ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Issue 4, April 2016

A NEW APPROACH TO RANK-BASED WEIGHTED ASSOCIATION RULE MINING BASED ON K-SVD ALGORITHM

M.N.SOWMIYA[1], R.KARTHIKEYAN[2]

(Anna Univ Affiliated)M.E Department of Computer Science and Engineering[1]
(Anna Univ Affiliated)HOD Department of Computer Science and Engineering[2]
Mohamed Sathak Engineering College,kilakarai, Ramanathapuram district, TN.
[1]sowmi.rathinam93@gmail.com
[2]karthikhonda77@gmail.com

***Abstract*— The Association Rule Mining is defined as a process of Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transactional databases, relational databases, and other information repositories. This method is commonly used in bioinformatics for the ranking of genes and genomes. Here there is a drawback, which makes the decision maker more confusion due to huge number of evolved rules. To avoid this a weighted association rule mining is propose called RANWAR (or) Rank based Weighted Association Rule Mining which uses the proposed rule interestingness measures, viz., rank-based weighted condensed support (WCS) and weighted condensed confidence (WCC). Based on these measures here assign weight to the each item, which generates less number of frequent item sets than state-of-the-art association rule mining, this process on Gene Expression and Methylation datasets. The resulted genes of the top rules are biologically validated by Gene Ontologies (GOs) and KEGG pathway analyses. The top ranked rules extracted from RANWAR that hold poor ranks in traditional Apriori , are highly biologically significant to the related diseases. Finally, report the top rules evolved from RANWAR that are not in Apriori.**

***Keywords*— Weighted association rule mining, wcs, wcc, Limma gene-weight, gene-ranking, RANWAR.**

I.INTRODUCTION

Gene Ranking is not always admitted in biomedical publications, the ranked lists obtained from univariate analyses should not be considered as fixed universal results. After giving a formal definition of the term 'ranking' as considered in the present article, we briefly review important sources of variations for gene rankings.

The term 'data set' denotes a pair $\mathcal{D} = (\boldsymbol{x}, \boldsymbol{y}),$ where the $n \times p$ matrix $\boldsymbol{x} = (x_{ij})_{i=1,\ldots,n \atop j=1,\ldots,p}$ contains n observations of the random vector $(X_1, \ldots, X_p)'$ (for instance, the expression levels of p genes), and $\boldsymbol{y} = (y_1, \ldots, y_n)$ stores either the response variable Y of interest for these n observations or, e.g. an experimental condition fixed by design. In this article, we define a *ranking* of the variables $X_1, \ldots, Xp$ as a permutation $\boldsymbol{r} = (r_j)_{j=1,\ldots,p}$ of $(1, \ldots, p)$, where r_j is the rank of the variableX_j with respect to its association with Y. A small rank indicates strong association between the considered gene and Y, either positive or negative. A ranking yields an ordered *list* $\boldsymbol{l} = (l_m)_{m=1,\ldots,p}$ defined by$l_m = j \Leftrightarrow r_j = m \quad \text{for all } j, m = 1, \ldots, p.$(1)For instance, in the case of differential gene expression, $r_{1786} = 1$ and $l_1 = 1786$ would mean that X_{1786} is identified as the most differentially expressed gene. If $\boldsymbol{l}$ is an ordered list, the k top genes $l_1, \ldots, l_k$ form the so-called top-k list (usually $k \ll p$). For example, biomedical articles often report top-20 or top-50 lists. Note that it do not consider rankings with ties for the sake of simplicity.

Association rule mining aims to explore large transaction databases for association rules. Classical Association Rule Mining (ARM) model assumes that all items have the same significance without taking their weight into account. It also ignores the difference between the

transactions and importance of each and every itemsets. But, the Weighted Association Rule Mining (WARM) does not work on databases with only binary attributes. It makes use of the importance of each itemset and transaction. WARM requires each item to be given weight to reflect their importance to the user. The weights may correspond to special promotions on some products, or the profitability of different items.

