Ranking Analysis of Gene Expression and Methylation Data for Identification of Weighted Association Rules

[1]Manju Priya.V,[2]Durai Kumar.D,[3]Balaji.S

[1]Student, [2]Associate Professor & Head, [3]Assistant Professor
Department of Information Technology,
Ganadipathy Tulsi's Jain Engineering College,Vellore-632102,India

***Abstract* - Association rule mining is an interesting topic in data mining and bioinformatics. The huge number of evolved rules by association rule mining algorithms makes confusion to the decision maker. In this paper, three techniques for mining association rules are proposed. For selection of perfectly Differentially Expressed (DE)/ Methylated(DM) genes the p-value and fold change value are used. For assigning weight to each gene the weighted condensed support (wcs) and confidence (wcc) is used along with various network weighting methods which reduces the complexity. For data discretization effective cluster algorithm i.e., Genetic Cluster Algorithm (GCLUS) is used. Thus, it saves time for execution of algorithm. The genes of the top rules are biologically validated by Gene Ontology (GOs) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analyses. Many top ranked rules extracted using these techniques hold poor rank in traditional Apriori algorithm which is significant to treat many diseases.**

***Index Terms* - rule mining, p-value, fold change value, wcs, wcc, network weighting methods, GCLUS**

I. INTRODUCTION

Knowledge Discovery and Data Mining (KDD) is an interdisciplinary domain that mainly focuses on the systematic ways of acquiring interesting rules and patterns from the data. The significant common patterns which are estimated by interestingness measures include association rules. Association Rule Mining (ARM), an important data mining technique is utilized for detecting interesting relationships between items. Huge number of rules always creates problem to select top among them. Therefore, the ranking of rules from the biological data is very important area for research.

The main objective is to analyze the gene expression and methylation data for mining association rules which holds poor rank in traditional Apriori algorithm. RANWAR (RANking analysis of gene expression and methylation data for identification of Weighted Association Rules) makes use of microarray dataset which contains the treated and normal samples of gene expression and methylation data and then pre-filtering process is carried out where low variance data are removed and then they are normalized using zero-mean normalization and the statistical test is applied in order to identify DE genes and then they are ranked and weighted using various weighting techniques and then the data's are discretized in order to extract the rules which holds poor rank in traditional algorithm and these rules are validated using GO[3] and KEGG pathway analyses. The rules which hold poor in traditional Apriori algorithm hold good in RANWAR which is highly significant to treat many diseases.

II. LITERATURE REVIEW

Huge number of rules always creates problem to select top among them. By further investigations, different limitations have been found in the traditional Apriori algorithm, like generation of huge number of frequent item sets, high elapsed time, multiple scan problem, load imbalance problem, importing same importance to each item etc. and for data discretization, the k-means clustering algorithm is used which focuses only on local optima and sensitive to noise and outliers, applicable only when number of clusters and mean value is defined and the standard database GO and KEGG pathway is used, which it is not updated for long time or if it is uncertain. For reducing those shortcomings, different enhancements have been performed on the original simultaneously. Therefore, we have mainly focused how to reduce elapsed time for rule mining in such way only top ranked items and their related highly significant rules is present as result in large transactional database.

III. PROPOSED APPROACH

The RANWAR performs analysis on gene expression and methylation data in order to identify the rule that holds poor rank in traditional Apriori algorithm but which holds good rank when these rules are biologically validated.

Step 1: Finding DE/DM genes and Ranking

Microarray dataset contains the diseased and normal samples information. At first, some pre-filtering process should be applied on the data (viz., 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. Thus, it is needed to check the overall variance of the data according for each gene and filter out the genes having very low variance.

The variance is calculated using: $V = V_g + V_e + V_{ge}$

where V_g is the genetic variation, V_e is the environmental variation and V_{ge} is both. Then the filtered data should be normalized as the normalization converts the data from different scales into a common scale.

The zero-mean normalization is applied: $x_{ij}^{norm} = \frac{x_{ij}-\mu}{\sigma}$

where μ and σ refer to mean and standard deviation of the expression/ methylation data of a gene *i* before normalization respectively and x_{ij} and x_{ij}^{norm} denote the value of *i*-th gene at *j*-th sample before and after normalization, respectively. For identifying DE/DM genes, a suitable non-parametric test should be applied correctly [3]. The Limma should be performed well for both normal and non-normal distributions for all sizes of data.

