USE OF POINT SYMMETRY BASED DISTANCE FOR GENE EXPRESSION DATA CLUSTERING

Sriparna Saha, Asif Ekbal, Neha Vinayak
Department of Computer Science and Engineering
Indian Institute of Technology Patna
Patna-800013 India
Tel: +91-8809559190; fax: +91-612-2277383
e-mail: {sriparna,asif}@iitp.ac.in

ABSTRACT

Introduction of microarray technology helps to study the expression profiles of thousands of genes across different experimental conditions or tissue samples simultaneously. Clustering techniques have been widely used for analyzing such microarray data, typical properties of which are its inherent uncertainty, noise and imprecision. In this paper we have developed some unsupervised approaches for clustering of tissue samples. In recent years some symmetry based clustering techniques have been developed. These clustering algorithms try to optimize total symmetry within a given partitioning. We have used these point symmetry based clustering techniques as the underlying unsupervised approach for gene expression data clustering. Here for grouping of different genes point symmetry based distance is used. The performance of the symmetry based clustering method is compared with that of several other clustering algorithms for some publicly available benchmark gene expression datasets. Biological significance tests have been conducted to analyze the biological relevance of the clustering solutions.

1 INTRODUCTION

Clustering [1,2,4] is an unsupervised classification technique. It has many applications in data mining, which helps to group objects, such that objects in the same group/cluster are similar to each other with respect to some criteria and objects in different clusters are different from each other with respect to the same criteria. Clustering [1,2,4] is performed based on an objective function which can be either minimized or maximized, depending on the algorithmic requirements. Many types of clustering algorithms have been developed, which can be broadly grouped as partitional, hierarchical and graph theoretic methods. Examples of these are K-means, single linkage and minimum spanning tree-based algorithms [2].

For partitioning a data set, at first some measure of similarity or proximity has to be defined based on which cluster assignments can be done. The measure of similarity is generally data dependent. In general, one of the fundamental features of shapes and objects is symmetry. This is considered to be important for enhancing the recognition of different objects [1]. As the concept of symmetry is common in the natural world, several researchers have utilized this property while clustering a data set. In the real world there are many objects which contain some form of symmetry; the human face, jellyfish, the human body, stars, etc. are some examples of symmetry. Symmetry mainly conveys balance and reflects perfection or beauty. As symmetry represents the well-defined concept of balance or "pattern self-similarity", it has been extensively used to describe many processes/objects in geometry and physics. Thus, we can assume that symmetry would be a desirable property of good clusters and that it should therefore be included as an objective for the clustering algorithm to pursue. Based on this concept, several different symmetry-based similarity measures/distances have been proposed in the literature [1].

A point symmetry based distance measure was proposed in [1] denoted as $d_{ps}(\mathbf{X},\mathbf{C})$, where **X** is the point and **C** is the centroid. The definition of d_{ps} is as follows: let a point be **X**. The symmetrical (reflected) point of **X** with respect to a particular centroid **C** is 2***C*X**. Let us denote this by **X'**. Let the first and the second unique nearest neighbors of **X'** be at Euclidean distances of d1 and d2, respectively. Then $d_{ps}(\mathbf{X},\mathbf{C})= \frac{(d1+d2)}{2}$ x $d_e(\mathbf{X},\mathbf{C})$, where $d_e(\mathbf{X},\mathbf{C})$ is the Euclidean distance between the point **X** and **C**.

The major characteristics of this distance is that in d_{ps}, two nearest neighbours are taken into consideration. So, the term $(d1 + d2)/2$ will never be equal to 0 and hence, the effect of the Euclidean distance d_e will always be taken into account. Also, considering only one nearest neighbour may be misleading in some cases, whereas on taking into account two nearest neighbours, if both d1 and d2 of a point **X** with respect to **C** are less, then the likelihood that **X** is symmetrical with respect to **C** increases.

Gene Expression Data Clustering -

Gene is the fundamental unit of storage of hereditary information in living beings [3][5]. Technically, it can be viewed as a distinct sequence of nucleotides forming part of a chromosome. Information from a gene is used in the synthesis of functional gene products like proteins and functional RNAs for non-protein coding genes. This process of synthesis is called Gene Expression, by which genotype gives rise to phenotype.

