

Mining Association Rules from Large Datasets

(Neeraj Koul, Alexei Zapari, Smruti Behera)

Abstract

Extracting information from large datasets is a well-studied research problem. As larger and larger data sets become available (e.g., from human genome project, gene expression data, customer behavior data from organizations such as Wal-Mart) it is getting essential to find better ways to extract relations (inferences) from them. We plan to use clustering and mutual information based on entropy [1] to generate association rules from biological datasets as well as non-biological datasets.

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 [3,4] 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.

We propose to use mutual information based on entropy for generating association rules [1]. Apart from the usual positive correlations between the genes, this criterion would also discover association rules with negative correlations in the data sets. We expect to find results of the form $\text{Gene1} \wedge \text{Gene 2} \rightarrow \neg \text{Gene3}$, which can be interpreted as follows: Gene1 and Gene2 are co expressed and have silencing effect on Gene 3. We will compare the results from our experiments to those obtained from clustering. Since the approach is a general one we will also apply it on some non-biological databases to show that it can be used to extract useful information from these datasets as well.

2 Objectives and Scope

We started by trying to apply the mentioned approach on the gene expression data of the entire yeast genome obtained under various experiments [3,4,5]. However, since the number of genes was large (2468), we found that the approach of studying them at once was not feasible. First, even after tuning the various parameters (like support, support fraction, 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 (which the software

was not designed to handle). Here, we need to know that for the problem under consideration, the genes are attributes, as we need to find relationships among them. To overcome this problem we used another approach. We took genes, which were already known to be related, using the results obtained from clustering [3,4,5]. This allowed us to decrease the number of attributes to manageable levels (both for program). We used the approach above to find the relationships (positive, negative) among the attributes.

Another problem that we studied was about 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 [3,4,5]. We believe one of the obvious places to look for is a subset of the genes responsible for Protein Synthesis and ATP Synthesis. Using the approach we hope to get a clue regarding the type of genes these silencers affect.

We also used the above-mentioned approach on certain non-biological databases and compared the results obtained from this approach with those of clustering.

3 Methods

We obtained data from experiments conducted at Stanford by Eisen et al. and DeRisi et al [3,4,5]. We took a cluster of genes responsible for ATP Synthesis and Protein Synthesis and obtained their gene expression values along with the four silencer genes (whose effect was unknown) under various experimental conditions [3].

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 our approach we are able to predict the effect of these silencers on some of the genes as well the relationship among the genes itself (see results).

For non-biological databases we obtained data from UCI Machine Learning Repository. The data obtained was about breast cancer, CPU-performance and automobile mileage.

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. For non-biological data sets the discretization was done in accordance with interpretation required. This discretization was done automatically using the written software. This software also formatted the data into the format required by the program. A finer level of discretization (or supporting the real values) would have been more appropriate, but the used approach also gave much of the useful results.

4 Results and Interpretation

4.1 Biological Data Sets

We considered genes whose functions are well known (Protein and ATP synthesis) and also took four silencer genes whose interaction is unknown. We got the expression data of these genes from the various experiments [3,4,5]. Then, we ran the MI program (that uses mutual information based on entropy to extract association rules) on the data set and obtained results. The complete results are given in the Appendix. Most of the relations have high mutual information values. This is expected since all the genes had similar function. From the data set we can infer relations regarding the various genes.

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 we can interpret that when YFR009W, YMR282C are turned on then YDR211W, YDR283C are turned off. We can conclude that there is a negative correlation between these pairs. As another example, consider a result from the *Cdc 15* experiment (ATP Synthesis):

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

We observe that information value is unusually high. Further, from the entropy values we can conclude 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

We observe that mutual information is 1. Further, from the entropy values we can conclude that these genes have positive correlation.

We can also draw 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 we can conclude 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 we can infer that when YDR363W is turned on,
 YHR154W and YJL076W are turned off.

To draw some inferences on a global scale, let us consider the following table that was obtained from the results:

ATP synthesis genes and silencers

	YMR219W	YDR363w	YHR154W	YJL076W
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	0.236-0.592 13 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

The entry in each cell 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. We can also observe that in these experiments YMR219W seems to be a major player as far as silencing is concerned.

Protein synthesis genes and silencers

	YMR219W	YDR363W	YHR154W	YJL076W
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	0.258-0.577 ~30
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 cases	0.317-0.514 10 cases
Cold experiments	No case	0.918-0.918 2 cases	No case	0.918-0.918 2 cases

The entry in each cell 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.