The concept of association rule mining proposes the support-confidence measurement framework and reduced association rule mining to the discovery of frequent item sets. WARM generalizes the traditional model to the case where items have weights. WARM requires for each item to be given weight to reflect their importance to the user. The weights may correspond to the profitability of different items. As more data is gathered, which are frequently getting updated, the construction of the graph should be dynamic instead of static. Using Online Hits algorithm, the graph can be constructed dynamically and the cost can be reduced by postponing updates whenever possible. By calculating Eigen values are enforcing the mutual reinforcement relationship between the items.

Limma is a package for differential expression analysis of data arising from microarray experiments. The package is designed to analyze complex experiments involving comparisons between many RNA targets simultaneously while remaining reasonably easy to use for simple experiments. The central idea is to fit a linear model to the expression data for each gene. The expression data can be log-ratios, or sometimes log-intensities, from two color microarrays or log-intensity values from one channel technologies such as Affymetrix. Empirical Bayes and other shrinkage methods are used to borrow information across genes making the analyses stable even for experiments with small number of arrays . Limma is designed to be used in conjunction with the affy or affyPLM packages for Affymetrix data. With two color microarray data, the marray package may be used for pre-processing. Limma itself also provides input and normalization functions which support features especially useful for the linear modeling approach.

Microarray-based gene expression profiling experiments, which are routine today, allow researchers to identify, for instance, genes differentially expressed (DE) between diseased and normal patient samples or genes that change in expression over time during a treatment. Unfortunately, the steady increase in the amount of data generated in the past decade from such experiments was not paralleled by the evolution of analytical methods used to extract knowledge from such datasets and, therefore, there is a gap between our ability to measure gene expression data and to extract workable knowledge from it.

Since the beginning of the microarray-based expression profiling experiments, researchers were interested in finding common “themes” among the genes identified as differentially expressed between two conditions. For instance the identification of Gene Ontology (GO) terms enriched in differentially expressed genes was used as early as 1999, but became widespread only [GO analysis tools were made available. As biological annotations started to include descriptions of gene interactions in the form of pathways, the identification of the pathways involved in various conditions has emerged as a ubiquitous bioinformatics task.

In general, biological pathways can be divided into gene signaling pathways, and metabolic pathways. Gene signaling pathways are graphs that use nodes to represent genes, or gene products, and edges to represent signals that go from one gene to another. Metabolic pathways are graphs that use nodes to represent biochemical compounds, and edges to describe biochemical reactions that involve such compounds. Since biochemical reactions are usually carried out by enzymes which are coded for by genes, in a metabolic pathway genes are associated with edges rather than nodes. Ideally, a comprehensive pathway analysis method would be able to take into consideration all aspects of the phenomena described by a pathway. These aspects would include the position and role of each gene in a pathway, the types of signals between genes, the efficiency with which a signal travels from one gene to another, or the efficiency with which a certain reaction is carried out, rate limiting conditions, etc. Such methods have been proposed for both signaling pathways, and metabolic pathways, but no method is currently available to analyze both types of pathways taking into consideration all the information available. Hence, even though they do not use all information available, methods that treat the pathways as simple gene sets are still popular because they can be applied equally well to signaling pathways, metabolic pathways, GO terms, as well as arbitrary sets of genes.

II.RELATED WORKS

Report a novel finding that KCC2 is widely expressed in several human cancer cell lines including the cervical cancer cell line (SiHa). Membrane biotinylation assays and immune staining showed that endogenous KCC2 is located on the cell membrane of SiHa cells[3]There is no gene signature for predicting relapse and survival of cervical cancer with early stage currently. investigate whether gene expression profiling of cervical cancer could be used to predict the prognosis of patient[1]. The gene expression profiles of ADC and SCC were downloaded from Gene Expression Omnibus under accession No. GSE10245. Accordingly, differentially expressed genes (DEGs) were identified by the limma package in R language[2]..The classical models ignore the difference between the transactions, and the weighted association rule mining does not work on databases with only binary attributes. introduce a new measure w-support, which does not require preassigned weights. It takes the quality of transactions into consideration using

link-based models. A fast mining algorithm is given, and a large amount of experimental results are presented[11].