This Gene Expression Data is generally very huge in size and to search for useful patterns within this data, genes have to be grouped into "clusters" on the basis of similar features. The Gene Expression data is in the form of a 2-Dimensional matrix of Gene Names and the corresponding expression levels for features exhibited by the genes.

Clustering of Gene Expression data has been done by various algorithms. Here we have analysed the performance of a symmetry based genetic clustering technique, GAPS, [1] with respect to Gene Expression data. We have also compared the performance of GAPS with respect to two popular clustering algorithms, e.g., GAK-means algorithm [4], average linkage clustering technique [2]. Results on five gene expression data sets including yeast sporulation, yeast cell cycle, rat CNS, human fibroblasts serum, Arabidopsis Thaliana [3] show the superior performance of the GAPS clustering technique. Clustering results are validated using an internal cluster validity index named Silhouette index [6]. Experimental results show the efficacy of GAPS over other well-known clustering algorithms in finding clusters of co-expressed genes efficiently. We have also carried out biological significance tests to check the biological relevance of the obtained clusters, i.e., consist of genes which belong to the same functional group. Results reveal that GAPS can be effectively used to identify co-expressed genes from gene expression data sets.

2. PROPOSED APPROACH OF GENE EXPRESSION DATA CLUSTERING

In this paper, we have applied the GAPS algorithm [1] on gene expression data, for the readily available datasets (Refer section 3) and analysed the performance of GAPS relative to other single objective clustering algorithms – GAK-Means [4] and Average Linkage [2]. The Biological Significance of GAPS has also been established, as compared to the above mentioned algorithms.

The GAPS algorithm uses a genetic algorithm based approach for clustering, when the value of K (No. of Clusters) is known. GAPS uses the above defined Point Symmetry distance measure d_{ps} [1] instead of the Euclidean distance to determine a clustering metric, M. The objective of the algorithm is to find the cluster centroids such that M is maximized.

The main steps of the GAPS algorithm are as follows:

String Representation and Population Initialization: Each chromosome in the population is represented by a string of K cluster centroids, which are initialized to K randomly chosen points from the dataset. Then after executing five iterations of K-means on each of the chromosomes, the cluster centroids are replaced by the result of K-means algorithm.

Fitness Computation: If the total symmetricity (**(d1+d2/2)** is less than a given threshold value (which is set depending on the data set; here we have used 0.6 for all the data sets), assignment of points to different clusters are done based on the point symmetry distance, otherwise Euclidean distance measure is used for assignment. The cluster centroids are then updated to the mean points of the respective clusters. Subsequently, the clustering metric, M is calculated for each chromosome, as

```
M=0
For k = 1 to K do
For all data points $\mathbf{Xi}$, i=1 to n and $\mathbf{Xi} \in$ kth cluster do
        $M = M + d_{ps}(\mathbf{Xi,Ck})$
```

The fitness function fit is defined as fit = 1/M. The function will be maximised by using GA.

Selection: Roulette wheel selection has been implemented.

Crossover: Single point crossover has been used. The crossover probability μ_c of each chromosome is such that when the better of the two chromosomes to be crossed is itself quite poor, μ_c is increased and when it is a good solution, μ_c is decreased.

Mutation: Each chromosome undergoes mutation with a probability $\mu_{m.}$ Like $\mu_{c,}$ μ_m will also get lower values for high fitness solutions and higher values for low fitness solutions.

In GAPS the processes of fitness computation, crossover, mutation, selection are executed for a maximum number of generations. The best string seen upto the last generation provides the solution to the clustering problem. Elitism has been implemented at each generation by preserving the best string seen up to a generation in a location outside the population. Thus, on termination, this location contains the centers of the final clusters. According to these center combinations we have to assign cluster labels to each point using the point symmetry based distance.

3. DATA SETS USED

In this paper we have used five gene expression data sets. These pre-processed datasets have been downloaded from the site mentioned in [3] (http://anirbanmukhopadhyay.50webs.com/mogasvm.html). A short description of the data sets is provided in Table 1. The description of these data sets are already available in [3] but we have included those here for the sake of completeness.

a. Yeast Sporulation

This data set consists of 6118 genes measured across 7 time points (0, 0.5, 2, 5, 7, 9 and 11.5 hours) during the sporulation process of budding yeast. The data are then log-transformed. The Sporulation data set is publicly available at the website http://cmgm.stanford.edu/pbrown/sporulation.

