Semantic Mining and Analysis of Gene Expression Data

Xin Xu **Gao Cong** **Beng Chin Ooi** **Kian-Lee Tan** **Anthony K. H. Tung**[*]

School of Computing, National University of Singapore
3 Science Drive 2, Singapore 117543
{xuxin, conggao, ooibc, tankl, atung}@comp.nus.edu.sg
[*]Contact Author

Abstract

Association rules can reveal biological relevant relationship between genes and environments / categories. However, most existing association rule mining algorithms are rendered impractical on gene expression data, which typically contains thousands or tens of thousands of columns (gene expression levels), but only tens of rows (samples). The main problem is that these algorithms have an exponential dependence on the number of columns. Another shortcoming is evident that too many associations are generated from such kind of data. To this end, we have developed a novel depth-first row-wise algorithm FARMER [2] that is specially designed to efficiently discover and cluster association rules into *interesting rule groups* (*IRGs*) that satisfy user-specified minimum support, confidence and chi-square value thresholds on biological datasets as opposed to finding association rules individually. Based on FARMER, we have developed a prototype system that integrates semantic mining and visual analysis of *IRGs* mined from gene expression data.

1 Introduction

Recent studies have shown that association rules can reveal the relationship between genes and environments / categories. For example, they help identify gene predictors for cancer diagnosis. In addition to their simplicity and ease of interpretation, association rules show much promise in the analysis of gene expression data.

However, gene expression data has a large number of columns which poses a great challenge for existing rule mining algorithms, since their basic approaches are the column-wise enumerations where combinations of columns are tested incrementally to search for frequent occurrences of certain combinations. Column-wise association rule mining algorithms generally have the following three problems on gene expression data:
Problem 1: *Extremely long running time due to the huge column enumeration space*,
Problem 2: *Too many association rules found due to the combinatorial explosion of frequent* itemsets, and
Problem 3: *No support of semantic navigation of the huge number of association rules for biologists.*

To address the first 2 problems, we propose a novel row-wise depth-first algorithm FARMER [2] that mines all the *interesting rule groups* (*IRGs*) satisfying user-specified minimum measure (support, confidence, chi square value) thresholds, instead of finding individual association rules. For the last problem, we introduce visualization technique to effectively interpret and compare the semantics of *IRGs*. The graphic interface enables users to conduct semantic explorations over the *IRGs* and identify the most discriminating *IRGs* rapidly.

In the next section, we will briefly introduce the *IRG* mining process with FARMER. *IRG* visualization techniques will be described in details in Section 3. We will discuss the promising applications of our demo system in Section 4. The description of the demo is given in Section 5. We will conclude our work in Section 6.

2 IRG Mining

To have a rough idea of FARMER [2] and *IRGs*, let's look at a simple example. Suppose there is a two-row discretized dataset, 1:{g_1, g_2, g_3, g_4, g_5, g_6, $Cancer$}, 2: {g_7, g_8, g_9, g_{10}, g_{11}, g_{12}, $\neg Cancer$}, where **item** g_i (i = 1, 2, ..., 12) is the discretized value of the original gene expression level. We could generate 63

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Data Base Endowment. To copy otherwise, or to republish, requires a fee and/or special permission from the Endowment.

Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004

association rules in the form of "$A \rightarrow Cancer$" from the same row set $\{1\}$, where A is any combination of g_1, g_2, ..., g_6, and 63 association rules in the form of "$B \rightarrow \neg Cancer$" from the same row set $\{2\}$, where B is any combination of g_7, g_8, ..., g_{12}. Obviously, many of them are redundant.

For the above example, FARMER utilizes the following two core techniques.