III.SYSTEM REPRESENTATION

To propose two novel rule-interestingnes smeasures, viz., rank-based *weighted condensed support* (wcs) and *weighted condensed confidence* (wcc) on basis of independency of the genes of a microarray dataset in *statistical scenario*. Suppose, input boolean matrix BIT is of size m× n,where m denotes #sample and n denotes #gene. The assigned weights to genes are assumed to be W = {w_1,w_2, ...,w_n}. A weight w_i is attached to each gene g_i (i.e., i = 1, 2, ..., n). This denotes a pair of (g_i,w_i) which is stated as a weighted gene. The weight of the g_i gene in the k-th sample/transaction is denoted by w_{ki}, where 1 _ k _ m. If the gene g_i presents in the k-th transaction (s_k), then value of w_{ki} will be the weight of the gene g_i, otherwise, value of w_{ki} becomes zero. In other words,

$$w_{ki} = \begin{cases} w_i, & \text{if } g_i \in s_k. \\ 0, & \text{otherwise.} \end{cases}$$

A.MicroArray Dataset

Microarray technique is a useful tool for measuring gene expression data across different experimental and control samples.In microarray data, it is mandatory to use some pre-filtering process like removal of genes having low variance.differential expression values (t-statistical values, or corresponding p-value in any statistical test) of genes are calculated on the assumption of independency of genes.If all the genes are dependent, then p-values (specified probabilities)of the genes in the test will be invalid. Therefore, on the basis of the independency of the genes of a microarray dataset, determined itemset-transaction weight. The weight of a gene is calculated through p-value ranking of the gene. Thus ,itemset-transaction weight can be defined as multiplication of weights of all the genes (items) of the itemset in a transaction(sample) for a microarray dataset. It is estimated as:

$$W_k(Z) = \prod_{i=1}^{Q}{}_{(\forall g_i \epsilon Z, Q=|Z|)} w_{ki}, \qquad (1)$$

where Wk(Z) denotes itemset-transaction weight of itemset Z for k-th transaction, wki refers to the weight of gene gi for k-the transaction, Q denotes multiplicative operator, Q refers to the total number of genes in the itemset Z. It should be mentioned that for rule mining using the wcs and wcc measures, have to set two thresholds, one for wcs, and other wcc. It is well-known that support is the property of an itemset, and confidence is the property of a rule. Therefore, here, the above statement signifies that if wcs of an itemset is greater than equal to a minimum support threshold (say, min wsupp), then the itemset is frequent and if the itemset is frequent, then wcc values of rules made from the item set need to validate. If wcc of any rule is greater than equal to minimum confidence threshold (say, min wconf), then the rule can be selected.

B. Calculating proposed.WCS and WCC

Calculating the WCC and WCS based on the following formula.

TABLE I: An example of calculating the proposed *wcs* and *wcc*.

Genes (Items) ⟶

Item weight (w_i)	g_1	g_2	g_3	g_4	g_5
	1.0	0.2	0.8	0.6	0.4

Genes (Items) ⟶ / Transactions ↑

	g_1	g_2	g_3	g_4	g_5
s_1	1	1	0	1	1
s_2	0	0	1	1	0
s_3	0	1	1	1	1
s_4	1	1	0	0	1

Suppose, Z is whole itemset of a rule, $\{g1, g2 \Rightarrow g5\}$; Here, antecedent, $A = \{g1, g2\}$; consequent, $C = \{g5\}$; and $Z = A \cup C = \{g1, g2, g5\}$; and
So, $W_1(Z) = w_{11} * w_{12} * w_{15} = w_1 * w_2 * w_5 = 1.00 * 0.2 * 0.4 = 0.08$;
$W_2(Z) = w_{21} * w_{22} * w_{25} = 0 * 0 * 0 = 0$;
$W_3(Z) = w_{31} * w_{32} * w_{35} = 0 * w_2 * w_5 = 0 * 0.2 * 0.4 = 0$;
$W_4(Z) = w_{41} * w_{42} * w_{45} = w_1 * w_2 * w_5 = 1.00 * 0.2 * 0.4 = 0.08$;

