A Methodology for Handling a New Kind of Outliers Present in Gene Expression Patterns

Anindya Bhattacharya[1] and Rajat K. De[2,⋆]

[1] Department of Computer Science and Engineering,
Netaji Subhash Engineering College, Kolkata 700152, India
[2] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India

Abstract. Performance of clustering algorithms is largely dependent on selected similarity measure. Efficiency in handling outliers is a major contributor to the effectiveness of a similarity measure. In the present work, we discuss the problem of handling outliers with different existing similarity measures, and introduce the concepts of a new kind of outliers present in gene expression patterns. We formulate a new similarity, incorporated in Euclidean distance and Pearson correlation coefficient, and then use them in various clustering algorithms to group different gene expression profiles. Assessment of the results are done by using functional annotation. Different existing similarity measures in their traditional form are also used with clustering algorithms for performance comparisons. The results suggest that the new similarity improves performance, in terms of finding biologically relevant groups of genes, of all the considered clustering algorithms.

1 Introduction

Clustering algorithm involves measuring similarity between a pair of objects. Some standard similarity measures used in various clustering algorithms include Euclidean distance, various correlation coefficients, Mahalanabis distance. Choice of a similarity measure plays an important role in the performance of a clustering algorithm.

If the objects in a dataset are evenly distributed over the space, the aforesaid similarity measures would be effective. On the other hand, if some of objects due to noise or other factors, called outliers, are included in a dataset, these similarity measures may not lead to good performance of the clustering algorithms. They may be biased towards these outliers. There exist various methods for handling such outliers. They include, among others, statistical approach [1]-[3], distance-based approach, clustering-based approach [4]-[6], density-based local outlier detection approach [7,5] and deviation-based approach [8].

If the expression value(s) of a single (both) gene(s) corresponding to a sample differ much from its (their) mean expression value(s) of the other samples, then the expression value(s) for this(ese) sample(s) differ drastically for the pair of

⋆ Corresponding author.

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 394–399, 2011.

genes. This gives rise to the notion of a different kind of outlier which is introduced here. That is, the sample is an outlier with respect to the gene pair. Distance/similarity measures used by different clustering algorithms are unable to treat an outlier sample and a normal sample differently. All the samples contribute equally during the measurement of distance/similarity.

In order to improve performance of the various similarity measures (including Euclidean distance and Pearson correlation coefficient), with respect to better ability of handling outliers, we introduce the concept of assigning weight values to samples. Instead of using a sample value for the similarity measure, we multiply a weight value with expression value of a sample and then use the resulting value. Weight values are determined in such a way that possible outliers are assigned smaller weight values (*i.e.*, nearly equal to zero). Weight values for non outliers are large, and are nearly equal to 'one'. With this new similarity, Euclidean distance or Pearson correlation coefficient involves low contribution of outlier samples and high contribution of non outlier samples.

In order to incorporate this varying contribution, we assign a weight value to gene expression samples. Any similarity measure that computes pair wise distance similarity, can use the weight assignment technique for better handling the outliers. For comparison, Euclidean distance with weight (WD) and without weight (D), Pearson correlation coefficient with weight ($WCorr$) and without weight ($Corr$), Spearman rank-order correlation coefficient ($RankCorr$) and Jackknife correlation coefficient ($JackCorr$) are used with clustering algorithms K-means [9,8,10], DCCA [11] and ACCA [12]. All the instances of these algorithms are applied to different gene expression datasets and performances are assessed.

2 A New Kind of Outlier

Due to error or other factors in microarray measurements, the expression profiles for a pair of genes may be similar over all the samples except for a few. For these few samples, expression values of the same pair of genes may differ drastically from the other samples. In other words, the expression value(s) of a single (both) gene(s) corresponding to the sample differ much from its (their) mean expression value(s) over the other samples. The sample(s) for which the expression values differ drastically for the pair of genes, give rise to the notion of a different kind of outlier. That is, the sample is an outlier with respect to the gene pair. It may be mentioned here that this outlier is different from the notion of outliers already available in literature [8]. In the later case, the gene as a whole needs to be treated as an outlier with respect to a group of genes, in contrary to the former one where a sample is considered as an outlier corresponding to a gene pair. This situation affects clustering if we consider similarity computation based on expression values only. We introduce a novel methodology to take care of the effect of such outliers.

3 A Methodology for Handling a New Kind of Outliers

Let us consider a set of n genes $X = \{g_1, g_2, \ldots, g_n\}$, for each of which m expression values are given. Let us also consider a set of m microarray experiments/ samples (measurements) $Y = \{e_1, e_2, \ldots, e_m\}$. For each experiment, we have n expression values corresponding to n genes in X. That is, for each gene g_i, there is an m-dimensional vector $\mathbf{x}_i$, where x_{il} is the expression value of g_i in l^{th} experiment e_l. Similarity between gene pair (g_i, g_j) may be computed using Euclidean distance $D(\mathbf{x}_i, \mathbf{x}_j)$ or Pearson correlation coefficient $Corr(\mathbf{x}_i, \mathbf{x}_j)$, and are defined, respectively, as

$$D(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{l=1}^{m} (x_{il} - x_{jl})^2} \tag{1}$$

and

$$Corr(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{l=1}^{m} (x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)}{\sqrt{\sum_{l=1}^{m} (x_{il} - \bar{x}_i)^2 \sum_{l=1}^{m} (x_{jl} - \bar{x}_j)^2}}. \tag{2}$$

Here $\bar{x}_i$ and $\bar{x}_j$ are mean values over m expression values of i^{th} and j^{th} genes respectively.