- *Mining Interesting Rule Groups*: All the above 126 rules of the running example belong to two **rule groups**. One *rule group* is identified with a unique *antecedent support set* [1] $\{1\}$, a unique *upper bound rule* $g_1g_2g_3g_4g_5g_6 \rightarrow Cancer$, and 6 *lower bound rules* $g_i \rightarrow Cancer$, $i = 1, 2, ..., 6$. The other *rule group* is identified with another *antecedent support set* $\{2\}$, a unique *upper bound rule* $g_7g_8g_9g_{10}g_{11}g_{12} \rightarrow \neg Cancer$, and 6 *lower bound rules* $g_i \rightarrow \neg Cancer$, $i = 7, 8, ..., 12$. The rules between the *upper bound rule* and the *lower bound rules* are the remaining members of the corresponding *rule group*. In this way, we only need to generate 2 *upper bound rules* and 12 *lower bound rules* instead of all the 126 rules. As can be seen, the rules in the same *rule group* share the same *antecedent support set* and the same consequent, thus the same support, confidence and chi square values. From this point of view, the *rule group* is a lossless compression of the association rules. FARMER only outputs **interesting rule groups (IRGs)**. For two *rule groups* of the same consequent, rg_1 and rg_2, if $rg_1.upperbound \subset rg_2.upperbound$ and rg_1 has a higher confidence, then FARMER only outputs rg_1, because rg_1 is defined to be more interesting.

- *Row Enumeration Combined with Efficient Pruning Strategies*: As the row enumeration space is orders smaller than the column enumeration space in gene expression data, FARMER performs search by a depth-first traversal of a **row enumeration tree**. Each node corresponds to a certain row enumeration, where a **transposed table** is set up and a new *IRG* may be identified. For the simple example, the *row enumeration tree* without applying pruning strategies is shown in Figure 1. The traversal starts from the root node $\{\}$, goes through node $\{1\}$ and node $\{1, 2\}$ in sequence, and ends at node $\{2\}$. Figure 2 lists the corresponding three non-empty transposed tables, where $R(g_i)$ represents the complete set of rows that contain item g_i. In this way, the *upper bound rule* $g_1g_2g_3g_4g_5g_6 \rightarrow Cancer$ is discovered at node $\{1\}$, and the *upper bound rule* $g_7g_8g_9g_{10}g_{11}g_{12} \rightarrow \neg Cancer$ is discovered at node $\{2\}$. To avoid redundancy and to comply with the minimum measure thresholds, efficient pruning strategies are applied to further speed up the mining process.

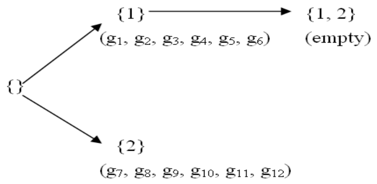


Figure 1: Row Enumeration Tree

g_i	$R(g_i)$
g_1	1
g_2	1
g_3	1
g_4	1
g_5	1
g_6	1
g_7	2
g_8	2
g_9	2
g_{10}	2
g_{11}	2
g_{12}	2

(a) $TT|_{\{\}}$

g_i	$R(g_i)$
g_1	1
g_2	1
g_3	1
g_4	1
g_5	1
g_6	1

(b) $TT|_{\{1\}}$

g_i	$R(g_i)$
g_7	2
g_8	2
g_9	2
g_{10}	2
g_{11}	2
g_{12}	2

(c) $TT|_{\{2\}}$

Figure 2: Transposed Tables

According to our experiments, FARMER is orders of magnitude faster than CHARM [4] and Bayardo's algorithm [1], two well-known column-wise mining algorithms on several bench mark gene expression datasets, as shown in [2].

3 IRG Visualization

Figures 3, 4, and 5 show our system interfaces. We ran the system on the Colon Tumor [2] dataset for demonstration purpose here. We split the original dataset to 47 training samples and 15 test samples randomly. The training dataset consists of 47 rows representing the tissue samples of patients and 2000 columns representing the expression levels of various genes.