$$m'(Z) = max\{\sum_{k=1}^{4} BIT_{k1}, \sum_{k=1}^{4} BIT_{k2}, \sum_{k=1}^{4} BIT_{k5}\} = max\{(1+0+0+1), (1+0+1+1), (1+0+1+1)\} = max\{2,3,3\} = 3$$

$wcs(Z) = \frac{W_1(Z)+W_2(Z)+W_3(Z)+W_4(Z)}{m'(Z)} = \frac{0.16}{3} = 0.053$;
Similarly, $W_1(A) = w_{11} * w_{12} = w_1 * w_2 = 1.00 * 0.2 = 0.2$;
$W_2(A) = w_{21} * w_{22} = 0 * 0 = 0$; $W_3(A) = w_{31} * w_{32} = 0 * w_2 = 0 * 0.2 = 0$;
$W_4(A) = w_{41} * w_{42} = w_1 * w_2 = 1.00 * 0.2 = 0.2$;

$$m'(A) = max\{\sum_{k=1}^{4} BIT_{k1}, \sum_{k=1}^{4} BIT_{k2}\} = max\{(1+0+0+1), (1+0+1+1)\} = max\{2,3\} = 3$$

$wcs(A) = \frac{W_1(A)+W_2(A)+W_3(A)+W_4(A)}{m'(A)} = \frac{0.4}{3} = 0.13$;
Therefore, $wcc(A \Rightarrow C) = \frac{wcs(Z)}{wcs(A)} = \frac{0.053}{0.13} = 0.41$;

The proposed support can be stated as:

$$wcs(Z) = \begin{cases} \dfrac{\sum_{k=1}^{m} W_k(Z)}{m'(Z)}, & \text{if } |Z| > 1 \\ \dfrac{\sum_{k=1}^{m} W_k(Z)}{m}, & \text{if } |Z| = 1 \end{cases}$$

The proposed confidence can be stated as:

$$wcc(A \longrightarrow C) = \frac{wcs(A \cup C)}{wcs(A)} = \frac{wcs(Z)}{wcs(A)}.$$

IV RULE MINING APPROACH

Proposed a rank-based weighted association rule mining (RANWAR) using the two proposed rule-interestingness measures (wcs and wcc). The steps of the

methodology shows that how the genes can be applied on microarray/beadchip data to extract association rules.

A.Prefiltering process

In microarray data, it is mandatory to use some pre-filtering process like removal of genes having low variance. In fact, due to the low variance of the gene, sometime lower p-value is produced which seems to be significant, but actually it is insignificant.Therefore, it is needed to check the overall variance of the data according for each gene and filter out the genes having very low variance. Here, used a matlab function "gene varfilter" by which some user-defined percentile (say, 5 or 10 or 20 percentile) of the genes having low variance can be eliminated from the gene list. The filtered data should be normalized gene-wise as normalization converts the data from different scales into a common scale. There are many normalization methods available.

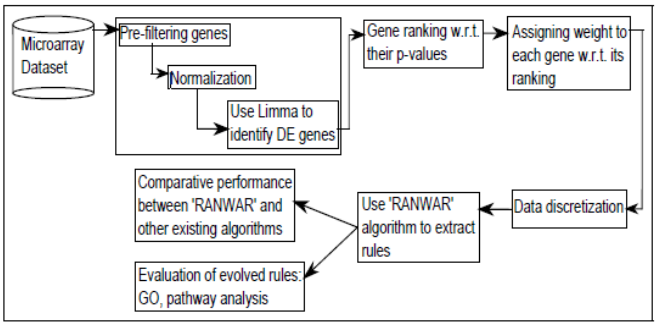


Fig. 1: The Rule mining approach from biological data.

