ISSN : 2231-0711

Clustering Gene Expression Datasets by Combinatorial Fusion Analysis

Karimulla.SK[1] ,Srinivasulu.P[1],Dr. Srinivasa Rao.V[1] ,Mastanvali Shaik[2]

[1] *Department of Computer Science and Engineering, V R Siddhartha Engineering College, Vijayawada, A.P-520007.*
[2]*Senior Software Engineer, SPAN InfoTech,Bangalore,Karnataka-560004*

***Abstract*-Genomic data that are available in large volumes will define the need for new data analysis techniques and tools. Gene expressions data sets analysis through clustering is a combinatorial optimization problem. Combining with the conception of Harmony search in Optimization theory on combinatrics and with genomic halving technique in game theory, a new algorithm is proposed for clustering the Gene expressions datasets by Combinatorial Fusion Analysis mechanism. Meta heuristic algorithms are one of the best optimization techniques that are useful for finding optimal or near optimal solutions for combinatorial optimization problems whose search space is very huge. The inter cluster gene expressions datasets optimized by harmony search to overcome the problem of local optima , the dynamic memory improvisation in initial cluster assignment and the cluster size, the redundancy over the genomic data will be cleared out by applying by the whole genome duplication events in molecular evolution with the genomic halving technique. In this paper, a new algorithm and a framework are proposed to analyze the gene-expressions in various situations with the help of graph theory technique and Optimization Theory.**

Keywords

Data mining, Combinatorial Optimization Problem, Genomic halving technique, Clustering Algorithms, Meta Heuristics, Information Retrieval.

1. INTRODUCTION

Clustering Gene expression analysis is a technique that allows the identification of groups of similar gene sequences in multi-dimensional space. It was related to *the* field of *Information Retrieval* (IR) [15] as a means of improving the efficiency of serial search. The *Cluster Hypothesis* is fundamental to the issue of improved effectiveness; it states that relevant gene expressions tend to be more similar to each other than to non-relevant ones, and therefore tend to appear in the same clusters. Various tests have been used to quantify the degree of similarity at which genes relevant to each other remain to the hypothesis [17]. If the Cluster Hypothesis holds for a particular gene expression collection, then relevant expressions will be well separated from non-relevant ones, and hence a cluster-based search strategy will likely be effective. The drawbacks with the conventional clustering methodologies [6] [9] are 1) Prior knowledge about cluster size is required. 2) Number of clusters should be known before. 3) Problem of local and global optima. For clustering, two measures of cluster quality are used. One type of measure compare different sets of clusters without reference to external knowledge and is called an *internal quality* measure we will use a measure of "overall similarity" based on the *pair wise similarity* of Gene expression datasets in a cluster. The other type of measures lets us evaluate how well the clustering is working by comparing the groups produced by the clustering techniques to known classes. This type of measure is called an *external quality* measure .one of the external quality external quality measure is the f-measure, which, as we use it here, is more oriented toward measuring the effectiveness of a hierarchical clustering. There are many different quality measures and the performance and relative ranking of different clustering algorithms can vary substantially depending on which measure is used. However, if one clustering algorithm performs better than other clustering algorithms on many of these measures. As we shall see in the results section, the proposed clustering algorithm has the best performance for the other quality measures algorithms. *Harmony search algorithm* had been very successful in a wide variety of optimization problems like *Data clustering;* presenting several advantages with respect to traditional optimization techniques such as the following (a) HS algorithm imposes fewer mathematical requirements and does not require initial value settings of the decision variables. (b) As the HS algorithm uses stochastic random searches, derivative information is also unnecessary. (c) The HS algorithm generates a new vector, after considering all of the existing vectors. These features increase the flexibility of the HS algorithm and produce better solutions.

2. PRELIMINARIES

2.1. Distance Measurement

Objective function clustering is dependent on the definition of similarity and dissimilarity according to distance measurement, and the data set is divided into subsets according to their similarities and dissimilarities. The objects in closer vicinity, which means the higher similarities they hold, could be grouped into a cluster according to this measurement.