If l^{th} expression values of a co-expressed, co-regulated gene pair (g_i, g_j), corresponding to an experiment e_l, are such that the sample is an outlier with respect to gene pair (g_i, g_j), both Equations 1 and 2 may be biased towards this outlier. That is, if we consider Equation 2 for measuring similarity, the value should ideally be closed to 1 for a pair of co-regulated genes. Due to this outlier, the correlation value will differ much from 1. In order to reduce this type of misleading contribution of outlier, we introduce the notion of weighting coefficient w_{lij} corresponding to l^{th} expression value and gene pair (g_i, g_j), for all l, i, j.

We determine the weight values so as to reduce the effect of such outliers by assigning lower weight values corresponding to the outlier samples of gene pair (g_i, g_j) and a higher weight values to the other samples. In other words, higher the difference in l^{th} expression values of the genes in the pair (g_i, g_j) from their means, lower is the value of the weight w_{lij}. Considering Euclidean distance for computing difference in l^{th} expression values of both the genes g_i and g_j from their means, we have

$$D_{ijl} = \sqrt{(t_{il} - \bar{t}_i)^2 + (t_{jl} - \bar{t}_j)^2}, \tag{3}$$

where t_{il} and t_{jl} are normalized expression values in $[0, 1]$ of x_{il} and x_{jl} respectively. Similarly, $\bar{t}_i$ and $\bar{t}_j$ are mean of normalized expression values, computed over all the samples, of gene g_i and g_j respectively. Here we have considered normalized expression values in Equation 3, for keeping D_{ijl} bounded to

$\sqrt{2}$ ($\sqrt{(1-0)^2+(0-1)^2} = \sqrt{2}$). For an outlier sample, measured value of D_{ijl} should be high. Weight value w_{lij} for an outlier sample e_l corresponding to a pair of gene (g_i, g_j) should be low. Thus relationship between D_{ijl} and w_{lij} should be such that, increase in D_{ijl} should cause decrease in w_{lij} and vice versa.

In order to reflect such relationship between D_{ijl} and w_{lij}, here we consider exponential function to define weight w_{lij} of a sample e_l corresponding to a pair of gene (g_i, g_j). Thus

$$w_{lij} = e^{-\alpha \times D_{ijl}}, \tag{4}$$

where $\alpha \geq 1$ is a constant. Here the value of α should be such that w_{lij} is nearly equal to zero for $D_{ijl} = \sqrt{2}$. On the other hand, w_{lij} should tend to one for a non-outlier sample. In fact this happens as D_{ijl} tends to zero for a non-outlier sample.

Thus the weight function incorporates w_{lij}, for each l^{th} experiment, in Equations 1 and 2. In these Equations, x_{il} and x_{jl} are replaced by $x_{ijl}^{(w)} = w_{lij} \times x_{il}$ and $x_{jil}^{(w)} = w_{lij} \times x_{jl}$ respectively. Similarly, mean values $\bar{x}_i$ and $\bar{x}_j$, in Equation 2, are replaced by $\bar{x}_{ijl}^{(w)} = \frac{1}{m}\sum_{l=1}^{m} w_{lij} \times x_{il}$ and $\bar{x}_{jil}^{(w)} = \frac{1}{m}\sum_{l=1}^{m} w_{lij} \times x_{jl}$ respectively. It is to be mentioned here that $\bar{x}_{ijl}^{(w)} \neq \bar{x}_{jil}^{(w)}$, although both of them involved the same w_{lij}s. It is further to be noted that the terms t_{il}, t_{jl}, $\bar{t}_i$ and $\bar{t}_j$ in Equation 3 are computed using x_{il}, x_{jl} only. Thus we get distance $WD(\mathbf{x}_i, \mathbf{x}_j)$ and correlation coefficient $WCorr(\mathbf{x}_i, \mathbf{x}_j)$ between a gene pair (g_i, g_j), based on Euclidian distance and Pearson correlation coefficient respectively.

Now the problem remains with the estimation of α-value in Equation 4. This is determined form a plot of WD or $WCorr$ for different pairs of genes with respect to α. From this plot, we have chosen α values at which similarity values (WD or $WCorr$) get saturated.

4 Results

The effectiveness of the weight assignment method along with comparative analysis with the aforesaid similarity measures is demonstrated with K-means [9,8,10], DCCA [11] and ACCA [12] clustering algorithms using five gene expression datasets of Yeast. Datasets are described in Table 1. The performance of all the algorithms is also demonstrated using P-value on functional annotation.

