International Journal of Computer Science and Informatics

Volume 3 | Issue 4 Article 4

April 2014

ASSOCIATION RULE MINING FOR GENE EXPRESSION DATA

O. V. KALE
Department of Computer Science & Engineering, Walchand College of Engineering, Sangli,
Ompriya.2007@gmail.com

B. F. MOMIN
Department of Computer Science & Engineering, Walchand College of Engineering, Sangli,
bfmomin@yahoo.com

ASSOCIATION RULE MINING FOR GENE EXPRESSION DATA

O. V. KALE[1] & B. F. MOMIN[2]

[1,2]Department of Computer Science & Engineering, Walchand College of Engineering, Sangli
E-mail: Ompriya.2007@gmail.com, bfmomin@yahoo.com

Abstract - Microarray technology has created a revolution in the field of biological research. Association rules can not only group the similarly expressed genes but also discern relationships among genes. We propose a new row-enumeration rule mining method to mine high confidence rules from microarray data. It is a support-free algorithm that directly uses the confidence measure to effectively prune the search space. Experiments on Leukemia microarray data set show that proposed algorithm outperforms support-based rule mining with respect to scalability and rule extraction.

Keywords - *Gene Expression Data, Data Mining, High Confident Association Rules, Bioinformatics.*

I. INTRODUCTION

One main objective of molecular biology is to develop a deeper understanding of how genes are functionally related and, more specifically, to explain how cells control and regulate the expression of their genes and other cellular functions. Deciphering gene relationships has the potential to assist biomedical research in identifying the underlying cause of a disease and developing specific gene-targeting treatments.

Association rule mining method [2] for mining high confident association rules, which describe interesting gene relationships from microarray data sets. The DNA microarray allows parallel genome-wide gene expression measurements of thousands of genes at a given time, under a given set of conditions, for a cell/tissue of interest. Here concentration is on analyzing perturbation microarrays as they are specifically designed to understand the relationships between genes. Perturbation experiments are based on the rationale that, if a gene or cell is no longer able to function normally, the expression levels of other genes that are functionally related may be altered.

II. GENE EXPRESSION DATA

The gene expression data in microarray are presented in M×N matrix where M is the number of microarray experiments and N being the number of genes. The number of experiments M can range from dozens to thousands. On the other hand, the number of genes N can range from hundred to tens of thousands. In some context, M can be referred to as number of transactions or item sets where each gene represents an item. To add to the complexity of representation, each gene is measured in terms of absolute values. However, biologists are more interested in how gene expression changes under different environments in each respective experiment. Thus, these absolute values are discretized according to some predetermined thresholds and grouped under three different levels, namely unchanged, up regulated and down regulated.

Analysis of these massive genomic data has two important goals: First goal is try to determine how the expression of any particular gene might affect the expression of other genes. Second goal of expression data analysis is try to determine what genes are expressed as a result of certain cellular conditions, e.g. what genes are expressed in diseased cells that are not expressed in healthy cells. In this paper, an attempt has been made to review the novel concepts and techniques proposed for mining association rule from the genomic data have been reviewed.

III. PROBLEM DEFINITION

A formal statement of the AR mining problem [2], [3] is as follows: Let the data set D = $\{t_1, t_2 \ldots . t_n\}$ be a set of n Microarray experiments and let $I = \{i_1, i_2 \ldots . i_m\}$ be the set of all genes (m). Each microarray experiment t consists of a set of genes I from *I*. The aim is to mine all ARs (implications) of the form $I_1 \Rightarrow I_2$ which describe strong relationships between the genes based on the microarray experiments in D. I_1 is referred to as the antecedent itemset and I_2 as the consequent itemset. The strength of an AR is measured by support and confidence and the goal is to identify rules that have a support and confidence greater than the user-specified thresholds minimum support (minsup) and minimum confidence (minconf), respectively.

Definition 1 (Support) : Let $I \subseteq I$ be a set of items from D. The support of an itemset I in D, denoted by σ (I), is the proportion of transactions that contain I

$$\sigma(I) = \frac{No\ of\ transactio\ ns\ containing\ \ I}{No\ of\ transactio\ ns} \quad (1)$$

The support of an AR $I_1 \Rightarrow I_2$ is $\sigma(I_1 \cup I_2)$. If σ (I) >= minsup, then I is a frequent itemset.