The Biweight Midcorrelation

Let $X=\{x_1,x_2,..x_i,..x_n\}$ and $Y=\{y1,y2,..yj,..yn\}$ be a gene expression datasets sample from a genomic database distribution. The biweight midcorrelation can be defined as fallows:

$$r_b = \frac{s_{bxy}}{\sqrt{s_{bxx} s_{byy}}} \quad \ldots. (1)$$

Where s (bxy) is the biweight midcovariance between X and Y and s (bxx) and s (byy) are the biweight midvariances of X and Y respectively.

2.2. Clustering as an optimization problem

In Cluster analysis we assume that we have been given a finite set of points X in the n-dimensional space R^n, $X=\{x^1,x^2,x^3,\ldots,x^m\}$,where $x^i \in R^n$,i=1,2,3…m.Clustering the datasets is a NP-hard Generalized Orienteering Problem (GOP).So that to solve the k-clustering problem. Here we defined a set of constraints in the form of Meta heuristics and used harmony search to resolve those heuristics.

$[X]=[x_1\ x_2\ \ldots\ x_n]$
$[P]=[p_1\ p_2\ \ldots\ p_n]$
$[R]=[r_1\ r_2\ \ldots\ r_n]$

Where p_i=precision $(x_i) \geq 0$,
and i=1, 2, 3,…,n.

$$\sum_{i=1\ldots n} p_i = 1. \qquad (2)$$

2.3.*Objective function*: *Weighted F-measure*

The general formula for non-negative real **α** is:

$F_\alpha(r, p)=(\alpha+1)rp/r+\alpha p$ (3)

Where r is recall

P is precision

α is non-negative constant.

Where, r have a weight of **α** *Є (0;+∞)*

and p have a weight of 1

We can calculate the recall and precision of that cluster for each given class.

More specifically, for cluster *j* and class *i,*

Recall is the fraction of the gene expression datasets that are relevant to the query that are successfully retrieved.

Recall(i, j) = *nij / ni*

Precision is the fraction of the gene expression datasets retrieved that are relevant to the user's information need.

Precision (i, j) = *nij / nj*

Where *nij* is the number of gene expressions of class *i* in cluster *j*,

nj is the number of gene expressions of cluster *j*

and *ni* is the number of gene expressions of class *i*.

For an entire hierarchical clustering the F measure of any class is the maximum value it attains at any node in the tree and an overall value for the F measure is computed by taking the weighted average of all values for the F measure as given by the following.

$F=\sum^{n}_{i=1}$ (ni/n) Max {F (i, j)} (4)

Where the max is taken over all clusters at all levels, and *n* is the number of gene expression datasets.

2.4. Genomic halving technique

Cycle Decomposition Problem. For a given duplicated genome P, find a perfect duplicated genome R©R maximizing cmax (G0 (P; R©R)). In order to solve the Cycle Decomposition Problem for a genome P, we will construct a Contracted breakpoint graph G' (P; ©) which achieves the upper bound. The genome P alone defines a vertex set of the graph G', an obverse matching, and black cycles in G0 so that G0 is black-obverse connected. A BO-graph [12] is a connected graph with black and obverse edges such that the black edges form black cycles and the obverse edges form an obverse matching (every duplicated genome P corresponds to a BO-graph).

3. A FRAME WORK FOR GENE EXPRESSION ANALYSIS BY CLUSTERING (THE DATA COMPRESSION APPROACH)

The goal of proposed framework is to compress the genomic data sets by finding relative gene datasets among them. The minimum description length (MDL) principle can be used to select among different expressions encoding.

Minimum description length principle: Use genomic halving distance as MDL in order to select the relative gene datasets to regularize the genomic database so as to rearrange them in a confined structure.

3.1. Description: The different phases of proposed *Framework* are depicted in the fig.1.

3.1.1.Phase1: All bio-medical data will be extracted from http://www.bioinformatics.org/pcgdb/. The data whatever we are considering as input should be not clustered and not properly arranged.

3.1.2.Phase2: Apply Genome Halving Problem [12] to rearrange the genomic data by avoiding replication among datasets (Apply MDL principle).