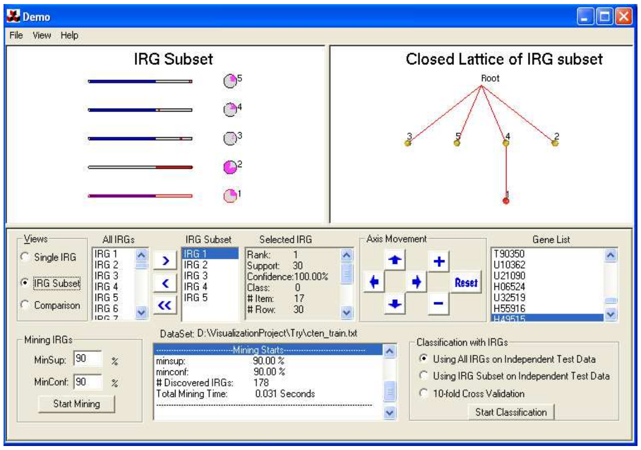
The *IRGs* are sorted based on their rank (descending) as evaluated first by confidence (descending), next by support (descending), and last by # item (ascending). The top 5 *IRGs* ($IRG_1 \prec IRG_2 \prec IRG_3 \prec IRG_4 \prec IRG_5$) are specified as the **IRG subset**. Meanwhile the order of the items in the specified *IRG* subset and the rows in the dataset are determined based on their memberships in the *itemsets*[3] and *antecedent support sets* of the *IRGs* respectively. An item i will be ranked higher than an item j if the highest ranked *IRG* that contain i is above the highest ranked *IRG* that contain j in the *IRG* ranking. Likewise, a row r will have a higher rank than a row s if the highest ranked *IRG* that is matched by r is above the highest

[1] The *antecedent support set* of a rule is the complete set of rows that contain the antecedent of the rule

[2] http://microarray.princetion.edu/oncology/affydata/index.html

[3] the *itemset* of an IRG is the complete set of items that appear in at least one of the antecedents of the association rules in the IRG

ranked *IRG* that is matched by s based on the *IRG* ranking.



All IRGs: IRG 1, IRG 2, IRG 3, IRG 4, IRG 5, IRG 6, IRG 7

IRG Subset: IRG 1, IRG 2, IRG 3, IRG 4, IRG 5

```
Rank:       1
Support:    30
Confidence:100.00%
Class:      0
# Item:     17
# Row:      30
```

Gene List: T90350, U10362, U21090, H06524, U32519, H55916, H49515

MinSup: 90 %

MinConf: 90 %

DataSet: D:\VisualizationProject\Try\cten_train.txt

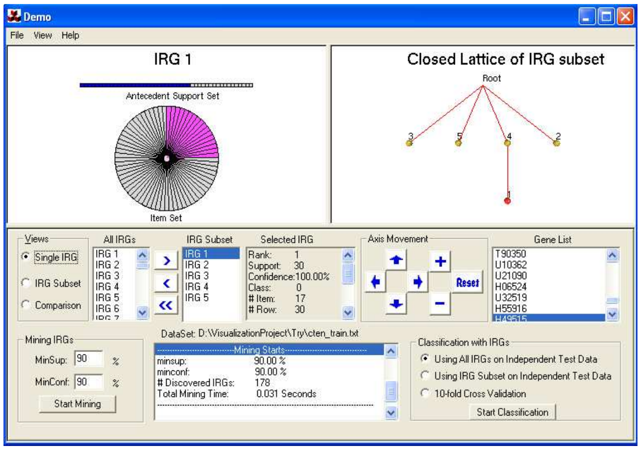
```
------------------------------Mining Starts------------------------------
minsup:                     90.00 %
minconf:                    90.00 %
# Discovered IRGs:          178
Total Mining Time:          0.031 Seconds
-------------------------------------------------------------------------
```

● Using All IRGs on Independent Test Data

○ Using IRG Subset on Independent Test Data

○ 10-fold Cross Validation

Figure 3: Semantic Visualization of the *IRG* Subset Using the Barcode View and the Flower View



All IRGs: IRG 1, IRG 2, IRG 3, IRG 4, IRG 5, IRG 6, IRG 7

IRG Subset: IRG 1, IRG 2, IRG 3, IRG 4, IRG 5

```
Rank:       1
Support:    30
Confidence:100.00%
Class:      0
# Item:     17
# Row:      30
```