Table 1: *Details of pre-processed datasets used.*

Data Set	No. Of Genes in pre-processed dataset	No. of Features
a. Yeast Sporulation:	474	7
b. Yeast Cell Cycle:	384	17
c. Rat Central Nervous System (CNS):	112	9
d. Human Fibroblasts Serum:	517	13
e. Arabidopsis Thaliana:	138	8

Among the 6118 genes, the genes whose expression levels did not change significantly during the harvesting have been ignored from further analysis. This is determined with a threshold level of 1.6 for the root mean squares of the log2-transformed ratios. The resulting set consists of 474 genes.

b. Yeast Cell Cycle

The yeast cell cycle dataset was extracted from a dataset that shows the fluctuation of expression levels of approximately 6000 genes over two cell cycles (17 time points). Out of these 6000 genes, 384 genes have been selected to be cell-cycle regulated. This data set is publicly available at the following website: http://faculty.washington.edu/kayee/cluster.

c. Arabidopsis Thaliana

This data set consists of expression levels of 138 genes of Arabidopsis Thaliana. It contains expression levels of the genes over 8 time points viz., 15 min, 30 min, 60 min, 90 min, 3 hours, 6 hours, 9 hours, and 24 hours. It is available at http://homes.esat.kuleuven.be/_thijs/Work/Clustering.html.

d. Human Fibroblasts Serum

This dataset contains the expression levels of 8613 human genes. The data set has 13 dimensions corresponding to 12 time points (0, 0.25, 0.5, 1, 2, 4, 6, 8, 12, 16, 20 and 24 hours) and one unsynchronized sample. A subset of 517 genes whose expression levels changed substantially across the time points have been chosen. The data is then log2-transformed. This data set can be downloaded from http://www.sciencemag.org/feature/data/984559.shl.

e. Rat CNS

The Rat CNS data set has been obtained by reverse transcription-coupled PCR to examine the expression levels of a set of 112 genes during rat central nervous system development over 9 time points. This data set is available at http://faculty.washington.edu/kayee/cluster.

All the data sets are normalized so that each row has mean 0 and variance 1.

4. PERFORMANCE MEASUREMENT METRICS

For evaluating the performance of clusters, Silhouette Index has been used as the metric..

Silhouette Index: Silhouette Index [6] is used as a cluster validity index, used to compare the quality of the clusters formed by the clustering algorithm. It gives a view of the compactness and separation of clusters. The value of silhouette index ranges between -1 to 1 and a good cluster will have a higher value of silhouette index.

Input parameters: The GAPS algorithm has been executed with a population size of 100 for 30 generations. As the no. of clusters is required to be provided as input in the algorithm the cluster size, selected as per [3] are listed in Table 2.

Table 2: Number of clusters used as input for different datasets

Dataset	**No. Of Clusters**
Yeast Sporulation	6
Yeast Cell Cycle	5
Rat CNS	6
Human Fibroblasts Serum	6
Arabidopsis Thaliana	4

RESULTS

Biological Significance testing was done for various runs of the Yeast Sporulation dataset, from the site (http://www.yeastgenome.org/cgi bin/GO/goTermFinder.pl).

It results in statistically significant Gene Ontology (GO) terms used to describe the genes in the list. Genes are considered to be statistically significant if the p-value < 0.01, i.e 1% Significance Level. This test has been carried out for three different Gene Ontologies, namely – Biological Processes, Molecular Functions and Biological Components. Out of these combined results, the three GO terms having the least p-values have been selected. Results show that GAPS attains minimum p-values for each cluster as compared to two other clustering techniques, GAK-Means and Average Linkage algorithms. For example the GO terms and the p-values attained by GAPS clustering technique for cluster 1 are :

cytoplasmic translation - GO: 2181

cytosolic ribosome - GO: 22626

structural constituent of ribosome - GO: 3735

A boxplot for the p-values has been drawn to compare them in **Figure 1**. The p-values have been converted to $\log_{10}$ for better visualization and ease of comparison (i.e new p-value = $-\log_{10}$(p-value)). Only clusters resulting in at least one significant GO term have been considered for this test. Clusters with lower p-values or higher $-\log_{10}$(p-value) values are considered to be better.

