International Journal of Information Technology and Knowledge Management
July-December 2008, Volume 1, No. 2, pp. 559-567

MINING GENE EXPRESSION DATA FOR ASSOCIATION RULES USING MUTUAL INFORMATION BASED ON ENTROPY

M. Anandhavalli, M. K. Ghose & K. Gauthaman

Abstract

Knowledge discovery from gene expression databases has become an important research area for biologists since the growing number of gene sequences was obtained. Using computation techniques such as data mining to find the association relationship among these gene data is of great interest and challenging aspect. Extracting information from large datasets (e.g., from human genome project, gene expression data) is a well-studied research problem. However, experimental research is time consuming, and does not use the currently available gene sequence data. Using data mining techniques, especially association rule techniques, it is getting essential to find better ways to extract relations (inferences) from them. In this paper, clustering and mutual information based on entropy has been used to generate association rules from gene expression data. Our results suggest that association rule may be well suited to predict the relationships among the silencer genes.

Keywords: Gene expression data, association rule, entropy, mutual information.

1. Introduction

Microarray-based genomic surveys and other high-throughput approaches (ranging from genomics to combinatorial chemistry) are becoming increasingly important in biology and chemistry. As a result, we need to develop our ability to "see" the information in the massive tables of quantitative measurements that these approaches produce.

Clustering [2,3] is an old studied technique used to extract this information from biological and other data sets. This follows from the fact that co-expressed genes have similar patterns of expression. Clustering groups records that are "similar" in the same group. It suffers from two major defects. It does not tell you how the two genes clusters are exactly related. Moreover, it gives you a global picture and any relation at a local level can be lost.

It is proposed to use mutual information based on entropy for generating association rules [6]. Apart from the usual positive correlations between the genes, this criterion would also discover association rules with negative correlations in the data sets.

An attempt has been made to find results of the form Gene1 $\wedge$ Gene 2 –> $\wedge$Gene3, which can be interpreted as follows: Gene1 and Gene2 are co expressed and have

silencing effect on Gene. The results from the experiments have been compared to those obtained from clustering.

2. OBJECTIVES AND SCOPE

Initially it was attempted to apply the mentioned approach on the gene expression data of the entire yeast genome obtained under various experiments [2,3,5]. However, since the number of genes was large (2468), it was found that the approach of studying them at once was not feasible. First, even after tuning the various parameters (like support, significance level), the program ran out of memory. This was because even with binary data, 2468 attributes may lead to the power (2, 2468) relations. For the problem under consideration, the genes are attributes and relationships among the genes need to be found. To overcome this problem an alternative approach as detailed below was followed. The genes, which were already known to be related, using the results obtained from clustering [2,3,5] were considered, resulting in decrease the number of attributes to manageable levels (both for program). The above mentioned approach was the used to find the relationships (positive, negative) among the attributes. Another problem studied was that the four Open Reading Frames (YMR219W, YDR363W, YHR154W, YJ076W), which are known to be silencers, however, the type of genes they turn off is not known [2,3,5]. One of the obvious places to look for is a subset of the genes responsible for Protein Synthesis and ATP Synthesis. Using the approach a clue regarding the type of genes with these silencers affect has been obtained. The above-mentioned approach was also used on certain non-biological databases and compared the results obtained from this approach with those of clustering.

3. ASSOCIATION RULES

Various algorithms exist to extract association rules, of which the apriori [4] algorithm is the most commonly used and we too shall use it in this study. It entails a two-step process, defined below, which consists of first generating the set of frequent itemsets, from which association rules are extracted that are above a certain confidence level.

(1) Given a set of items I, the input consist of a set of transactions D, where each transaction T is a non-empty subset of items taken from the itemset I, so $T\ I$.

(2) Given an itemset $T \in I$ and a set of transactions D, we define the support of T as support $D(T)$ equals the proportion of transactions that contain T to all transactions $|D|$.

(3) By setting a minimum support level α, where $0 \leq \alpha \leq 1$, we define frequent itemsets to be itemsets where support $D(T) \geq \alpha$.

