MIB: USING MUTUAL INFORMATION FOR BICLUSTERING HIGH DIMENSIONAL DATA

Neelima Gupta, Seema Aggarwal
Department of Computer Science, University of Delhi, Delhi, India

ABSTRACT

Most of the biclustering algorithms for gene expression data are based either on the Euclidean distance or correlation coefficient which capture only linear relationships. However, in gene expression data, non linear relationships may exist between the genes. Mutual Information between two variables provides a more general criterion to investigate dependencies amongst variables. In this paper, we propose an algorithm that uses mutual information for biclustering gene expression data. We present the experimental results on synthetic data. None of the distance based biclustering algorithms will identify the biclusters in our synthetic data which our algorithm is able to report. In future we intend to use our algorithm on gene expression data.

KEYWORDS

Biclustering and Mutual Information.

1. INTRODUCTION

Technological advances have led to collection of vast amount of data like microarray data, stock prices data, customer databases etc. The large scale of the data makes it challenging to analyze it to extract any significant information from it. Standard clustering algorithms like k-means clustering work well for small data sets but fair poorly for high dimensional data as they cluster the genes based on their expressions under all the conditions whereas the genes will cluster well under a small subset of conditions. Most of the other conditions which do not belong to the cluster add to the background noise. Moreover, these algorithms compute non-overlapping clusters but a gene may be responsible for more than one cellular activity and hence must be included in more than one cluster. One approach to reduce the set of dimensions is *Projected clustering.* However, the algorithms for projected clustering by (Kevin, Y. et al, 2004, Aggarwal, C. C. et al, 1999, Aggarwal, C. C. and Philip, S. Y., 2000) also compute non-overlapping clusters. The clusters overlap on the dimensions but not on the data points.

Cheng, Y. and Church, G. M., 2000 introduced the notion of biclustering for gene expression data in which the clusters are defined to be a set of genes and a set of conditions under which these genes are most tightly regulated. By definition, biclusters are overlapping. The existing algorithms by (Cheng, Y. and Church, G. M., 2000, Wang, H. et al, 2002, Getz, G. et al, 2000, Kluger, Y. et al, 2003, Ihmels, J. et al, 2002, Bergman, S. et al, 2003, Kloster, M. et al, 2005, Liu, X. and Wang, L., 2007) for biclustering/projected clustering use some kind of similarity measure like Euclidean distance or correlation coefficient. Though these measures have been successfully and satisfactorily used for several years they capture only the linear relationships between the objects. In particular, a vanishing correlation coefficient implies absence of linear dependencies. However, in many applications, like gene expression data and word-document data, non linear relationships may exist between the objects. Moreover, with advances in experimental technology, increasing methodologies are available for unveiling more complex relationships. Hence, we need similarity measures which exploit non linear dependencies.

Steur, R. et al, 2002 have shown that mutual information can be used as a measure of similarity to cluster data. They show that mutual information provides a better and more general criterion to investigate relationships (positive, negative correlation and non linear dependencies) between variables by showing that higher correlation coefficient implies higher mutual information but two variables having very low values of

correlation coefficient (implying no linear relationship) may still be related to each other (non linear dependencies).

Many researchers (Priness, I. et al, 2007, Butte, A. J. and Kohane, I. S., 2000, Michaels, G. S. et al, 1998, Zhou, X. et al, 2004, Slonim, N. et al, 2005) have used mutual information for one way clustering (clustering of genes on the entire set of conditions). These algorithms also support that information theoretic measure is responsive to any type of dependencies, including strongly non linear structures as compared to traditional measures which search only for linear relations. Priness, I. et al, 2007 show that mutual information is a more generalized measure of statistical dependence as compared to both Euclidean distance and correlation coefficient and is resistant to outliers and missing data. Butte, A. J. and Kohane, I. S., 2000 compute pairwise mutual information for all genes against each other. They hypothesize that an association between two genes indicated by a high amount of mutual information between them would also signify biological relationship.

In this paper, we present an algorithm for biclustering using mutual information. Let G be the set of genes and C be the set of conditions in the expression data. We define a bicluster as a pair (G', C'), where G' is the subset of genes (objects) which are most closely related to each other under the subset C' of conditions (dimensions) and C' is the subset of conditions (dimensions) under which the genes (objects) of G' are more closely related to each other as compared to other conditions (dimensions).

We assume that if there is a group of genes which exhibit some relationships in their expressions under a subset of conditions, then these conditions are also related to each other in some sense. We generate synthetic data to show that such an assumption is not implausible. One way then to choose relevant conditions is to find pairwise mutual information amongst all pairs of conditions and select those pairs whose mutual information is above some threshold. The problem in this approach is that if two pairs of conditions c_1, c_2 and c_3, c_4 have high mutual information then all four will be selected whereas there is no relation between c_1 and c_3. In our algorithm, we use a reference condition say c* and select conditions which have high degree of relation with c*. Since we do not know which c* is most suitable we do it for all the conditions one by one. Here we would like to mention that though we have defined the problem in the context of gene expression data, it has its applications in other places like word document data, stock prices monitoring data and others.

