objective function (i.e., the so-called fitness function) that guides the evolutionary search.

RESULTS AND DISCUSSION

Complete result of Soft C-means clustering for machine learning and bioinformatics data has been given in Appendix D and Appendix E. Appendix D contains the details of data distribution after simulation of SCM clustering algorithm and Appendix E contains the details of data point wrongly clustered in each cluster after simulation of SCM clustering algorithm (Table 1).

Soft C-means clustering for machine learning and bioinformatics data has been given in Appendix D and Appendix E (Figure 1). Appendix D contains the details of data distribution after simulation of SCM clustering algorithm and Appendix E contains the details of data point wrongly clustered in each cluster after simulation of SCM clustering algorithm.

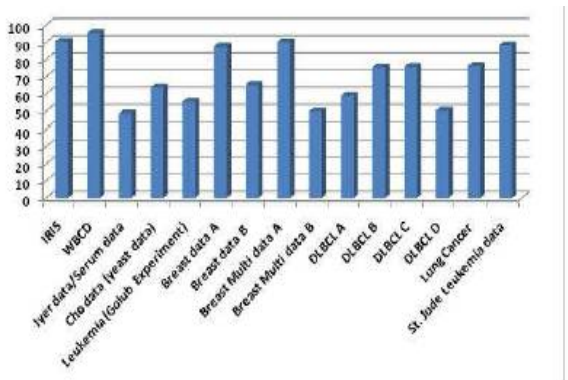
BASIC STEPS INVOLVED IN GA BASED CLUSTERING

The basic steps of GAs (shown in Figure 2); which are also followed in the GA-clustering algorithm,

Table 1: Parameter Used in SCM

S. No.	Datasets	MLU (M)	Maximum Laceration No. Taken for Convergence
1	Iris	4.2	1000
2	WBCD	1.58	20
3	Iyer data/Serum data	1.92	336
4	Cho data(yeast data)	2.15	241
5	Leukemia(Golub Experiment)	1.28	1000
6	Breast data A	1.6	311
7	Breast data B	1.15	194
8	Breast Multi data A	1.3	71
9	Breast Multi data B	1.795	576
10	DLBCLA	1.4	138
11	DLBCL B	1.285	419
12	DLBCL C	1.345	114
13	DLBCL B	1.25	423
14	Lung Cancer	1.07	63
15	St. Jude Leukemia data	1.26	165

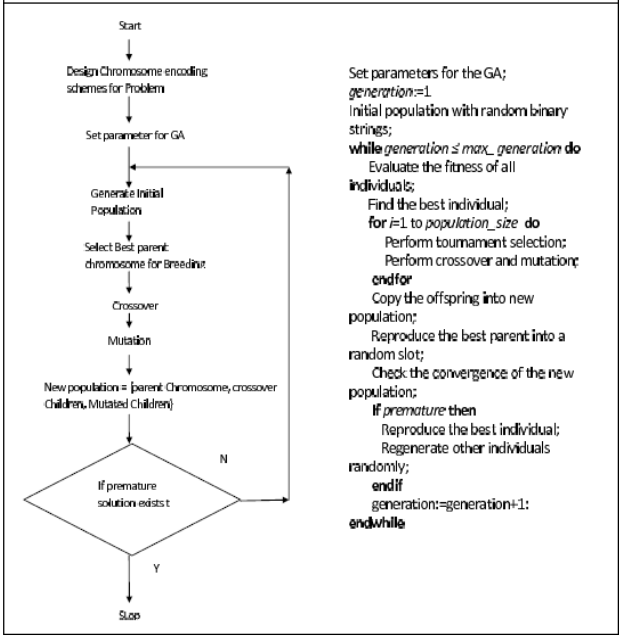
Figure 1: Clustering Accuracy of SCM for Pattern Recognition as Well as Microarray Data



are now described in details in this section. Initially, a random population is created, which represents different data point in a cluster or cluster center in the problem search space. Then fitness value of each chromosome is computed. Based on the principle of survival of the fittest, a few of the

chromosome are selected and each is assigned a number of copies that go into the mating pool. Then crossover operator are applied on these string and that results in set of crossover children. Next, mutation was applied on the crossover children that results set of mutated children, parent chromosome, crossover children and mutated children yields set of chromosome for the new generation. The process of selection, crossover and mutation continues for a fixed number of generation or till a termination condition is satisfied. In this section; discussion has been made on four different representation schemes for encoding of a chromosome.