In this paper, the association pattern *ai* consisting of items $\{i_1, i_2, \ldots i_p\}$ is considered α-significant if it satisfies the following conditions:

(1) The support for *ai*, defined as Pr(*ai*), is at least α; i.e., Pr(*ai*) $\geq \alpha$.

(2) The interdependency of $\{i_1, i_2, \ldots i_p\}$ as measured by mutual information measure MI(*ai*) = Log2 Pr(i_1, i_2,…i_p)/ Pr(i_1)Pr(i_2)…Pr(i_p) is significant.

4. METHODS

The data were obtained from experiments conducted at Stanford by Eisen *et al*. and DeRisi *et al.* [2,3,5]. A cluster of genes responsible for ATP Synthesis shown in Table 1 and Protein Synthesis shown in Table 2 was taken to obtain their gene expression values along with the four silencer genes (whose effect was unknown) under various experimental conditions [2].

Table 1
ATP Synthesis Genes and Silencers

	YMR219W	*YDR363W*	*YHR154W*	*YJLO76W*
Alpha Factor experiments	2.00-6.01 9 cases	0.207 – 0.506 5 cases	0.405-0.409 4 cases	0 cases
Cdc15 experiments	0.236-0.5921 3 cases	0.371-0.592 4 cases	0.226-0.553 11 cases	0.226-0.557 14 cases
Elutriation experiments	0.235-0.736 25 cases	0.314-0.736 20 cases	0.236-0.547 13 cases	0.253-0.736 16 cases

Table 2
Protein Synthesis Genes and Silencers

	YMR219W	*YDR363W*	*YHR154W*	*YJLO76W*
Alpha Factor experiments	0.33-0.630 23 cases	0.32-0.630 9 cases	0.297-0.630 17 cases	0.32-0.673 16 cases
Cdc15 experiments	0.322-0.654 60 cases	0.228-0.577 40 cases	0.228-0.577 70 cases	0.258-0.577 30 cases
Elutriation experiments	0.228-0.577 11 cases	0.228-0.577 9 cases	0.228-0.577 14 cases	0.258-0.577 21 cases
Heat experiments	0.317-0.650 10 cases	No case	0.317-0.650 10 case	0.317-0.514 10 cases
Cold experiments	No case	0.918-0.918 2 cases	No case	0.918-0.918 2 cases

ATP and Protein Synthesis are two major functions of the cell and the genes responsible for this should be first to be considered in order to study the effect of the silencers. Using this approach it is able to predict the effect of these silencers on some of the genes as well the relationship among the genes itself.

For non-biological databases, breast cancer data was obtained from UCI Machine Learning Repository. The data was discretized into binary values. For gene expression data a value of one represented that the gene was on and a value of zero represented that the gene was off. This was done by finding the average value of gene expression. Any expression value less than the average was supposed to mean that the gene was turned off and a value greater than the average was taken as the gene being turned on.

The program takes as input the data file having the various attributes in binary form (0 or 1). Further only certain attributes of the data set were relevant. Using domain knowledge the data set was pruned to represent the relevant attributes. Further the data was discretized into binary values. The binary values were chosen in accordance with interpretation required. The data was formatted in the way expected by the program.

5. RESULTS AND INTERPRETATION

5.1. Biological Data Sets

Genes whose functions are well known (Protein and ATP synthesis) and also took four silencer genes whose interaction is unknown was considered. The expression data of these genes from the various experiments [2,3,5] were obtained. Then, the MI program (that uses mutual information based on entropy to extract association rules) on the data set has been executed and results were obtained. The complete results are given in the Table 1 and Table 2. Most of the relations have high mutual information values. This is expected since all the genes had similar function. From the data set relations regarding the various genes were inferred.