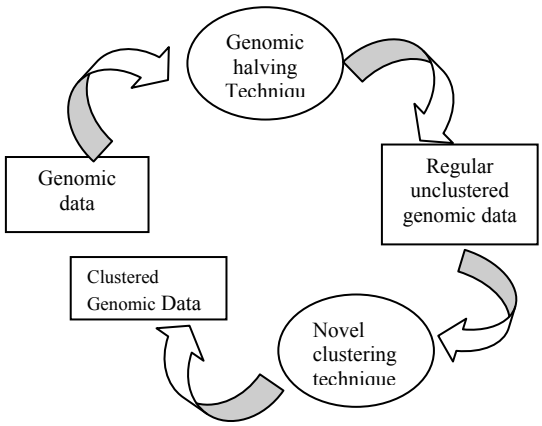

Fig.1 phases of proposed clustering framework

Genome Halving Problem: Given a duplicated genome P, recover an ancestral pre-duplicated genome R minimizing the reversal distance from the perfect duplicated genome R © R to the duplicated genome P.

3.1.3. Phase3: Apply the fallowing proposed algorithm on datasets to cluster the data. (A modified version to the clustering Algorithm)

MAXIMUM WEIGHTED F-MEASURE CLUSTERING ALGORITHM

Step1: Initialize the problem and algorithm parameters. (Cluster initialization)

The optimization problem is specified as follows:

Minimize f(x),Subject to xi Є Xi=1,2,..,N (5)

Where *f(x)* is an objective function; **x** is the set of each decision variable x_i; *N* is the number of decision variables, Set a random value for K (Update dynamically by number of improvisations),Set winning parameter η for deciding stopping criteria.

Step2: Distance computation

Objective function clustering is dependent on the definition of similarity and dissimilarity according to distance measurement, and the data set is divided into subsets according to their similarities and dissimilarities. The objects in closer vicinity [14], which means the higher similarities they hold, could be grouped into a cluster according to this measurement.

The Biweight Midcorrelation *:* Let $X=\{x_1,x_2,..x_i,..x_n\}$ and $Y=\{y_1,y_2,..y_j,..y_n\}$ be a gene expression datasets sample from a genomic database distribution[11].The biweight midcorrelation can be defined

$$r_b = \frac{a_{bcg}}{\sqrt{a_{bac}\, a_{bgg}}}$$

Step3: Initial cluster Assignment (Initialize the harmony memory)

In this Step, the HM (Harmonic Memory) matrix is filled with as many randomly generated solution vectors as the HMS

$$HM = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_{N-1}^1 & x_N^1 \\ x_1^2 & x_2^2 & \cdots & x_{N-1}^2 & x_N^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^{HMS-1} & x_2^{HMS-1} & \cdots & x_{N-1}^{HMS-1} & x_N^{HMS-1} \\ x_1^{HMS} & x_2^{HMS} & \cdots & x_{N-1}^{HMS} & x_N^{HMS} \end{bmatrix}.$$

Step4: Cluster center computation (Improvise a new harmony)

A new harmony vector, $x' = (x'_1, x'_2, \ldots, x'_N)$, is generated based on three rules:

(1) Memory consideration, (2) Object Function and (3) random selection.

Generating a new harmony is called 'improvisation'. In the memory consideration, the value of the first decision variable (x'_1) for the new vector is chosen from any of the values in the specified HM range $(x'_1 - x_1^{new})$. Values of the other decision variables $(x'_2, x'_3, \ldots, x'_N)$ are chosen in the same manner.

Step5: Cluster Assignment (Update harmony memory)

If the new harmony vector, $x' = (x'_1, x'_2, \ldots, x'_N)$ is better than the worst harmony in the HM, judged in terms of the objective function value, the new harmony is included in the HM and the existing worst harmony is excluded from the HM.

Step6: Check stopping criterion

Check the Winning parameter value (η).

If the stopping criterion (η=maximum number of improvisations) is satisfied, computation is terminated. Otherwise, Steps 3 and 4 are repeated.

3.1.4. Phase4: Evolve the output datasets.

4. EMPERICAL DEMOSTRATION OF MAX.WEIGHTED F-MEASURE ALGORITHM

We hypothesize that ***weighted f-measure*** evaluates the inter cluster distance efficiently and pervasively, even in unbalanced large gene expressions datasets. We demonstrate it by a detailed analysis of existing similarity measures. We maximize our heuristics' performance [13] using *dynamic updated thresholds* and *Winning parameter*. The threshold is the minimum level of certainty that must be reflected by the consensus of survey updations.

