International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-5, Issue-3, July 2016

Adaptive K-means clustering for Association Rule Mining from Gene Expression Data

Deepti Ambaselkar, Dr. A. B. Bagwan

Abstract— **Mining of association rules in the field of bioinformatics is essential for the researchers to discover the relationship between different genes. A large number of association rules generatetd by the conventional rule mining approaches creates a problem for selecting top among them. Rank-Based Weighted Association Rule-Mining (RANWAR) is used for rule ranking by implementing two measures i.e. weighted condensed support (wcs) and weighted condensed confidence(wcc) which is used to solve this problem. The proposed system helps to find out top gene among all the present gene in the system. The results are generated through rule generation algorithm which enhances the system outcome, by implementing RANWAR algorithm. This system implements Adaptive K-means clustering for rule mining in the proposed system. This technique for clustering relies on K -means such that the knowledge is partitioned into K clusters. During this technique, the number of clusters is predefined and therefore the technique is extremely dependent on the initial identification of components that represent the clusters well.**

Index Terms— **RANWAR, Gene-rank, gene-weight, wcc, wcs, weighted association rule generation.**

I. INTRODUCTION

In the area of data mining the rule generation methods are utilized to determine relationships between different items. The ranking of genes results is utilized to research about the bioinformatics for the avoidance of particular health problems. In the proposed framework rule mining is done with the help of weighted support and and weighted confidence of the record of generated outcomes. In this paper a weighted rule mining method i.e. RANWAR has been used and adaptive k-means clustering has been proposed. In some cases it may happen that various rules hold the same value for support as well as confidence. In such circumstance, if it is required to select couple of them, it becomes difficult . This issue is solved with the help of RANWAR. The benefit of this method is that it generates very less number of item-sets as compared to traditional algorithms used for association rule mining. It has been observed that there's no such ARM method that generates less amount of frequent itemsets than RANWAR. Along these lines, it requires less measure of time contrasted with different algorithms. One more point of preference of this algorithm is that a some of the generated rules that hold poor rank in traditional rule generating algorithms, hold top position i.e. rank in RANWAR.

In this framework basically the differentially expressed (DE) and differentially methylated (DM) genes have been used. For this, Limma is the related accommodating implemented mathematics that performs great, for normal and non normal distribution of information for each type of information size (i.e. little, medium, vast). Along with this rank is provided to each gene in the genelist obtained from Limma based on their p-values. After providing rank to genes weight is assigned to each genes which is calculated with the help of their respective ranks [1]. In this way, importance to each gene is provided. The measures used here i.e. wcs and wcc are condensed style of the ordinary support and confidence. A near examination of this framework [1] has been made with the Apriori algorithmic rule and other rule mining algorithms.

Selection of top rules between huge amounts of rules is always a problematic procedure.This issue is overcome by applying RANWAR algorithm. But there is a need of improvement in the accuracy of generating rules. Hence, we have proposed the adaptive k-means clustering approach in order to form the initial clusters more accurately as compared to k-means clustering.

II. LITERATURE SURVEY

Authors have composed [1] two new weighted condensed rule-interestingness measures on the basis of ranks such as wcs and wcc. RANWAR which is one of the weighted algorithm for mining of rules has been composed with these measures. To ascertain p-value of gene it makes utilization of statistical test, Limma and some weight is given to every gene on the basis of their p-value ranking. Authors utilized couple of gene expression and additionally two methylation datasets to obtain the contrast between efficiency of RANWAR with the other ARM algorithms. It delivers low measure of frequent sets of item contrasted with rest of the algorithms. In such way the RANWAR decreases time of the algorithm execution.

Authors approached [2] a problem consist in the mining association rules in client transaction in an immense dataset as in sets of items. Authors have made an exploration on finding down legitimate rules that have minimum transactional and minimum certainty. They found the rules that have single item inside the subsequent and a union of any number of items in the precursor. They have tackled this issue by decaying it inside two sub issues as searching those item-sets which are available at minimum.

Authors led an observation [4] for the prediction of Uterine Leiomyoma of statistical methods and association rules on mRNA expression and DNA methylation datasets. Authors likewise proposed a rule-base classifier which is one of a kind. Authors utilized a Genetic Algorithm based rank aggregation approach over the association rules which is the outcome from the training information by Apriori association rule mining algorithm on the basis of 16 different rule-interestingness measures.

Author attempted [6] to expand regarded the unique concept and methods utilized for mining association rule from the genomic information have been clarified additionally as of

Deepti Ambaselkar, RajarshiShahu College of Engineering, Pune, India
Dr. A. B. Bagwan, RajarshiShahu College of Engineering, Pune, India

late exhibited association rule mining methods have been checked on and talked about.

III. IMPLEMENTATION DETAILS

- Identifying Differentially Expressed/ Methylated Genes and their Ranking

As it is known that gene dataset contains a large amount of genes, so, some pre-filtering method is applied on the dataset (i.e. genes with low change are eliminated). In case, due to low variance of genes, regularly lesser p-value is produced that seem to be significant, however in real world it is not. Subsequently, the general contrast in the information relatively for each previous gene must be inspected and strain the genes having to massively least difference. The separated data must be normalized gene-wise since normalization changes the information from totally various scales into a standard scale. Here, zero-mean normalization has been utilized that changes over information in such a way that mean of each gene gets to be zero and standard deviation of each gene gets to be one.