For example, consider a result from the *alpha factor arrest* experiments (Protein Synthesis):

{YDR211W, YDR283C, YFR009W, YMR282C} (0.517)
0.232,0,0,0,0,0.232,0.513,0,0.232,0,0.352,0,0.352,0,0.352,0.482

From this it is interpreted that when YFR009W, YMR282C are turned on thenYDR211W, YDR283C are turned off. It is concluded that there is a negative correlation between these pairs. As another example, consider a result from the *Cdc15* experiment (ATP Synthesis):

{YPL078C, YDR298C} (0.837) 0.508,0,0,0.328

It is observed that information value is unusually high. Further, from the entropy values it is concluded that both of these genes are turned off together. Similarly, consider a result from the *heat experiments* (Protein Synthesis): {YOR133W, YDR385W} (1) 0.5,0,0,0.5

It is observed that mutual information is 1. Further, from the entropy values it is concluded that these genes have positive correlation. It is also drawn similar interpretation regarding silencer genes as follows:

Cold experiments (Protein Synthesis): {YDR363W, YJL076W} (0.918) 0,0.528,0.39,?

Interpretation: there are very few instances when YDR363W and YJL076W are off.

Heat experiments (Protein Synthesis): {YMR219W, YHR154W} (0.650) 0.431,0,0,0.219

Interpretation: YMR219W and YHR154W are highly related. They seem to be turned on/off together. An interesting observation was that for other group of genes the mutual information was about 3.33, which was much less than the above pair.

Cdc 15 experiments (Protein Synthesis): {YOR133W, YDR385W} (0.997) 0.484,0,0,0.513

Interpretation: YOR133W and YDR385W are positively correlated.

{YDR363W, YJL076W} (0.506) 0.328,0,0.260,0.464

Interpretation: from this it is concluded that when the silencer YDR363W is turned on, the silencer YJL076W is turned off.

Alpha factor arrest experiments (ATP Synthesis): {YDR363W, YHR154W, YJL076W} (0.315) 0.53,0.232,0.352,0.352,0,0.232,0.352,0.431

Interpretation: from this it is inferred that when YDR363W is turned on, YHR154W and YJL076W are turned off.

To draw some inferences on a global scale, consider the table1 and table 2 that was obtained from the results. *The entry in each cell of table 1 represents the range of mutual information and the number of relationships where the corresponding gene occurs.* The interesting observation here is that the silencer YJL076W does not occur in any case. So, it seems that this gene does not play any significant role in the ATP Synthesis. It is observed that in these experiments YMR219W seems be a major player as far as silencing is concerned.

The entry in each cell of table 2 represents the range of mutual information and the number of relationships where the corresponding gene occurs. The interesting observation is that YDR363W seems to have no role in shutting of the cell machinery in response to heat experiments. Further in cold experiments YMR219W and YHR154W seem to play no role, while YDR363W and YJL076W seem to play a significant role (mutual information ~ 1).

Further while comparing across experiments the studied silencers seem to have a much more impact on Protein synthesis than on ATP Synthesis. This is even more apparent in the case of *Cdc 15* experiments.

5.2. Non-Biological Data Sets

For the Breast-cancer-Wisconsin data we have been able to run the same program that generates the association rules. The process of association rules generation and the results are given below.

Process

(1) Get data (file: *breast-cancer-wisconsin.data*): 11 attributes, 699 samples.

(2) Remove unique attributes (IDs). Here, *sample_code_number* attribute has been removed.

(3) Remove those samples (total 16) that contain "?" (missing data) as a value for some of their attributes (so, we are left with 10 attributes and 683 samples).

(4) Since all the attributes in this database are multiple-valued and integers, we discretize them based on their average values (which is (maximum attribute value + minimum attribute value) / 2) shown in Table 3.

Table 3
Attributes discretization based on their average values

Attribute	*0*	*1*
Clump_Thickness	1-5.5	5.5-10
Uniformity_of_Cell_Size	1-5.5	5.5-10
Uniformity_of_Cell_Shape	1-5.5	5.5-10
Marginal_Adhesion	1-5.5	5.5-10
Single_Epithelial_Cell_Size	1-5.5	5.5-10
Bare_Nuclei	1-5.5	5.5-10
Bland_Chromatin	1-5.5	5.5-10
Normal_Nucleoli	1-5.5	5.5-10
Mitoses	1-5.5	5.5-10
Classes	2 (benign)	4 (malignant)

(5) Run the program to generate association rules using *mutual information based on entropy* metric. The following output has been obtained (where, MI = mutual information value based on entropy, E's are the entropy values) which is shown in Table 4.