C. Gene weight calculation

The ranges of weight lie in between 0 and 1. Suppose, n is number ofgenes. Thus, the weight of each gene (denoted by wi, 1 _ i _ n) is estimated from a function of the above rank (denoted by ri,1 _ i _ n) and number of genes as described below:

$$w_i = \frac{1}{n} * (n - (r_i - 1)).$$

D. Discretization Process

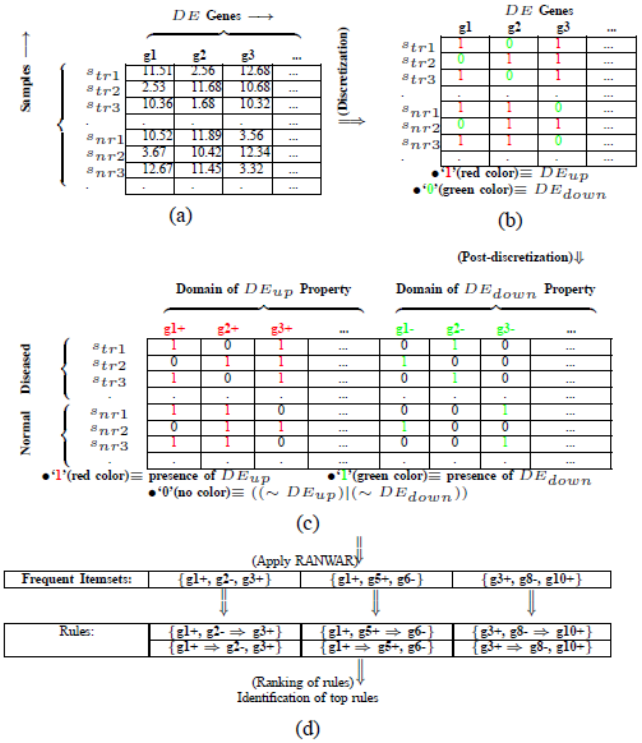
Suppose, I[r, c] is input data matrix. Here, r denotes genes,and c denotes samples. First of all, the matrix I is transposed .Suppose, IT be the resulting matrix.Now, discretization of the input data matrix is mandatory for applying association rule mining.

1.K-SVD Algorithm

The K-SVD algorithm for training of dictionaries.This algorithm is flexible and works in conjunction with any pursuit algorithm.It is simple and designed to be a truly direct generalization of the k -means. The K-SVD is highly efficient, due to an effective sparse coding and a Gauss–Seidel-like accelerated dictionary update method. The algorithm's steps are coherent with each other, both working towards the minimization of a clear overall objective function. While this may seem superfluous, will use the very description of the -means to derive the K-SVD as its direct extension. Here then discuss some of the K-SVD properties and implementation issues.Just like the k-means, the K-SVD algorithm is sussceptible to local minimum traps.

E. Determining Frequent Pattern

After the post-discretization technique as stated in the last subsection, then identify frequent itemsets. For this, at first, evaluate wcs of the 1-itemsets, and then identify the frequent singleton itemsets (i.e., wcs of them are greater than equal to min wsupp).



Thereafter, similarly, calculate their supersets 2-itemsets and then determine frequent 2-itemsets. After that, rules are extracted from the frequent 2-itemsets.Then, wcc of each rule is computed. The rules having wcc greater than min wconf value, are selected for resulting list of rules. Then, determine their supersets 3-itemsets and then determine frequent 3-itemsets, and then extract significant rules from these, and so on.

The algorithm terminates if there is no further successful extensions of frequent itemsets to be identified. Finally, the evolved rules are ranked w.r.t. wcs or wcc.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Issue 4, April 2016

Finally, we report many top ranked rules produced by RANWAR that hold poor ranks in traditional Apriori, but are highly biologically significant to related diseases.

V.ORGINAL DATASETS

Two real datasets are used which are described in Table II.

TABLE II: Information of used Datasets (DS).