Possible caveats:

- discretizing the data,
- Bigger size relations seem to have larger MI. This is expected since all the genes (except silencers) in the experiment have similar function. To ameliorate this we can try to see if there is a significant difference in the MI for a relation of given size.

4.1 Non-Biological Data Sets

For this kind of data we have been able to run the same program that generates the association rules. Three different databases have been considered. The process of association rules generation and the results are given below.

1. *CPU-performance* database.

Process:

- 1) Get data (file: *machine.data*): 10 attributes, 209 samples.
- 2) Remove unique attributes (IDs).
Here, *model_name* attribute has been removed (so, we are left with 9 attributes and 209 samples).
- 3) Randomly discretize multiple-valued string attributes. Here, we did it as follows:

Attribute	0	1
<i>vendor_name</i>	HP	Not HP

- 4) Discretize real-valued attributes based on their average values (which is (maximum attribute value + minimum attribute value) / 2):

Attribute	0	1
<i>MYCT</i>	17 – 758.5	758.5 – 1500
<i>MMIN</i>	64 – 16032	16032 – 32000
<i>MMAX</i>	64 – 32032	32032 – 64000
<i>CACH</i>	0 – 128	128 – 256
<i>CHMIN</i>	0 – 26	26 – 52
<i>CHMAX</i>	0 – 88	88 – 176
<i>PRP</i>	6 – 578	578 – 1150
<i>ERP</i>	15 – 626.5	626.5 – 1238

- 5) Run the program to generate association rules using *mutual information based on entropy*. The following output has been obtained (where, MI = mutual information value based on entropy, E's are the entropy values):

Correlation-set	MI	$E(a^b)$	$E(ab)$	$E(a^b)$	$E(ab)$
{vendor_name, MYCT}	(0.004)	0.164	0	0.156	0.294
{vendor_name, MMAX}	(0)	0.164	0	0.074	0.109
{vendor_name, CACH}	(0.003)	0.164	0	0.125	0.237
{vendor_name, CHMIN}	(0)	0.164	0	0.074	0.109
{vendor_name, CHMAX}	(0.002)	0.164	0	0.093	0.164
{vendor_name, PRP}	(0)	0.164	0	0.074	0.109
{vendor_name, ERP}	(0)	0.164	0	0.074	0.109
{MYCT, MMAX}	(0.002)	0.137	0.109	0.294	?
{MYCT, CACH}	(0.007)	0.186	0.237	0.294	?
{MYCT, CHMIN}	(0.002)	0.137	0.109	0.294	?
{MYCT, CHMAX}	(0.004)	0.156	0.164	0.294	?
{MYCT, PRP}	(0.002)	0.137	0.109	0.294	?
{MYCT, ERP}	(0.002)	0.137	0.109	0.294	?
{MMAX, CACH}	(0.022)	0.093	0.21	0.064	0.064
{MMAX, CHMIN}	(0.013)	0.048	0.088	0.088	0.037
{MMAX, CHMAX}	(0.030)	0.061	0.129	0.064	0.064
{MMAX, PRP}	(0.137)	0.027	0	0	0.109
{MMAX, ERP}	(0.137)	0.027	0	0	0.109
{CACH, CHMIN}	(0.006)	0.1	0.088	0.224	0.037
{CACH, CHMAX}	(0.013)	0.112	0.129	0.21	0.064
{CACH, PRP}	(0.022)	0.093	0.064	0.21	0.064
{CACH, ERP}	(0.022)	0.093	0.064	0.21	0.064
{CHMIN, CHMAX}	(0.030)	0.061	0.129	0.064	0.064
{CHMIN, PRP}	(0.013)	0.048	0.088	0.088	0.037
{CHMIN, ERP}	(0.013)	0.048	0.088	0.088	0.037
{CHMAX, PRP}	(0.030)	0.061	0.064	0.129	0.064
{CHMAX, ERP}	(0.030)	0.061	0.064	0.129	0.064
{PRP, ERP}	(0.137)	0.027	0	0	0.109

Results:

By observing the output from the program we can see that a few relationships between the attributes had high values of mutual information. Namely, the highest MI-value was observed for:

- 1) *MMAX* (maximum main memory in kilobytes) and *PRP* (published relative performance). Further, by observing the entropy values we conclude that there are very few models that have high *MMAX* and low *PRP*, or high *PRP* and low *MMAX*.
- 2) *MMAX* and *ERP* (estimated relative performance). By observing the entropy values we conclude that there are very few models that have high *MMAX* and low *ERP*, or high *ERP* and low *MMAX*.