Our algorithm works in three stages. In the first stage we take a gene as a seed and find the subset of genes which are most closely related to the input seed gene. For this we compute the pair wise mutual information of all the genes with the seed gene over all the conditions and select the ones having mutual information above some threshold. In the second stage, the algorithm identifies the subset of conditions under which the selected subset of genes is most related to each other. In the third and the final stage the algorithm refines the gene set based on the condition subset. It selects those genes which are most related with the seed gene under the reduced set of conditions identified in stage two. The gene subset and the condition subset then define a bicluster. At each stage we select only those genes or conditions which have mutual information greater than certain threshold.

We compared the performance of our algorithm with three other algorithms (each using a different similarity measure) on synthetic data.

2. THE MUTUAL INFORMATION

The mutual information between two random variables X and Y is a measure of information contained in X about Y or the information contained in Y about X. If given a value of X, it is easy to predict the value of Y then X contains good amount of information about Y. Clearly with this definition, if X and Y are independent the mutual information between them is zero and it is high if they are highly dependent or closely related to each other. Thus Kullback has defined mutual information between two random variables as a measure of divergence from the hypothesis that X and Y are independent.

2.1 The Kullback Divergence

An experiment performed on a system A puts the system in one of the states a_1, a_2 …a_{NA}, each with its corresponding probability $p(a_i)$ The information gained by the system through a series of experiments is the amount of surprise one feels on reading the outcomes of the experiments. Thus if one hypothesize that

probability distribution observed by the outcomes is $\{p^0\}$ and the actual densities are $\{p\}$, the Kullback divergence $K(p/p^0)$ between the two probability distributions is given by

$$K(p/p^0) = \sum_i p_i \log (p_i / p_i^0)$$

Kullback divergence can be interpreted as the information gained when the assumed probability distribution $\{p^0\}$ is replaced by the final distribution $\{p\}$. $K(p/p^0)$ is always greater than or equal to zero according to Haykin S. 2007. It equals zero if and only if $\{p^0\}$ and $\{p\}$ are same. In our case the assumed probability distribution $\{p^0\}$ is given by the hypothesis that two variables X and Y are statistically independent. Thus $p^0_{XY}(x_i, y_j)$ is given by

$$p^0_{XY}(x_i, y_j) = p_X(x_i)\, p_Y(y_j)$$

The final distribution $\{p\}$ is given by the observed joint probability densities $p_{XY}(x_i, y_j)$. Thus using Kullback divergence mutual information is defined as

$$I(X, Y) = \sum_{i=1}^{nx} \sum_{j=1}^{ny} p_{XY}(x_i, y_j) \log (p_{XY}(x_i, y_j) / (p_X(x_i)\, p_Y(y_j)))$$

X takes values $x_1, x_2 .. x_{nx}$, Y takes values $y_1, y_2 \ldots y_{ny}$,), $p_{XY}(x_i, y_j)$ represents the joint probability distribution of X, Y and, $p_X(x_i)$ and $p_Y(y_j)$ are the marginal distributions of X and Y respectively. The mutual information is zero if and only if X and Y are statistically independent i.e. vanishing mutual information does imply that the two variables are independent.

Since Kullback divergence is computed using discrete probabilities, we estimate probability densities using the widely used histogram method as explained in Michaels G. S. et al, 1998. Let the bins of the random variable X be denoted by a_i, $i = 1..N_x$ and the bins of the random variable Y be denoted by b_j, $j = 1..N_y$. Let $f_X(i)$ and $f_Y(j)$ denote the number of observations of X and Y falling in the bin a_i and b_j respectively. The probabilities $\{p(a_i)\}$ and $\{p(b_j)\}$ are then estimated as

$$p(a_i) = f_X(i) / n \text{ and } p(b_j) = f_Y(j) / n.$$

Let $f_{XY}(i, j)$ denote the number of observations such that X falls in bin a_i and Y falls in bin b_j. Then the mutual information between X and Y is estimated as

$$I(X, Y) = \log n + (1/ n) \sum_{i=1}^{nx} \sum_{j=1}^{ny} f_{XY}(i, j) \log ((f_{XY}(i, j))) / (f_X(i)\, f_Y(j)))$$

3. THE MUTUAL INFORMATION BASED BICLUSTERING ALGORITHM

Traditional distance based or correlation coefficient based clustering/biclustering algorithms capture only linear relationships amongst the objects. However, in complex data sets, objects may exhibit non-linear complex relationships. In this section, we present an algorithm to find biclusters using mutual information. Since mutual information captures more general relations, our algorithm is able to group objects possessing more complex relationships.