Gene List: T90350, U10362, U21090, H06524, U32519, H55916, H49515

MinSup: 90 %

MinConf: 90 %

DataSet: D:\VisualizationProject\Try\cten_train.txt

```
------------------------------Mining Starts------------------------------
minsup:                     90.00 %
minconf:                    90.00 %
# Discovered IRGs:          178
Total Mining Time:          0.031 Seconds
-------------------------------------------------------------------------
```

● Using All IRGs on Independent Test Data

○ Using IRG Subset on Independent Test Data

○ 10-fold Cross Validation

Figure 4: Semantic Visualization of a Single *IRG* Using the Barcode View and the Flower View

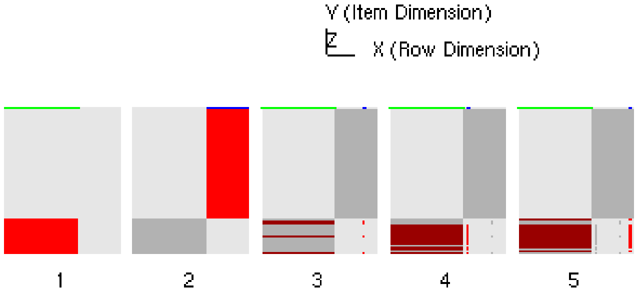


Figure 5: *IRG* Comparisons Using the Matrix View

For each *IRG*, we can visualize its *antecedent support set* and its *itemset* with a "**barcode**" and a "**flower**" separately, or with a "**matrix**" jointly. A "**closed lattice**" graph is also proposed to summarize the *IRGs* in the *IRG* subset based on the subset/superset relationship of their *antecedent support sets*.

- *Antecedent Support Set Visualization*: The "**barcode**" (left hand of Figures 3 and 4) is the identification number of the *IRG*. The "bar" consists of several small grids, each mapping to one ordered row of the dataset. If the mapped row is a member of the IRG's *antecedent support set*, the grid is dyed according to the class label of the row (i.e., red for "negative", blue for "positive"). In this way, the semantics of the *IRG*, like support and confidence, can be obtained by a snapshot. The overall "barcode" view (left hand of Figure 3) suggests that the *antecedent support set* of IRG_1 occupies only the "negative" tissue samples (all red, no blue), while the *antecedent support set* of IRG_2 occupies only the "positive" tissue samples (all blue, no red). They are the only two *IRGs* of confidence 100% in the *IRG* subset. The "**closed lattice**" (right hand of Figures 3 and 4) is another summarization based on the superset/subset relationships of the *antecedent support sets* of *IRGs* in the *IRG* subset. Each node in the lattice except the root node maps to the *antecedent support set* of one *IRG* in the *IRG* subset. The *antecedent support set* of the parent node includes that of the child node. The root node corresponds to the set of all the 47 rows.

- *Itemset Visualization*: We visualize the *itemset* of the *IRG* in the user-specified *IRG* subset as a "**flower**" (left hand of Figures 3 and 4). Each "flower" corresponds to the same set of ordered items that appear in the *IRG* subset and each item is represented by a "petal" of the "flower". The "petal" is dyed if the corresponding item appears in the current *IRG*, otherwise it is left blank.

- *Joined Visualization*: The x-dimension of the "matrix" represents the set of rows in the dataset while the y-dimension of the "matrix" represents the set of items in the *IRG* subset. The items and rows along each dimension are ordered. Given a "matrix" representing a *rule group* IRG_i, a cell valued (x, y) in the "matrix" will be colored red if item y is in the antecedent of the *upper bound rule* for IRG_i and row x matches the *upper bound rule* of IRG_i. Due to the ordering of the items and rows, the red cells in the "matrix" of the highest ranked *IRG* (i.e. IRG_1) will always be clustered at the bottom left corner of the "matrix" as can be seen from Figure 5.