Definition 2 (Confidence): The confidence of an AR $I_1 => I_2$ is denoted by conf ($I_1 => I_2$) refers to the strength of the association and is given by

$$\frac{\sigma (I_1 \cup I_2)}{\sigma (I_1)}$$

IV. ASSOCIATION RULE MINING

A. Preprocessing Of Data

A gene expression profile can be seen as a single transaction, and each gene, transcript or protein can be thought as an item. The gene expression data can be considered to be a matrix, denoted as G in real expression numbers, which is shown in Table I. The columns denote different samples or conditions. The rows denote genes. In applying association rules to gene expression data, traditional technique would be to first convert each gene expression data into one of three items, down-regulated, up-regulated, or normal expression, which can be denoted as -1, 1 and 0, respectively, as shown in Table II. This is performed by binning the 2 log of the expression level into the three classes [2] with bounds $\leq -r$, $\geq r$, or in between, where r is a threshold defined by user.

Table I
An Example of Microarray

	Cln3 Exp1	Cln3 Exp2	Clb2 Exp2	Clb2 Exp1	Alpha 0min
YAL001C	0.15		-0.22	0.07	-0.15
YAL002W	-0.07	-0.76	-0.12	-0.25	-0.11
YAL003W	-1.22	-0.27	-0.1	0.23	-0.14
YAL004W	-0.09	1.2	0.16	-0.14	-0.02
YAL005C	-0.6	1.01	0.24	0.65	-0.05

Table II
Converted Microarray

	Cln3 Exp 1	Cln3 Exp 2	Clb2 Exp 2	Clb2 Exp 1	Alpha 0min
YAL001C	1	0	0	1	0
YAL002 W	0	0	0	0	0
YAL003 W	0	0	0	1	0
YAL004 W	0	1	1	0	0
YAL005C	0	1	1	1	0

B. Association Rule Extraction

In this section, we introduce our row-enumeration approach to mining high confident association rules efficiently. This approach addresses the two main shortcomings of AR mining: support pruning and itemset explosion. The main challenge is that no support pruning can take place to reduce the search space. A naive approach would be to grow the entire enumeration tree with no support pruning [3] until no more itemsets can be formed. This would be equivalent to generating all closed itemsets, including those that cannot produce confident rules.

Recently, support-based row-enumeration methods have emerged to facilitate the mining of microarray data. These include FARMER [7], TOPKRGS [11], CARPENTER [4], CHARM [7], CLOSET [10] and RERII [3]. These algorithms effectively prevent itemset explosion by only expanding closed itemsets and enumerating the rows (transactions) rather than the items.

C. Grow Entire Enumeration Tree with no Support Pruning

When applying AR mining to microarray data, each microarray experiment is considered to be a single transaction. Consider a sample transaction set as shown in Table III. We will concentrate on algorithm RERII [3] to provide a strong foundation and motivation for our approach. In RERII [3], each node X in Figure 1 will be represented with a three-element group X = {itemlist, sup, childlist}, where itemlist is the closed pattern corresponding to node X, sup is the number of rows at the node and childlist is the list of child nodes of X. For example, the root of the tree can be represented with {{}, 0, {1, 2, 3, 4, 5, 6, 7, 8}} and the node "12" can be represented with {{1, 2}, 2, {3, 4, 5, 6, 8}}.

Given a node X in the row enumeration tree, we will perform an intersection of the itemlist of node X with the itemlist of all its sibling nodes after X. Each intersection will result in a new node whose itemlist is the intersection, whose sup is X.sup + 1 and whose childlist will be available at next level intersection. And each new node will be intersected with its afterward siblings. In this way, the row enumeration tree will be recursively expanded in a depth-first way. The search space (without support pruning) for the transactions in Table 3 is represented as a row-enumeration tree in Fig. 1a.