- 3) *PRP* and *ERP*. By observing the entropy values we conclude that there are very few models that have high *PRP* and low *ERP*, or high *ERP* and low *PRP*.

These results confirm our intuition (and knowledge) about the relationships of the described attributes.

We have also run clustering on this data set (without data discretization). The clustering tree obtained is given below:

`(vendor_name,(MYCT,(((MMIN,(MMAX,(PRP,ERP))), (CACH,CHMIN)),CHMAX)))`

We can see that while clustering gives us the degree of relations between attributes, it does not give us the explicit relation. With association rules, we are able to tell the exact relation between such attributes. For example, clustering gives us that *PRP* and *ERP* are closely related, but the association rules also give us the relation as described above. At the same time, we do not have to forget that in obtaining association rules we use a “coarse” discretization of data before actually searching for the rules, which can also give us some inaccuracy at the association rules obtained.

2. Breast-cancer-Wisconsin database.

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):

Attribute	0	1
<i>Clump_Thickness</i>	1 – 5.5	5.5 – 10
<i>Uniformity_of_Cell_Size</i>	1 – 5.5	5.5 – 10
<i>Uniformity_of_Cell_Shape</i>	1 – 5.5	5.5 – 10
<i>Marginal_Adhesion</i>	1 – 5.5	5.5 – 10
<i>Single_Epithelial_Cell_Size</i>	1 – 5.5	5.5 – 10
<i>Bare_Nuclei</i>	1 – 5.5	5.5 – 10
<i>Bland_Chromatin</i>	1 – 5.5	5.5 – 10
<i>Normal_Nucleoli</i>	1 – 5.5	5.5 – 10
<i>Mitoses</i>	1 – 5.5	5.5 – 10
<i>Class</i>	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):

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, Class}	(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, Class}	(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, Class}	(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, Class}	(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, Class}	(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	0.102	0.482	0.16
{Bare_Nuclei, Class}	(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

{ <i>Bland_Chromatin, Class</i> }	(0.306)	0.412	0.419	0.068	0.458
{ <i>Normal_Nucleoli, Mitoses</i> }	(0.041)	0.274	0.096	0.429	0.165
{ <i>Normal_Nucleoli, Class</i> }	(0.278)	0.415	0.428	0.082	0.451
{ <i>Mitoses, Class</i> }	(0.06)	0.406	0.522	0.025	0.207

Results:

From the obtained output we have been especially interested in the relations that contain *Class* attribute that determines to which class each sample belonged. Here, we were lucky since 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 *Uniformity_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 we conclude 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).

We have also run clustering on this data set (without data discretization). 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)

3. *Auto-mpg* database.

Process:

- 1) Get data (file: *auto-mpg.data*): 9 attributes, 398 samples.
- 2) Remove unique attributes (IDs).
Here, *car_name* attribute has been removed.
- 3) Remove those samples (total 5) that contain “?” (missing data) as a value for some of their attributes (so, we are left with 8 attributes and 393 samples).
- 4) Discretize real-valued attributes based on their average values (which is (maximum attribute value + minimum attribute value) / 2)

Attribute	0	1
<i>mpg</i>	9 – 27.8	27.8 – 46.6
<i>cylinders</i>	3 – 5.5	5.5 – 8
<i>displacement</i>	68 – 261.5	261.5 – 455
<i>horsepower</i>	46 – 138	138 – 230
<i>weight</i>	1613 – 3376.5	3376.5 – 5140
<i>acceleration</i>	8 – 16.4	16.4 – 24.8
<i>model_year</i>	70 – 76	76 – 82

<i>origin</i>	1, 2	3
---------------	------	---

- 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):