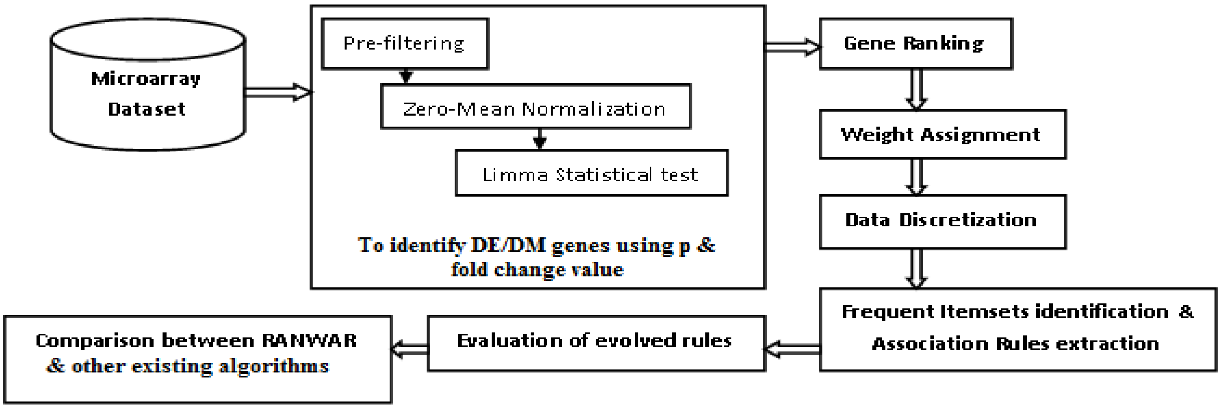
Microarray Dataset
Pre-filtering
Zero-Mean Normalization
Limma Statistical test
To identify DE/DM genes using p & fold change value
Gene Ranking
Weight Assignment
Data Discretization
Frequent Itemsets identification & Association Rules extraction
Evaluation of evolved rules
Comparison between RANWAR & other existing algorithms

Fig.1 RANWAR Architecture

The moderated t-statistic [3] of Limma to calculate p-value is stated as:

$$\tilde{t}_g = \left(1/\sqrt{(1/n_1)+(1/n_2)}\right) * (\hat{\beta}_g/\tilde{s}_g)$$

where sample size n = n1 + n2 and and denote the contrast estimator and posterior sample variance for the gene g, respectively. If p-value of a gene is less than the entered value, then the gene is called DE/DM, otherwise not.

The fold change value is calculated using: F = Fitness value of normal gene – V

The fold change value [1] is also calculated to find the deviation on the gene expression from its original expression, suppose if p-value entered is 0.05 and fold change value is 5.The genes are then ranked w.r.t. their p-values.

Step 2: Assigning Weight to Each Gene

Some weight is assigned to each gene w.r.t. their p-value ranking, and the weight of the first ranked gene is always 1. The ranges of weight lie in between 0 and 1. The number of genes and its weight is calculated using:

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

where n is number of genes and r_i is its rank.

Step 3: Data Discretization

The data is discretized using GENCLUS which adopts the notion of density based approach. The GENCLUS clustering algorithm runs sample-wise (i.e., row-wise) on each row to form the cluster and the distance metric is the Euclidean distance. The cluster having higher centroid value is the cluster of up- regulated / hyper- methylated genes (DE_{up}) and the other cluster is the down-regulated/hypo–methylated (DE_{down}).

Step 4: Identification of Frequent Itemsets and Rule Mining

After the discretization [2], the frequent itemsets should be identified. For this, at first, the *wcs* of the 1-itemsets is evaluated. Similarly, their supersets 2-itemsets and frequent 2-itemsets are determined. Then the rules are extracted from the frequent 2-itemsets. The rules having greater than equal to value, are selected for resulting list of rules and so on.

suppose, { gene1+, gene2-, $gene3_{nde}$ => gene4+ }

is an association rule which states that if gene1 is up-regulated ("+"), gene2 is down-regulated ("-"), and $gene3_{nde}$ is non-differentially expressed (depicted as "nde") simultaneously, then it is likely that gene4 will be up-regulated. The algorithm terminates if there is no further successful extensions of frequent item sets to be identified. Finally, the evolved rules are ranked w.r.t. wcs or wcc. The proposed support can be stated as:

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

It is estimated as:

$$W_k(Z) = \prod_{i=1}^{Q}{}_{(\forall g_i \epsilon Z, Q=|Z|)} w_{ki}, \quad w_{ki} = \begin{cases} w_i, & \text{if } g_i \in s_k. \\ 0, & \text{otherwise.} \end{cases}$$

where weight w_i of the gene g_i in the k-th sample/transaction is denoted by $w_{ki.}$ 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.W_k(Z) denotes item set-transaction weight of item set Z for k-th transaction, w_{ki} refers to the weight of gene g_i for k–th transaction, Π denotes multiplicative operator, and Q refers to the total number of genes in the item set Z. The proposed support of the item set (i.e., *wcs*(Z)) is defined in two folds. If the size of item set is one, then the *wcs*(Z) is the ratio of summation of all the item set -transaction weights of the item set (i.e., summation of all W_k(Z),for all k) to the total number of transactions/samples (i.e., m) in the database. But if the size of item set is greater than one, then the *wcs*(Z) is the ratio of summation of all the item set-transaction weights of the item set (i.e., summation of all W_k(Z), for all k) to the frequency of the highest frequent gene/item of the item set (i.e.,m^1(Z)) instead of considering m. where $m^1(Z)$ is described as follows:

$$m'(Z) = \max_{(\forall g_i \epsilon Z, Q=|Z|)} \left\{ \sum_{k=1}^{m} BIT_{k1}, \sum_{k=1}^{m} BIT_{k2}, \ldots, \sum_{k=1}^{m} BIT_{kQ} \right\},$$

where, Q denotes the total number of genes in the item set Z(|Z| > 1) , and BIT_{ki} denotes Boolean value of the gene g_i for k-the sample/transaction(here i = 1,2,..Q). Here, the Boolean sub-matrix of the item set Z having size |m X Q| is considered. The proposed confidence of a rule (viz., *wcc* (A→ C) is defined as the ratio of the support of the item set (i.e., *wcs(Z)* to the support of the antecedent(i.e., *wcs(A)*).

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

RANWAR is compared with other existing rule mining methods [4][5].The number of rules is compared with RANWAR, and other methods at different minimum support values. For validation of the rules, GO terms and KEGG pathways of the genes in the rules are identified. The report contains many top ranked rules produced by RANWAR that hold poor ranks in traditional Apriori, but are highly biologically significant to related diseases.

IV. CONCLUSION

The existing system generates huge number of evolved rules by association rule mining algorithms which make confusion to the decision maker. The proposed system makes use of three techniques to mine the association rules comparatively less than Apriori algorithm but these rules hold top rank in RANWAR whereas not in Apriori. For selection of perfectly differentially expressed / methylated genes the p-value and fold change value are used. For assigning weight to each gene various network weighting methods are used, which reduces the complexity. For data discretization effective cluster algorithm i.e., Genetic Cluster Algorithm is used. Thus, it saves time for execution of algorithm. Since it runs on gene expression and methylation datasets. The genes of the top rules are biologically validated by Gene Ontology (GOs) and KEGG pathway analyses. Many top ranked rules extracted using these techniques hold poor ranks in traditional Apriori, are highly biologically significant to the related diseases. Finally, the top rules evolved using these techniques which are not been evaluated by Apriori are reported. And it can be enhanced by finding the symptoms and appropriate medicine for the related diseases. In future various network weighting methods can be used to form the final composite network that results from a search. Such as Query dependent weighting, Gene-Ontology (GO) based weighting and equal weighting to view the relationship in different patterns.

REFERENCES

[1] Anthony Deeter, Mark R Dalman, Gayathri Nimishakavi, Zhong-Hui Duan “Fold change and p-value cutoffs significantly alter microarray interpretations” Dalman et al. BMC Bioinformatics 2012, 13 (Suppl 2):S11 http://www.biomedcentral.com/1471-2105/13/S2/S11

[2] Claudia Marinica and Fabrice Guillet “Knowledge-Based Interactive Post mining Of Association Rules Using Ontology’s”, IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 6, June 2010

[3] Giuseppe Agapito, Mario Cannataro, Pietro Hiram Guzzi,and Marianna Milano “Extracting Cross-Ontology Weighted Association rules from Gene Ontology Annotations” 1545-5963 (c) 2015 IEEE

[4] Yuande Tan and Yin Liu proposed “Comparison of methods for identifying differentially expressed genes across multiple conditions from microarray data” ISSN 0973-2063 Bioinformation 7(8): 400-404 (2011) /www.bioinformation.net

[5] Yuan-De Tan, Myriam Fornage, Yun-Xin Fu “Ranking Analysis Of Microarray Data: A Powerful Method for Identifying Differentially Expressed Genes”, Science Direct Genomics 88 (2006) 846–854 www.Elsevier.Com/Locate/Ygeno