5. EXPERIMENTAL EVALUATION OF FRAMEWORK

To illustrate the effectiveness of our proposed algorithm, we compare the empirical results between our algorithm and the conventional clustering algorithms like Modified Global K-means algorithm [1][3][5], Fuzzy C-means (FCM) algorithm [8] and maximum weight entropy clustering algorithm [9].Our algorithm doesn't need the prior knowledge of inter cluster distance and cluster size, more over it doesn't required any static value initialization for number of clusters. The comparative results are furnished by applying all these techniques on http://www.bioinformatics.org/pcgdb/ datasets.

6. CONCLUSION AND FUTUREWORK

We proposed a new unsupervised clustering method based on Optimization theory on combinatrics, the weighted F-measure as the objective function to harmony search, which considers the genomic datasets as a N-dimensional space and evaluates the inter cluster distance parameters efficiently. The empirical evaluation of the objective function illustrates the efficiency of the algorithm in detecting clusters of different kinds. The algorithm doesn't need the prior knowledge about the cluster number and the initialization and updation of cluster center. The proposed optimization framework provides a systematic and simple way of analyzing gene expressions. Genomic halving technique is applied on all genomic datasets before applying the clustering mechanism, yields good results computationally. As a part of future work, the gene expression analysis via clustering as an optimization problem, one can go for tree mining techniques or information retrieval measures for graph clustering instead of genomic halving problem as the genomic halving technique requires more mathematical background.

7. ACKNOWLEDGEMENT

This research was supported by grant from Department of Computer Science and Engineering in V R Siddhartha Engineering College, Vijayawada.

REFERENCES

[1] Adil M. Bagirov Karim Mardaneh, Modified global k-means algorithm for clustering in gene expression data sets, Australian Computer Society, Inc

[2] Anastasios Tombros, C.J. van Rijsbergen, Query-Sensitive Similarity Measures for the Calculation of Interdocument Relationships, 2001 ACM I-581 l3-436-3/Ol/OOl

[3] Aristidis Likasa, Nikos Vlassisb, JakobJ. Verbeekb,The global k-means clustering algorithm, 2002 Pattern Recognition Society. Elsevier Science Ltd. 0031-3203/02.

[4] Carlos Ordonez, Edward Omiecinski,FREM: Fast and Robust EM Clustering for Large Data Sets, 2002 ACM 1-58113-492-4/02/0011

[5] David Arthur ,Sergei Vassilvitskii, k-means++: The Advantages of Careful Seeding,

[6] Gao Cong, Kian-Lee Tan Anthony K.H.Tung, Xin Xu, Mining Top-k Covering Rule Groups for Gene Expression Data, 2005 ACM 1-59593-060-4/05/06

[7] Greg Hamerly, Charles Elkan, Alternatives to the k-means algorithm that find better clusterings, 2002 ACM 1-58113-492

[8] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981.

[9] Li Lao, Xiaoming Wu, Lingpeng Cheng and Xuefeng Zhu, Maximum Weighted Entropy Clustering Algorithm

[10] M. Granzow, D. Berrar, W. DubitzkyA. Schuster,A.Schuster@ulst.ac.uk, F.J.Aguaje,Francisco.Azuaje@cs.tcd.ie,R. Eils, Tumor Classification by Gene Expression Profiling:Comparison and Validation of Five Clustering Methods

[11] Marla D. Curran , Hong Liu , Fan Long , Nanxiang Ge ,Statistical Methods for Joint Data Mining of Gene Expression and DNA Sequence Database, SIGKDD Explorations

[12] Max A. Alekseyev and Pavel A. Pevzner,Colored de Bruijn Graphs and the Genome Halving Problem, 2007 IEEE1545-5963/07.

[13] Mihail Popescu, James M. Keller, and Joyce A. Mitchell, Fuzzy Measures on the Gene Ontology for Gene Product Similarity, 2006 IEEE1545-5963/06

[14] Silvia Selinski and Katja Ickstadt, Similarity Measures for Clustering SNP Data.

[15] Van Rijsbergen, C. J. 1979. Information Retrieval. London, U.K.: Butterworths

[16] Weiss, Sholom M. & Casimir A. Kulikowski. 1991. Computer Systems that Learn:Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. San Francisco, Calif.: Morgan Kaufmann

[17] Wolfgang Huber, Anja von Heydebreck, Martin Vingron,Analysis of microarray gene expression data, April 2, 2003.