When applying AR mining to microarray data, each microarray experiment is considered to be a single transaction. Consider a sample transaction set as shown in table 3. We will concentrate on algorithm RERII [3] to provide a strong foundation and motivation for our approach. In RERII [3], each node X in Fig. 1 will be represented with a three-element group X = {itemlist, sup, childlist}, where itemlist is the closed pattern corresponding to node X, sup is the number of rows at the node and childlist is the list of child nodes of X. For example, the root of the tree can be represented with {{}, 0, {1, 2, 3, 4, 5, 6, 7, 8}} and the node "12" can be represented with {{1, 2}, 2, {3, 4, 5, 6, 8}}.

Given a node X in the row enumeration tree, we will perform an intersection of the itemlist of node X with

the itemlist of all its sibling nodes after X. Each intersection will result in a new node whose itemlist is the intersection, whose sup is X.sup + 1 and whose childlist will be available at next level intersection. And each new node will be intersected with it's afterward siblings. In this way, the row enumeration tree will be recursively expanded in a depth-first way. The search space (without support pruning) for the transactions in Table III is represented as a row-enumeration tree in Fig. 1a.

D. Confidence Pruning

This pruning will remove nodes that cannot generate confident I-spanning rules. This pruning is based on an observation of the row enumeration tree's structure. For each node in the tree, we can predict the maximum support [4] and confidence its corresponding itemset can exhibit based on its location within the tree. It is based on the following definitions.

Definition 3 (Maximum support) : Given a node n with k sibling nodes, the maximum support of the itemset at n, represented as σ_{max} (n) or any of n's potential child nodes is

$$\sigma_{\max}(n) = n \cdot initial\_\sup port + k$$

Table III
Transaction Set

Transaction	Items
1	A B C D E G
2	A C D E G
3	C D E F G H I
4	B C D E G
5	A C E G I
6	A D I
7	D I J
8	A B C D G

Definition 4 (Minimum feature): The item i_1 in the itemset I is the minimum feature if

$$\sigma(i_1) \le \sigma(i_2) \;\Big|\; \forall i_2 \in I$$

Definition 5 (I-spanning rule): Given an itemset I, a rule r is an I-spanning rule if

$$antecedent(r) \cup consequent(r) = I \quad and$$
$$|antecedent(r)| = 1$$

Definition 6 (Maximum confidence): Given a node n with minimum feature i, the maximum confidence of any spanning rule of the itemset at n is

$$Conf_{\max}(n) = \frac{\sigma_{\max}(n)}{\sigma(i)}$$

If $Conf_{max}$ (n) < minconf, then n can be pruned as any further enumeration below the node will only generate less or equally confident child rules. This is because the maximum support of any child node is bounded above by σ_{max} (n) and the support of its minimum feature can only be greater than or equal to the minimum feature of n. Thus, the child node is bounded above by $Conf_{max}$ (n). If the current parent node is not pruned by this approach, it is expanded to form a subtree of child nodes following the approach of RERII [4]. Tree after confidence pruning is shown in Fig. 2.

Rules generated by this approach are shown in Table IV.

Table IV
Association rules (minsup>3 and minconf>4/5)

Association Rules	Confidence	Support
C=>DEG	4/6	4
E=>CDG	4/5	4
G=>CDE	4/4	4
A=>CG	4/5	4
C=>AG	4/6	4
G=>AC	4/6	4
A=>D	4/5	4
B=>CDEG	2/3	2
B=>CDG	3/3	3
I=>D	3/4	3
J=>DI	1/1	1
F=>CDEGHI	1/1	1
H=>CDEFGI	1/1	1