Table 3 gives the detailed $-\log_{10}$(p-value) values of each stage of the boxplot (Min, Lower Quartile, Median, Upper Quartile, Max) for all the three algorithms. It can be observed from this table that the Median value for GAPS is better than that of GAK-Means and Average Linkage. Hence, it can be established that GAPS produces significant and biologically relevant clusters.

Table 3: *Numerical Values for each step of the boxplots comparing GAPS, GAK-Means and Average Linkage algorithms, establishing that GAPS produces biologically significant clusters which are functionally enriched*

	GAPS	**GAK-Means**	**Average Linkage**
Minimum	3.79588	3.568636	3.68987
Lower Quartile	11.65956	13.289037	14.84968
Median	27.44782	26.976456	27.02503
Upper Quartile	35.35655	35.237321	36.2214
Maximum	60.96257	58.378824	47.70774

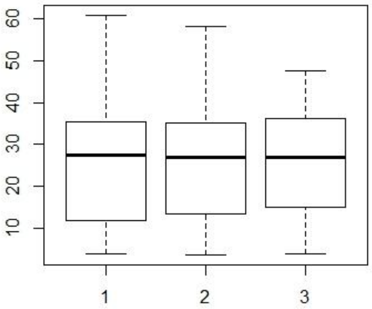

Figure 1: Boxplot for the values indicated in Table 2; Here 1: GAPS, 2: GAK-means, 3: Average Linkage

Next, the value of Silhouette Index has been calculated for each of the datasets for 10 runs of GAPS with different combinations of input parameters. The best results have been listed below in **Table 4**.

It can be clearly seen that GAPS gives better results than GAK-Means and Average Linkage for most of the Datasets. For Yeast Sporulation dataset, although the maximum value of Silhouette index observed is 0.6424 (Result-1), but the Silhouette index value for the most biologically significant result has been found to be 0.6310 (Result-2). Also, it was found that no gene was placed in one of the clusters in Result-1.

***Table 4**: Comparison of Algorithms based on Silhouette Index*

Data Sets	GAPS	GAK-means	Average Linkage
Sporulation (K=6)	0.6424	0.5681	0.6366
Cell Cycle (K=5)	0.4393	0.3661	0.3938
Arabidopsis(K=4)	0.3595	0.34	-0.1792
Serum (K=6)	0.3506	0.3467	0.2898
Rat CNS (K=-6)	0.411	0.3442	0.3075

5. DISCUSSIONS AND CONCLUSION

In this paper, we have compared the performance of GAPS with GAK-Means and Average Linkage for gene expression data clustering and concluded that the point symmetry based GAPS algorithm gives better performance than the other algorithms. The performance comparison has been done on the basis of Silhouette Index values.

We have also established that GAPS gives biologically significant clusters by finding out the most significant Gene Ontology (GO) terms for each cluster and plotting their p-values (at 1% Significance Level) with respect to the other two algorithms. GAPS assumes number of clusters apriori.

In future we would like to apply some automatic clustering techniques for gene expression data clustering which can automatically determine appropriate number of clusters and appropriate partitioning.

6. REFERENCES

[1]. Sanghamitra Bandyopadhyay, Sriparna Saha - GAPS: A clustering method using a new point symmetry-based distance measure, Pattern Recognition 40 (2007) 3430 – 3451

[2]. A. K. Jain, M. Murthy, P. Flynn, Data Clustering: a review, ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[3]. Ujjwal Maulik, Anirban Mukhopadhyay, Sanghamitra Bandyopadhyay - Combining Pareto-Optimal clusters using supervised learning for identifying co-expressed genes, BMC Bioinformatics 2009, 10:27, doi: 10.1186/1471-2105-10-27.

[4]. Ujjwal Maulik, Sanghamitra Bandyopadhyay - Genetic Algorithm-based clustering technique, Pattern Recognition 33 (2000) 1455 – 1465.

[5]. Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, AND David Botstein- Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA, Vol. 95, pp. 14863–14868, December 1998.

[6]. Rousseeuw P: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comp App Math 1987, 20:53-65