DS id	Dataset information	Treated samples	Control samples
1	Genome-wide DNA methylation dataset of Uterine cervical carcinogenesis (**NCBI Ref. id: GSE30760**) with cancerous uterine cervix (CUC) and normal uterine cervix (NUC).	**63 (CUC)**	**152 (NUC)**
2	Gene expression dataset of cigarette smokers of lung adenocarcinoma (**NCBI Ref. id: GSE10072**) with current smoker (CS) and never smoker (NS) for Tumor samples.	**24 (CS)**	**16 (NS)**
3	Expression dataset of Uterine Leiomyoma, belonging Uterine Leiomyoma tumor (UL) and normal myometrial (MM) samples (**NCBI Ref. id: GSE31699**).	**16 (UL)**	**16 (MM)**
4	Methylation dataset of Uterine Leiomyoma having the UL and MM samples (**NCBI Ref. id: GSE31699**).	**18 (UL)**	**18 (MM)**

VI.CONCLUSION AND FUTURE WORKS

In the Real world, there are huge numbers of evolved rules of items (or genes) by Association Rule Mining algorithms makes confusion to the decision maker. Here introduced the two new rank based weighted condensed rule-interesting measures called WCC,WCS. A weighted rule mining algorithm called RANWAR, which has been developed using the measures especially for micro array data. . It saves time of execution of the algorithm. Finally report that some top rules extracted from RANWAR that are not present in Apriori, which have high biological significance. In Further, Using Fuzzy c-means clustering in the Data Discretization process, thus points on the edge of a cluster may be in the cluster to a lesser degree than points in the center of cluster.The comparison results provide the efficient clustering process in RANWAR.

VII.REFERENCES

[1]Saurav Mallik, Anirban Mukhopadhyay and Ujjwal aulik, "RANWAR:RankBased Weighted Association Rule Mining from Gene Expression andMethylation Data" in 2014.

[2]Thomas *et al.*, "Expression profiling of cervical cancers in Indian women at different stages to identify gene signatures during progression of the disease," *Cancer Med.*, vol. 2, no. 6, pp. 836–848, Dec. 2013.

[3]J. Liu *et al.*, "Identifying differentially expressed genes and pathways in two types of non-small cell lung cancer: Adenocarcinoma and squamous cell carcinoma," *Genet. Mol. Res.*, vol. 13, pp. 95–102, 2014.

[4]W. Wei *et al.*, "The potassium-chloride cotransporter 2 promotes cervical cancer cell migration and invasion by an mission transport-independent mechanism," *J Physiol.*, vol. 589, pp. 5349–5359, 2011.

[5]J. Pavon, S. Viana, and S. Gomez, "Matrix Apriori: Speeding up the search for frequent patterns," in *Proc. IASTED, 24th Multi-Conf. Appl. Informat.*, Innsbruck, Austria, 2006.

[6]Y. Hong *et al.*, "Incrementally fast updated frequent pattern trees," *Expert Syst. Appl.*, vol. 34, pp. 2424–2435, 2008.

[7]D. Oguz and B. Ergenc, *Incremental Itemset Mining Based on Matrix Apriori Algorithm*. Berlin/Heidelberg, Germany: Springer, 2012, pp. 192–204.

[8]S. Orlando *et al.*, "Enhancing the apriori algorithm for frequent set counting," in *Data Warehousing and Knowledge Discovery*. Berlin/ Heidelberg, Germany: Springer , 2013, pp. 71–82.

[9]U. Yun *et al.*, "WIP: Mining Weighted Interesting Patterns with a strong weight and/or support affinity," in *Proc. SDM*, 2006, vol. 6, pp. 3477–3499.

[10]K. M. Yu and J. L. Zhou, "A Weighted Load-balancing parallel apriori algorithm for association rule mining," in *Proc. IEEE Int. Conf. Granular Computing (GrC 2008)*, pp. 756–761.

[11]K. Sun and F. Bai, "Mining weighted association rules without preassigned weights," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 4, pp. 489–495, 2008.