To compare IRG_i against other higher ranked *IRGs*, a cell in the "matrix" for IRG_i will be colored dark grey if it has been colored red in any "matrix" of higher ranked *IRGs*. For example, the dark grey patch in the "matrix" of IRG_2 indicates that these cells have been colored red in the "matrix" of IRG_1. In the case in which the cell also has to be painted red to represent IRG_i, the color of dark red will be used to paint the cell. Finally, the top most cells in each "matrix" are used to represent the class labels of the corresponding rows. By looking at the highest cells in the "matrix" of IRG_1, we can see that IRG_1 has a 100% confidence prediction for a certain class. Overall, we can see that IRG_1 and IRG_2 are the most discriminating *IRGs* with the largest number of non-overlapped red cells.

4 IRG Application

With the effective visualization techniques in Section 3, we can identify the most discriminating *IRGs*, which can be of great value in understanding the mechanics of disease and identifying new pathways by describing what genes are expressed as a result of certain cellular environments.

One promising application of *IRG* is disease diagnosis. As an example, 14 out of the 15 colon tumor test samples have been classified correctly using only the *upper bound rules* of IRG_1 and IRG_2. In [2], we made a first try to build a simple classifier by aggregating the discriminating powers of the *upper bound rules* of *IRGs* on five benchmark gene expression datasets. The simple classifier is competitive with SVM as well as being efficient.

5 Description of the Demo

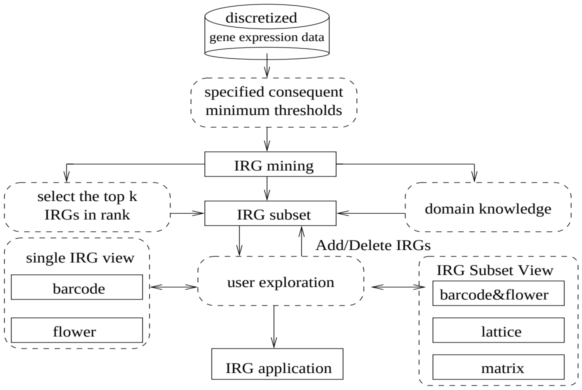


Figure 6: System Framework

In this demo, we will demonstrate an interactive prototype system that specifically involves the following three components (Figure 6).

• *IRG Mining*: *For each user-specified consequent, mine IRGs that satisfy user-specified minimum measure (support, confidence and chi square value) thresholds.*

• *IRG Exploration*: *Users select/adjust the IRG subset of interest, analyze and compare the IRGs in the IRG subset interactively.*

• *IRG Application*: *Output the most discriminating IRGs for disease diagnosis and so on.*

We will showcase (1) how a user can interact with the system with the specified minimum measure thresholds and how the system can find *IRGs* efficiently with FARMER; (2) how the *IRG* summarizes the set of association rules effectively; (3) how the semantics and the discriminating powers of the discovered *IRGs* can be interpreted and compared using our visualization techniques effectively and efficiently; and (4) how the discovered *IRGs* can be used to build an accurate rule-based classifier.

6 Conclusion

In this paper, we used the concept of *IRG* so that numerous rules discovered from gene expression data are clustered into limited number of *IRGs* that encapsulate the complete information about the set of globally significant rules and that we avoid generating billions of redundant rules. From another point of view, *IRGs* could be considered as clusters of emerging patterns [3], an important concept for discovering significant rules from bio-medical data.

Our prototype system not only finds the discriminating associations completely and efficiently, but also provides an interactive graphic interface to identify the associations of the highest biological meanings. Furthermore, it shows great promise in the clinical application, i.e., disease diagnosis.

References

[1] R. J. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining on large, dense data sets. In *Proc. 1999 Int. Conf. Data Engineering (ICDE'99)*.

[2] G. Cong, Anthony K. H. Tung, X. Xu, F. Pan, and J. Yang. Farmer: Finding interesting rule groups in microarray datasets. *In the 23rd ACM SIGMOD International Conference on Management of Data*, 2004.

[3] G. Dong, X. Zhang, L. Wong, and J. Li. Caep: Classification by aggregating emerging patterns. In *Proc. 2nd Int. Conf. Discovery Science (DS'99)*.

[4] M. J. Zaki and C. Hsiao. Charm: An efficient algorithm for closed association rule mining. In *Technical Report 99-10*, Computer Science, Rensselaer Polytechnic Institute, 1999.