Table 4
Association Rules Generation using *Mutual Information Based on Entropy* Metric

Correlation_set	*MI*	*E(∧a∧b)*	*E(∧ab)*	*E(a∧b)*	*E(ab)*
{ Clump_Thickness, Uniformity_of_Cell_Size}	0.171	0.388	0.251	0.365	0.41
{ Clump_Thickness, Uniformity_of_Cell_Shape }	0.169	0.393	0.266	0.36	0.413
{ Clump_Thickness, Marginal_Adhesion }	0.073	0.394	0.269	0.429	0.337
{ Clump_Thickness, Single_Epithelial_Cell_Size }	0.085	0.380	0.228	0.433	0.331
{ Clump_Thickness, Bare_Nuclei }	0.186	0.406	0.302	0.329	0.434
{ Clump_Thickness, Bland_Chromatin }	0.142	0.390	0.259	0.383	0.394
{ Clump_Thickness, Normal_Nucleoli }	0.133	0.390	0.259	0.39	0.387
{ Clump_Thickness, Mitoses }	0.041	0.341	0.075	0.488	0.179
{ Clump_Thickness, Classes }	0.351	0.427	0.352	0.149	0.493
{Uniformity_of_Cell_Size, Uniformity_of_Cell_Shape }	0.371	0.319	0.202	0.175	0.441
{Uniformity_of_Cell_Size, Marginal_Adhesion }	0.152	0.332	0.24	0.329	0.357
{Uniformity_of_Cell_Size,Single_Epithelial_Cell_Size }	0.175	0.315	0.189	0.331	0.355
{Uniformity_of_Cell_Size, Bare_Nuclei }	0.163	0.382	0.352	0.266	0.402
{Uniformity_of_Cell_Size, Bland_Chromatin }	0.197	0.341	0.262	0.283	0.392
{Uniformity_of_Cell_Size, Normal_Nucleoli }	0.204	0.337	0.251	0.283	0.392
{Uniformity_of_Cell_Size, Mitoses }	0.043	0.287	0.089	0.442	0.17
{Uniformity_of_Cell_Size, Classes }	0.367	0.407	0.4	0.034	0.471
{Uniformity_of_Cell_Shape, Marginal_Adhesion }	0.143	0.341	0.24	0.345	0.357
{Uniformity_of_Cell_Shape, Single_Epithelial_Cell_Size}	0.132	0.332	0.215	0.362	0.34
{ Uniformity_of_Cell_Shape, Bare_Nuclei }	0.207	0.376	0.326	0.251	0.421
{ Uniformity_of_Cell_Shape, Bland_Chromatin }	0.211	0.344	0.247	0.289	0.4
{ Uniformity_of_Cell_Shape, Normal_Nucleoli }	0.211	0.341	0.24	0.293	0.398
{ Uniformity_of_Cell_Shape, Mitoses }	0.051	0.293	0.075	0.448	0.179
{ Uniformity_of_Cell_Shape, Classes }	0.367	0.41	0.392	0.052	0.476
{ Marginal_Adhesion, Single_Epithelial_Cell_Size}	0.11	0.301	0.262	0.305	0.305
{ Marginal_Adhesion, Bare_Nuclei }	0.168	0.359	0.376	0.198	0.381
{ Marginal_Adhesion, Bland_Chromatin }	0.149	0.326	0.317	0.247	0.352
{ Marginal_Adhesion, Normal_Nucleoli }	0.116	0.332	0.329	0.273	0.334
{ Marginal_Adhesion, Mitoses }	0.038	0.248	0.109	0.4	0.154
{ Marginal_Adhesion, Classes }	0.266	0.409	0.448	0.043	0.431
{ Single_Epithelial_Cell_Size, Bare_Nuclei }	0.098	0.367	0.415	0.224	0.334
{ Single_Epithelial_Cell_Size, Bland_Chromatin }	0.090	0.331	0.36	0.259	0.308
{ Single_Epithelial_Cell_Size, Normal_Nucleoli }	0.103	0.324	0.347	0.251	0.314
{ Single_Epithelial_Cell_Size, Mitoses }	0.054	0.223	0.096	0.367	0.165
{ Single_Epithelial_Cell_Size, Classes }	0.21	0.411	0.469	0.060	0.404
{ Bare_Nuclei, Bland_Chromatin }	0.2	0.368	0.228	0.342	0.41
{ Bare_Nuclei, Normal_Nucleoli }	0.139	0.380	0.262	0.372	0.385
{ Bare_Nuclei, Mitoses }	0.026	0.334	0.120	0.482	0.16
{ Bare_Nuclei, Classes }	0.444	0.411	0.34	0.060	0.498
{ Bland_Chromatin, Normal_Nucleoli }	0.172	0.338	0.276	0.286	0.376
{ Bland_Chromatin, Mitoses }	0.035	0.281	0.102	0.436	0.16
{ Bland_Chromatin, Classes }	0.306	0.412	0.419	0.068	0.458
{ Normal_Nucleoli, Mitoses }	0.041	0.274	0.096	0.429	0.165
{ Normal_Nucleoli, Classes }	0.278	0.415	0.428	0.082	0.451
{ Mitoses, Classes }	0.06	0.406	0.522	0.025	0.207