Let G be the set of genes and C be the set of conditions in the expression data. The number of genes in G is denoted by N_g and the number of conditions in C be denoted by N_c. We find biclusters which are pairs of (G', C'), where G' is the subset of genes which are most closely related to a gene seed g* under the subset C' of conditions and C' is the subset of conditions under which subset G' of genes are most closely related to the gene seed. Different biclusters are obtained by varying the gene seed (chosen randomly) and biclusters with more than 75% overlap are merged.

Our algorithm takes a gene (g*) as a seed and proceeds in three steps. In the first step we find the set of genes G_0 which is most related to the input seed gene. Gene threshold t_g is used to select genes with high mutual information with g*. In the second step, the algorithm identifies the conditions under which the set of genes in G_0 show maximum dependence. Condition threshold t_c is used to select the related conditions. In the third and the final step the algorithm selects from the whole expression data those genes which are most related to the gene seed under the reduced set of conditions identified in step two. Since we do not know which c^* is best, second and third steps are repeated for every condition. The algorithm is run for all seed genes and highly overlapping biclusters are merged. The procedure MIB () and other procedures in Figure 1 summarize the algorithm. Procedure mi (x_1, x_2, n) computes the pairwise mutual information between the two variables x_1 and x_2 of length n according to the formula given in section 2.

MIB (g^*, G, C, t_g, t_c, N_g, N_c)
1) G_0 = Compute-G0 (g^*, G, C, t_g , N_g, N_c)
2) for all c^* in C do
2.1) C' = Compute-conditions (c^*, G_0, C, t_c, N_c)
2.2) G' = Compute-biclus-genes (g^*, C', t_g, N_g, $N_{c'}$)
2.3) output (G', C')

Compute-G0 (g^*, G, C, t_g , N_g, N_c)
1) for i = 1 to N_g do
mi_g [i] = mi (g_i, g^*, N_c)
2) $\mu = \sum_i$ mi_g [i] /N_g
3) $\sigma^2 = \sum_i$ (mi_g [i] - μ)^2/N_g
4) $G_0 = \{g_i : ((mi\_g [i] - \mu)/\sigma) > t_g\}$
5) return G_0

Compute-condititons (c^*, G_0, C, t_c, N_c)
1) for j = 1 to N_c do
mi_c [j] = mi (c_j, c^*, N_{g0})
2) $\mu = \sum_j$ mi_c [j] /N_c
3) $\sigma^2 = \sum_j$ (mi_c [j] - μ)^2/N_c
4) $C' = \{c_j : ((mi\_c [j] - \mu)/\sigma) > t_c\}$
5) return C'

Compute-biclus-genes (g^*, C', t_g, N_g, $N_{c'}$)
1) for i = 1 to N_g do
mi_g [i] = mi (g_i, g^*, $N_{c'}$)
2) $\mu = \sum_i$ mi_g [i] /N_g
3) $\sigma^2 = \sum_i$ (mi_g [i] - μ)^2/N_g
4) $G' = \{g_i : ((mi\_g [i] - \mu)/\sigma) > t_g\}$
5) return G'

Figure 1. Procedure MIB () and other procedures. N_{g0} represents the number of genes in G_0 and $N_{c'}$ represents the number of conditions in C'

Figure 2. Synthetic data showing two overlapped biclusters M1 and M2

4. EXPERIMENTAL RESULTS

In order to study the performance of our algorithm we used computer generated synthetic data. The main idea behind the synthetic data was to model nonlinear relationships between genes of the bicluster over a subset of conditions in such a way that the subset of conditions of the bicluster is also related over the subset of genes. We say that two genes g_i and g_k (having expression values e(i, j) and e(k, j) for condition j) have circular relationship and additive relationship if for all j they satisfy $e(i, j)^2 + e(k, j)^2$ = constant and e(i, j) = e(k, j) + constant, respectively.

We created two types of synthetic expression data each with two overlapping biclusters for 100 genes and 100 conditions (refer to Figure 2). The two expression data differ in the kind of relationships. In the first expression data the first bicluster M1 consists of genes g_1 to g_{60} and conditions c_1 to c_{50}. These 60 genes had circular relationships with a gene g1 chosen as seed under the first 50 conditions. The second bicluster M2 consisted of genes g_{51} to g_{100} and conditions c_{41} to c_{100}. These genes had circular relationship with g_{61} under conditions c_{41} to c_{100}. In the second expression data the circular relationships were replaced by additive relationships. The rest of the rows and columns (D1 and D2 in the figure 2) in both the expression data were given high constant value 10 to make them independent of the rows and columns in the biclusters.