Correlation-set	MI	E(^a^b)	E(^ab)	E(a^b)	E(ab)
{ <i>mpg, cylinders</i> }	(0.286)	0.492	0.514	0.517	0.067
{ <i>mpg, displacement</i> }	(0.139)	0.519	0.503	0.52	0.022
{ <i>mpg, horsepower</i> }	(0.132)	0.508	0.482	0.521	?
{ <i>mpg, weight</i> }	(0.210)	0.531	0.527	0.521	?
{ <i>mpg, acceleration</i> }	(0.023)	0.507	0.480	0.408	0.411
{ <i>mpg, model_year</i> }	(0.117)	0.527	0.518	0.212	0.5
{ <i>mpg, origin</i> }	(0.138)	0.482	0.43	0.311	0.472
{ <i>cylinders, displacement</i> }	(0.351)	0.488	0	0.478	0.504
{ <i>cylinders, horsepower</i> }	(0.291)	0.488	0	0.501	0.482
{ <i>cylinders, weight</i> }	(0.455)	0.489	0.022	0.414	0.526
{ <i>cylinders, acceleration</i> }	(0.047)	0.510	0.503	0.531	0.358
{ <i>cylinders, model_year</i> }	(0.036)	0.452	0.529	0.511	0.466
{ <i>cylinders, origin</i> }	(0.328)	0.441	0.530	0.519	0.135
{ <i>displacement, horsepower</i> }	(0.583)	0.322	0.022	0.180	0.480
{ <i>displacement, weight</i> }	(0.472)	0.393	0.289	0.080	0.497
{ <i>displacement, acceleration</i> }	(0.147)	0.530	0.531	0.499	0.067
{ <i>displacement, model_year</i> }	(0.051)	0.515	0.516	0.441	0.295
{ <i>displacement, origin</i> }	(0.212)	0.531	0.531	0.504	?
{ <i>horsepower, weight</i> }	(0.424)	0.386	0.345	0.039	0.478
{ <i>horsepower, acceleration</i> }	(0.154)	0.527	0.531	0.480	0.022
{ <i>horsepower, model_year</i> }	(0.065)	0.52	0.507	0.427	0.240
{ <i>horsepower, origin</i> }	(0.177)	0.528	0.531	0.482	?
{ <i>weight, acceleration</i> }	(0.043)	0.531	0.522	0.504	0.26
{ <i>weight, model_year</i> }	(0.027)	0.508	0.527	0.459	0.383
{ <i>weight, origin</i> }	(0.23)	0.523	0.531	0.525	0.054
{ <i>acceleration, model_year</i> }	(0.015)	0.527	0.523	0.390	0.491
{ <i>acceleration, origin</i> }	(0.016)	0.524	0.466	0.461	0.438
{ <i>model_year, origin</i> }	(0.007)	0.524	0.408	0.525	0.486

Results:

By observing the output from the program we can see that a few relationships between the attributes had high values of mutual information. Namely, the highest MI-values were obtained for:

- 1) *displacement* and *horsepower*. Further, by observing the entropy values we may notice that there are very few cars that have small *displacement* and high *horsepower*.

- 2) *displacement* and *weight*. Further, by observing the entropy values we may notice that there are very few cars that have large *displacement* and light *weight*.
- 3) *cylinders* and *weight*. Further, by observing the entropy values we may notice that there are very few cars that have small number of *cylinders* and heavy *weight*.
- 4) *horsepower* and *weight*. Further, by observing the entropy values we may notice that there are very few cars that have large *horsepower* but heavy *weight*.

These results confirm our intuition (and knowledge) about the relationships of the described attributes.

We have also run clustering on this data set (without data discretization). The clustering tree obtained is given below:

```
((((mpg,(((cylinders,displacement),weight),horsepower)),origin),acceleration),model_year)
```

5 Contributions

We believe the approach we proposed (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 we used to predict the relationships among the silencer genes and other genes could be extended to genes of unknown function.

6 Future Work

Presently we have studied the problem in which the attributes can take only binary values. It would be more useful to study similar problem with the multi-valued and real valued attributes. The software [1] 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] Asok Tiyagura, "Mining Association Rules Based on Mutual Information". M.S. thesis dissertation, Iowa State University, 1999.
- [2] Rakesh Agrawl, Tomasz Imielinski, 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.
- [3] 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.
- [4] 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):680-6.
- [5] Spellman et al., "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization". *Mol. Biol. Cell Online*, Vol. 9, Issue 12,3273-3297, December 1998.
- [6] UCI Machine Learning Repository,
<http://www.ics.uci.edu/~mlearn/MLRepository.html>, December 13, 2000.
- [7] CPU-Performance Data,
<ftp://ftp.ics.uci.edu/pub/machine-learning-databases/cpu-performance/>, December 13, 2000.
- [8] Breast Cancer Data,
<ftp://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin/>,
December 13, 2000.
- [9] Automobile Characteristics Data,
<ftp://ftp.ics.uci.edu/pub/machine-learning-databases/auto-mpg/>, December 13, 2000.

Appendix

Appendix can be found at: <http://www.cs.iastate.edu/~neeraj/projectfiles/appendix>.