Results

From the obtained output it is observed that the relations that contain *Class* attribute that determines to which class each sample belonged. The highest MI-values were obtained for almost all of those relations that included *Class attribute.* Namely the highest MI-values were observed between:

(1) *Bare_Nuclei* and *Class*. Further, by observing the entropy values we conclude that almost all benign tumors would have low *Bare_Nuclei*. The same follows and for U*niformity_of_Cell_Shape, Bland_Chromatin, Normal_Nucleoli, Marginal_Adhesion, and Clump_Thickness* attributes.

(2) *Uniformity_of_Cell_Size* and *Class*. By observing the entropy values it is concluded that *Uniformity_of_Cell_Size* would not be a good indicator of whether a tumor is benign or malignant.

These results show us that there is a large correlation between almost every single attribute and the *Class* (benign or malignant). The clustering tree obtained is given below:

((Clump_Thickness,(((((((Uniformity_of_Cell_Size, Uniformity_of_Cell_Shape), Class), Bland_Chromatin), Bare_Nuclei), Marginal_Adhesion), Normal_Nucleoli), Single_Epithelial_Cell_Size)), Mitoses).

6. CONCLUSIONS

The proposed approach (using clustering along with the mutual information) has some merit in extracting information from huge data sets by pruning the initial information (to bring it down to the manageable levels) and then finding the association rules among the attributes. Further, the approach used to predict the relationships among the silencer genes and other genes could be extended to genes of unknown function.

7. FUTURE WORKS

Presently the problem in which the attributes can take only binary values has been studied. It would be more useful to study similar problem with the multi-valued and real valued attributes. The software needs to be extended so that it could handle real valued attributes as well as work with a large number of attributes that is often the case for the large datasets. It would also be helpful to explore different classes of correlation metrics with corresponding algorithms to build association rules and compare the results obtained from this.

REFERENCES

[1] *Breast Cancer Data*, ftp://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin.

[2] Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D, Cluster *Analysis and Display of Genome-wide Expression Patterns*, *Proc. Natl. Acad. Sci.* USA 95: 14863-14868, (1998).

[3] Joseph L. DeRisi,Vishwanath R. Iyer, Patrick O, Brown Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale, Science (1997) Oct 24, 278, 5338, 680–686.

[4] Rakesh Agrawl, Tomasz Imielinski and Arun Swami, Mining Association Rules between Sets of Items in Large Databases, In *Proc of ACM SIGMOD Conference on Management of Data* Washington D.C. (May 1993).

[5] Spellman *et al*., Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharmomyces cerevisiae by Microarray Hybridization, Mol, Biol.Cell Online **9**, (12), (December 1998), 3273–3297.

[6] Tiyagura, Mining Association Rules Based on Mutual Information, *M. S thesis Dissertation* at Iowa State University (1999).

M. Anandhavalli & M. K. Ghose
Department of Computer Science Engineering
SMIT, Majitar
East Sikkim, INDIA
E-mail: *anandhigautham@gmail.com*

K. Gauthaman
Department of Pharmacognosy
HPI, Majitar
East Sikkim, INDIA.