This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2014.2359494, IEEE Transactions on NanoBioscience

RANWAR: Rank-Based Weighted Association Rule Mining from Gene Expression and Methylation Data

Saurav Mallik*, Anirban Mukhopadhyay†, *Senior Member, IEEE,* and Ujjwal Maulik‡, *Senior Member, IEEE*

***Abstract*—Ranking of association rules is currently an interesting topic in data mining and bioinformatics. The huge number of evolved rules of items (or, genes) by association rule mining (ARM) algorithms makes confusion to the decision maker. In this article, we propose a weighted rule-mining technique (say, $RANWAR$ or rank-based weighted association rule-mining) to rank the rules using two novel rule-interestingness measures, viz., rank-based weighted condensed support (wcs) and weighted condensed confidence (wcc) measures to bypass the problem. These measures are basically depended on the rank of items (genes). Using the rank, we assign weight to each item. $RANWAR$ generates much less number of frequent itemsets than the state-of-the-art association rule mining algorithms. Thus, it saves time of execution of the algorithm. We run $RANWAR$ on gene expression and methylation datasets. The genes of the top rules are biologically validated by Gene Ontologies (GOs) and KEGG pathway analyses. Many top ranked rules extracted from $RANWAR$ that hold poor ranks in traditional Apriori, are highly biologically significant to the related diseases. Finally, the top rules evolved from $RANWAR$, that are not in Apriori, are reported.**

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

I. Introduction

KNOWLEDGE Discovery and Data Mining (KDD) is an interdisciplinary domain that mainly focus 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) [3], 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. For this, different rule-interestingness measures (viz., support, confidence, lift, conviction etc.) were proposed. But, these still generate huge number of frequent itemsets, and thereby these generate huge number of association rules. Thus, lot of time is taken to run these algorithms. In this article, we propose a weighted rule-mining technique (viz., $RANWAR$ or Rank-based Weighted Association Rule-Mining) which has been developed using two novel measures rank-based weighted condensed support (say, wcs) and rank-based weighted condensed confidence (say, wcc) measures for extracting rules from the data. Sometime it happens that a lot of rules have same support and same confidence. At this moment, if we need some of them, it is difficult to differentiate among them. Therefore, if we apply the wcs and wcc, we can easily categorize them. The major benefit of $RANWAR$ is that it generates much less number of frequent itemsets than state-of-the-art association rule mining algorithms for same minimum support value. There is no such ARM method which generates lesser number of frequent itemsets than $RANWAR$. Thereby, it takes much less time than the other algorithms. Another benefit of $RANWAR$ is that some of the rules which hold low rank in traditional rule mining algorithms, hold good rank in $RANWAR$. Some evidences of biological significance of the genes of the evolved rules are also found.

As we know that if the number of genes in data is large, the number of itemsets will be also large, thus, using Limma [2] statistical test, we have just taken into account the top differentially expressed (i.e., DE) or differentially methylated (i.e., DM) genes. Limma is an useful statistical test which performs well for both normally and non-normally distributed data for all types of sample-size (i.e., small, medium, large). Our proposed measures are basically rank-based weighted measures. Therefore, ranking of genes has a significant role here. Limma test provides a rank-wise gene-list according to their p-values from best to worst cases. Thereafter, we assign weight to each item/gene w.r.t. their p-value ranking, and include these into the measures. Therefore, our measures give importance to each item (gene). Our proposed measures (viz., wcs and wcc) are condensed form of the traditional support and confidence measures. Furthermore, two gene expression datasets and two methylation datasets are used to test the performance of $RANWAR$. We have made a comparative analysis of it with the traditional Apriori algorithm [1] and other state-of-the-art rule mining algorithms. For validation of the rules, GO terms and KEGG pathways of the genes in the rules are identified. The genes of the evolved rules involving highest number of GOs/pathways are reported for biologically benevolent aspects. 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.

The rest of the article is organized as follows. Section II presents literature review. Section III and section IV elaborate the proposed measures, and the proposed ARM method, respectively. The source and description about the dataset are given in section V. Section VI presents the experimental results and discussion, where section VII shows utility of ARM in biology and our novel findings. Finally, Section VIII concludes the article.

II. Literature Review

ARM [1] is a popular technique to estimate interesting relationships among different items (i.e., genes). Suppose, $Itmset = \{i_1, i_2, ..., i_n\}$ be an itemset (i.e., set of genes) and $S = \{s_1, s_2, ..., s_m\}$ be a set of transactions (samples). Thus, a rule might be described as $A \Rightarrow C$, where $A, C \subseteq Itmset$ and $A \bigcap C = \phi$. Here, A is called as antecedent and C is called as consequent. In a transaction database, a transaction may consist of a set of items purchased in it. In a similar sense, in gene expression / methylation [3] dataset, in any tissue sample (trans-

*S. Mallik is with Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. E-mail: sauravmtech2@gmail.com, chasaurav_r@isical.ac.in

†A. Mukhopadhyay is with the Department of Computer Science and Engineering, University of Kalyani, Nadia, India. E-mail: anirban@klyuniv.ac.in

‡U. Maulik is with the Department of Computer Science and Engineering, Jadavpur University, Kolkata, India. E-mail: umaulik@cse.jdvu.ac.in