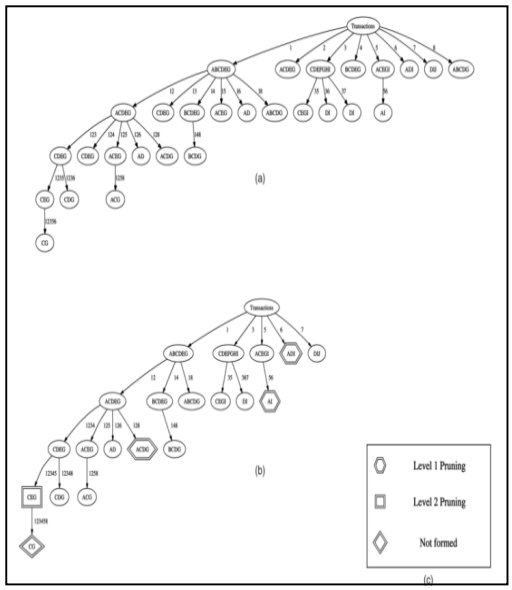
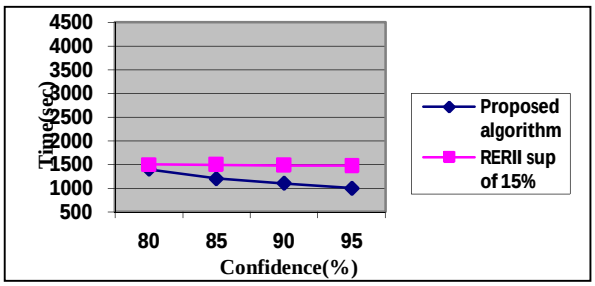


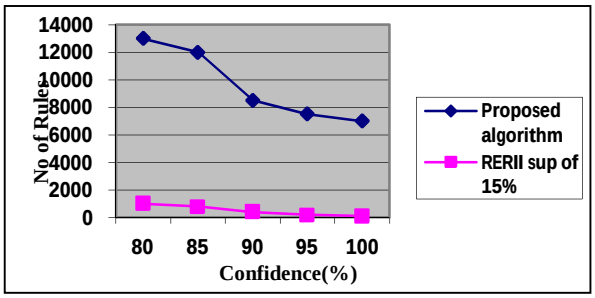
Fig.1. (a) Complete row-enumeration tree (b) Pruned row-enumeration tree. (c) Key.

V. RESULT ANALYSIS

Our experiments are performed on real-life dataset, which is the clinical data on ALL-AML leukemia (ALL). In this dataset, there are 78 tissue samples and each sample is described by the activity level of 12600 genes. Fig. 2 shows the experimental results on this datasets.



(a)



(b)

Fig. 1. Performance on the data set leukemia of RERII with 15% supports and proposed algorithm as confidence is increased (a) Scalability. (b) Number of rules discovered.

REFERENCES

[1] Tara McIntosh and Sanjay Chawla "High-Confidence Rule Mining for Microarray Analysis" IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 4, NO. 4, OCTOBER-DECEMBER 2007.

[2] C. Creighton and S. Hanash, "Mining Gene Expression Databases for Association Rules," Bioinformatics, vol. 19, no. 1, pp. 79-86, 2003.

[3] G. Cong, K.-L. Tan, A. Tung, and F. Pan, "Mining Frequent Closed Patterns in Microarray Data," Proc. Fourth IEEE Int'l Conf. Data Mining (ICDM), vol. 4, pp. 363-366, 2004.

[4] F. Pan, G. Cong, K. Tung, J. Yang, and M. Zaki, "CARPENTER: Finding Closed Patterns in Long Biological Datasets," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 637-642, 2003.

[5] Rakesh Agrawal Tomasz Imielinski Arun Swami, "Mining Association Rules between Sets of Items in Large Databases" IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120.

[6] Gao Cong, Anthony K. H. Tung, Jiong Yang, "FARMER: Finding Interesting Rule Groups in Microarray Datasets" Dept. of Computer Science Natl. University of Singapore.

[7] Mohammed J. Zaki and Ching-Jui Hsiao, "CHARM: An Efficient Algorithm for Closed Association Rule Mining" Computer Science Department Rensselaer Polytechnic Institute, Troy NY 12180.

[8] Tim BeiBbarth and Terence P. Speed, "GOstat: find statistically overrepresented Gene Ontologies within a group of genes" Walter and Eliza Hall Institute of medical Research, 1G Royal Parade, Parkville, Vic 3050,Australia.

[9] G. Cong, K.-L. Tan, A.K. Tung, and X. Xu, "Mining TOP-K Covering Rule Groups for Gene Expression Data," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 670-681, 2005.

[10] J. Pei, J. Han, and R. Mao, "CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets," Proc. ACM SIGMOD Int'l Workshop Data Mining and Knowledge Discovery (DMKD), pp. 21-30, 2000.

◈◈◈