Figure 2: Flow Chart of Simple GA



BRUTE FORCE BASED CLUSTERING AND SIMULATED ANNEALING BASED PREPROCESSING OF DATA

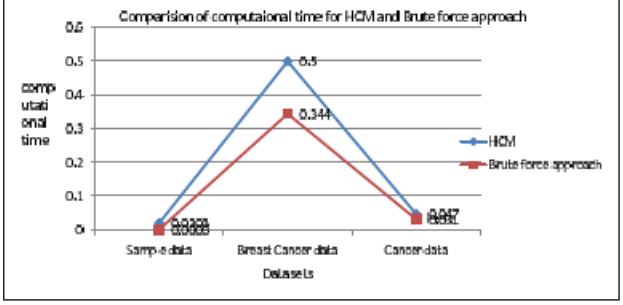
The Brute-Force approach for clustering was intended for three specific purposes. These are as follows:

- It should not require information like no. of clusters from user, i.e., it should help in automatic evolution of clusters.
- It should be computationally faster so that it may be useful for high dimensional data.
- It should be efficient in nature in terms of cluster formation

COMPARISON OF COMPUTATIONAL TIME FOR HCM AND BRUTE-FORCE APPROACH BASED CLUSTERING

Brute Force based clustering approach is computationally quite faster than conventional HCM clustering algorithm and thus it may be very useful while dealing with high dimensional microarray data (Figure 3). Another important feature of the proposed Brute-Force based clustering approach in opposite to HCM clustering algorithm is that, unlike HCM clustering algorithm it does not require the value of number of cluster, i.e., 'C' from user.

Figure 3: Comparison of Computational Time For HCM and Brute-force Approach Based Clustering



CONCLUSION

Extensive study has been carried out on cluster formation algorithms and their application to

machine learning and gene expression microarray data. Microarray data analysis is a novel topic in current biological and medical research, especially when using this technology for cancer diagnosis. Microarray data possess some important characteristic features which machine learning data do not possess. Gene Expression Analysis problem can be formalized as a machine learning data classification (supervised / unsupervised) problem having high-dimension-low-sample data set with lots of noisy/ missing data. Two set of datasets have been considered in this paper for the simulation study. The first set contains 15 datasets with class-labeled information and second set contains three datasets without class-labeled information.

REFERENCES

1. Hruschka E R, Campello R J G B A A F and de Carvalho A C P F (2009), "A Survey of Evolutionary Algorithms for Clustering", *IEEE Transaction on Systems, Man, and Cybernetics Part C: Applications and Reviews*, Vol. 39, No. 2, pp. 133-155.
2. Bhattacharjee A, Richards W G, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark E J, Lander E S, Wong W, Johnson B E, Golub T R, Sugarbaker D J and Meyerson M (2001), "Classification of Human Lung Carcinomas by MRNA Expression Profiling Reveals Distinct Adenocarcinomas Sub-classes", *PNAS*, Vol. 98, pp. 13 790-13 795.
3. Jiang D, Tang C and Zhang A (2004), "Cluster Analysis for Gene Expression Data: A Survey", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 11, pp. 1370-1386.
4. Xu R and Wunsch D (2005), "Survey of Clustering Algorithms", *IEEE Transactions on Neural Networks*, Vol. 16, No. 3, pp. 645-678.
5. Z W *et al*. (2005) "Improved K-means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property", *IEEE Transactions on Nanobioscience,* Vol. 4, No. 3, pp. 255-265.
6. Likas A, Vlassis N and Verbeek J (2003), "The Global k-Means Clustering Algorithm", *Pattern Recognition Letter*, Vol. 36, No. 2, pp. 451-461.
7. Michael L and Mukherjee S (2006), "A Genetic Algorithm Using Hyper-quadtrees for Low Dimensional k-Means Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 28, No. 4, pp. 533-543.
8. Gungor Z and Unler A (2008), "K-harmonic Means Data Clustering with Tabu-Search Method," *Elsevier Applied Mathematical Modelling,* Vol. 32, pp. 1115-1125.
9. Lu Y, Lu S and Fotouhi F (2004), "FGKA: A fast genetic k-means clustering algorithm", *in SAC'04*. Nicosia, Cyprus: ACM, March, ISBN:1-58113-812-1/03/04.
10. Lu Y, Lu S, Fotouhi F, Deng Y, Susan D and Brown J (2004), "An Incremental Genetic K-means Algorithm and Its Application in Gene Expression Data Analysis," *BMC Bioinformatics*.
11. Hamerly G and Elkan C (2003), "Learning the k in k-Means," in Proceedings of the 17th Annual Conference on Neural Information

Processing Systems (NIPS), December, pp. 281-288.

12. Ordonez C and Omiecinski E (2004), "Efficient Disk-based K-means Clustering for Relational Databases," *IEEE Transactions on Knowledge and Data Engineering,* Vol. 16, No. 8, pp. 909-921.

13. Wagstaff K (2002), "Intelligent Clustering with Instance-level Constraints," Ph.D. Dissertation, Cornell University.

14. Wang W, Wang C, Cui X and Wang A (2008), "A Clustering Algorithm Combine the FCM Algorithm With Supervised Learning Normal Mixture Model," in *The 19th IEEE International Conference on pattern Recognition (ICPR 08)*, December 2008, pp. 1-4.

15. Rayward-Smith V J (2005), "Metaheuristics for Clustering in KDD," in *Proc. IEEE Congress on Evolutionary Computation*, pp. 2380-2387.

International Journal of Engineering Research and Science & Technology
Hyderabad, INDIA. Ph: +91-09441351700, 09059645577
E-mail: editorijlerst@gmail.com or editor@ijerst.com
Website: www.ijerst.com