For gene expression data analysis, P-value of GO functional category/ attribute represents the probability of observing at least a given number of genes, in a cluster, from a specific GO functional category/attribute. A specific GO functional category is said to be "enriched" if the corresponding P-value is less than a predefined threshold value. A low P-value indicates that the genes belonging to the enriched functional categories are biologically significant in the corresponding clusters. In the present work, only attributes with P-value $< 1.0 \times 10^{-7}$ are reported as enriched. A clustering solution is considered to be more reliable if the number of enriched functional attributes obtained from a cluster is high. In order to compare the performance of different clustering

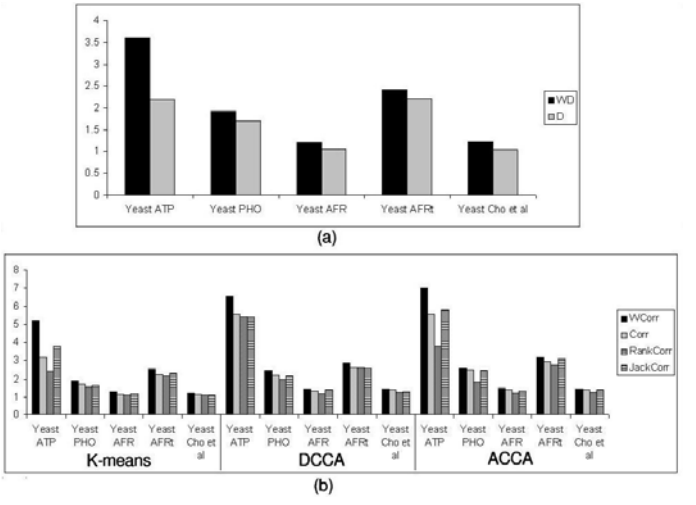

Fig. 1. Average number of functionally enriched attributes per cluster. (a) K-means algorithm with WD and D, (b) K-means, DCCA and ACCA with $WCorr$, $Corr$, $RankCorr$ and $JackCorr$.

algorithms for a microarray gene expression dataset, we can use average number of functionally enriched attributes found per cluster.

Fig. 1 shows that the performance of K-means clustering algorithm with WD (Euclidean distance with weight) is much larger compared to K-means algorithm with D (Euclidean distance but without weight) for all the five datasets. Similarly, Fig. 1 provides the comparative analysis of the clustering algorithms with four similarity measures and shows that $WCorr$ (Pearson correlation coefficient with weight) provides higher number of enriched attributes compared to algorithms with all the other correlation based measures for all the five datasets.

Table 1. Short description of the datasets considered

Name (Organism)	Number of genes	Number of samples
Yeast ATP (Yeast)	6215	3
Yeast PHO (Yeast)	6013	8
Yeast AFR (Yeast)	6184	8
Yeast AFRt (Yeast)	6190	7
Yeast Cho et al. (Yeast)	6457	17

5 Conclusions

Here we have introduced the concepts of a new kind of outlier and a methodology to handle such outliers. New outliers are the samples with respect to a gene pair, for which sample values show large difference from the other samples corresponding to the gene pair. Incorporation of the notion of weight helps in dealing with such outliers while measuring similarity between a pair of gene.

The results suggest that assignment of weight with a similarity measure improves the performance of clustering algorithms in obtaining more biologically significant clusters. The main advantage of weight assignment method is that it is able to deal with outliers without deleting them. Weight assignment method described here is a general framework. Thus it can be used with any distance based similarity measure without changing the basic formulation of that similarity measure.

References

1. Hawkins, D.: Identification of Outliers. Chapman and Hall, London (1980)
2. Rousseeuw, P., Leory, A.: Robust Regression and Outlier Detection. Wiley, New York (1987)
3. Barnett, V., Lewis, T.: Outliers in Statistical Data. Wiley, New York (1994)
4. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
5. Shekhar, S., Chawla, S.: A Tour of Spatial Databases. Prentice-Hall, New Jersey (2002)
6. Hu, T., Sung, S.Y.: Detecting pattern-based outliers. Pattern Recognition Letters 24, 3059–3068 (2003)
7. Schiffman, S.S., Reynolds, M.L., Young, F.W.: Introduction to Multidimensional Scaling: Theory, Methods and Applications. Academic Press, New York (1981)
8. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2001)
9. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, New Jersey (1988)
10. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M.: Systematic determination of genetic network architecture. Nature Genetics 22, 281–285 (1999)
11. Bhattacharya, A., De, R.K.: Divisive correlation clustering algorithm (DCCA) for grouping of genes: Detecting varying patterns in expression profiles. Bioinformatics 24, 1359–1366 (2008)
12. Bhattacharya, A., De, R.K.: Average correlation clustering algorithm (ACCA) for grouping of co-regulated genes with similar pattern of variation in their expression values. Journal of Biomedical Informatics 43, 560–568 (2010)