For circular data, at t_c = -0.5 and at a very low gene threshold, the biclusters reported had almost all the genes. As we increase t_g to -.1 we were able to find M1 or M2 depending upon the gene seed. As we increase t_g further to 1.5 we were able to find the genes in M1∩ M2. For the additive biclusters also at very low thresholds we found almost all the genes. At t_c = -0.5 and t_g = -.5 we were able to find M1 or M2 depending upon the gene seed. As we increase t_g to 1 we were able to find the genes in M1∩ M2.

We also implemented distance based projected clustering algorithm PROCLUS by Aggarwal, C. C. et al, 1999, MSB by Liu, X. and Wang L., 2007 and score based biclustering algorithm ISA by Bergmann, S. et al, 2003. PROCLUS and MSB always find D1 and D2 as the distance between expression value of genes of D1 and D2 is zero (as they are all same). They never find M1 or M2 which our algorithm is able to identify.

When the expression values in M1 and M2 are kept low as compared to 10 (in D1 and D2), ISA also finds D1 and D2 and not M1 or M2 as ISA favors genes and conditions with higher expression values.

5. CONCLUSION AND FUTURE WORK

As the mutual information captures more general relationships, our algorithm will be able to discover more and better biclusters as compared to algorithms using traditional similarity measures like Euclidean distance and correlation coefficient. The major challenge in designing such an algorithm was to detect the set of relevant conditions. We have assumed that if there is a group of genes which exhibit some special type of relationships in their expressions under a subset of conditions, then these conditions are also related to each other in some sense.

In the present work we were able to find biclusters in which largely one type (only circular or only additive for example) of relationship exist. Currently we are working on applying this algorithm for gene expression data for yeast. Also, we would like to find more general biclusters with mixed relationships. The thresholds t_g and t_c only affect the granularity of the biclusters. In future, we would like to study the effect of input parameters in more detail and try to eliminate them or reduce their number.

REFERENCES

Aggarwal, C. C. et al, 1999, *Fast Algorithms For Projected Clustering*. *Proceedings of ACM SIGMOD Int'l Conf. Management of Data*, pp 61--72.

Aggarwal, C. C. and Philip, S. Y.*, 2000, Finding generalized projected clusters in high dimensional spaces, Proceedings of ACM SIGMOD Int'l Conf. Management of Data,* pages 70--81.

Bergmann, S. et al, 2003, *Iterative signature algorithm for the analysis of large-scale gene expression data. Physical Review,* American Physical Society., volume 67, pp 1--18.

Butte, A. J. and Kohane, I. S., 2000, *Mutual Information relevance networks:Functional Genomic clustering using pairwise entropy measurements, Proceedings of PSB*, volume 5, pp 415--426.

Cheng, Y. and Church, G. M. 2000, *Biclustering Of Gene Expression Data, System Molecular Biology*, volume 8, pp 1--93.PNAS , volume 97, pp 12079--12084.

Ihmels, J. et al, 2002, *Revealing Modular Organization In The Yeast Transcription Network, Nature Genetics, Nature Publishing Group*, volume 31, pp 1--370.

Kevin, Y. et al, 2004, *HARP: A Practical Projected Clustering Algorithm, IEEE Transactions On Knowledge And Data Engineering*, *IEEE Computer Society*. Volume 16.

Kloster, M. et al, 2005, *Finding regulatory modules through large-scale gene-expression data analysis. Bioinformatics*, volume 21, pages 1172--1179.

Kluger, Y. et al, 2003, *Spectral Biclustering Of Microarray Data: Coclustering Genes And Conditions, Genome Research*, Cold Spring Harbor Laboratory Press.

Liu, X. and Wang, L 2007, *Computing the maximum similarity bi-clusters of gene expression data, Bioinformatics*, volume 23, pp 50--56.

Michaels, G. S. et al, 1998. *Cluster Analysis and Data Visualization of Large Scale Gene Expression Data, Proceedings of PSB*, volume 3, pages 42--53.

Priness, I. et al, 2007. *Evaluation of gene expression clustering via mutual information distance measure. BMC Bioinformatics*.

Slonim ,N. et al, 2005, *Information based clustering, PNAS*, volume 102, pp 18297--18302.

Steuer, R. et al, 2002, *The mutual information:Detecting and evaluating depencies between variables, Bioinformatics*, volume 18 Suppl 2, pp S231--S240.

Wang, H. et al, 2002, *Clustering by Pattern Similarity in Large Data Sets*, *Bull, Math. Biol. 46*, pp 515--527. ACM Press.

Zhou, X. et al, 2004, Gene Clustering Based on Clusterwide Mutual Information, Jouranl of Computational Biology, volume 11, Number 1, pp 147--161.