With a specific end goal to perceive DE and DM genes, a non-parametric test is needed, to be appropriately implemented. In this manner, Limma has been chosen since it performs better for every typical as well as non-typical conveyance for the whole volumes of information.

In any case, on the idea of value got from t-statistic, correlated p-value is calculated from either aggregate circulation function (cdf) [8] or t-table. If p-value is less than 0.05, then that gene is DE/DM, else it is not. At that point rank is given to these genes on the basis of their acquired p-values.

- Assignment of Weight to Every Gene

In this technique, each gene doesn’t have same importance. Subsequently, some weight or value is allocated to each gene according to their ranking. The weights of all genes are calculated in a way such that the variation in weights of any of the two genes that is consecutively ranked is similar and the weight of the gene that is basically ranked first is one. The points of range of assigned weights lie between zero and one. If the total number of genes in the system are “n” . Then, weight of each gene (shown by $1 \leq i \leq n$) is computed from a function of rank (indicated by r_i $1 \leq i \leq n$) and number of genes.

- Discretization of Data

Assume that input is the information matrix shown as I[r,c] where, r shown genes and c shown samples. To start with, the transpose of matrix is executed. Presently, procedure of discretization for the input information mtrix is need for deploying association rule mining. With the end goal of discretization, here Adaptive k-implies clustering rule is utilized by us.Reducing the iterations is the aim behind the use of Adaptive using K - means clustering. The concept is to define k-centres for every cluster. In next step association of every data point with its nearest centre after computing the distance from every centres selected. The distance is known as Euclidean distance. Early age group is the stage which comes after completion of first iteration (when every point is covered and no points are pending). Here we compute k-new centroids as centre of clusters resulting from the last step and once more novel binding is finished by computing the distance.The k-means algorithm starts at beginning cluster centroids, and iteratively alters these centroids to minimize the objective function. The k-means always collect to a local minimum. The specific local minimum found based on the beginning cluster centroids.

Steps involved in adaptive k-means clustering are given bellow,

Step1: Choose number of clusters for clustering.
Step2: Get all elements for clustering.
step3: Select an element randomly to get count between two elements from each existing elements.
step4: Form clusters where minimum distance of count found of each elements.
step5: Get all the initial clusters by sorting each element into existing clusters and get mean(centroid).
step6: Calculate mean of each clusters.
step7: Compare each mean of cluster with other elements of cluster.
step8: If distance is less between mean and element of other cluster shift the element into specific cluster.
step9: Calculate again mean of existing cluster.
step10: Repeat the previous step as per given iterations.

- Rule Mining and Identifying Frequent Itemset

After the post-discretization procedure, it is expected to spot frequent item-sets. For this, at in the first place, assessment of 1-itemsets is done, so to decide the frequent singleton item-sets (i.e. wcs of these singleton item-sets is more prominent than the support threshold min_wsupp). Essentially, their supersets i.e. 2-itemsets are computed to affirm the 2-itemsets that are repetitive. At that point extraction of rules is done from these itemsets. From that point, weighted condensed confidence for every single rule is calculated. The rule with their values of wcc being greater than or equivalent to the defined confidence threshold (min_wconf). Then, their supersets i.e. 3-itemsets are resolved after which visit 3-itemsets are resolved and afterward from these item-sets noteworthy or say, essential rules are separated. The algorithmic rule stops if there are no extra augmentations of repeated item-sets to be considered. At long last, the rules included are ranked depends on their wcs or wcc. For points of interest, see algorithmic rule 1, which is fundamentally an upgraded and weighted version of Apriori [1].

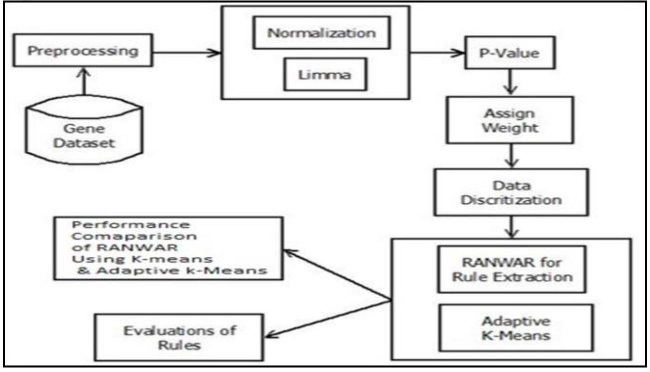

Fig. 1.The proposed approach for rule mining from biological data.

A. *Algorithm*

Input: Data matrix, D(genes, samples),original gene-list A1 as per the D, rank-wisegene-list A2(as per p-value), flag to sort theevolved rules sort flag (based on either wcs orwcc), minimum value of threshold for supportmin_wsupp , minimum value of threshold for confidence min_wconf.

Output: Rule-set Rules, support RuleSupp,confidence RuleConf.

```
1) start
2) Normalize D with zero-mean normalization.
3) Evaluate gene-rank (i.e.rankk(:) ) as per listof genes i.e. A1.
4) Allot weight wt(:) to each gene as per theirrank rankk(:).
5) Take transpose of the normalized matrix.
6) Select initial values of seed for usingAdaptive k-means clustering.
7) Discretize transposed data-matrix byapplying adaptive k-means clusteringsample-wise.
8) Apply the technique of post-discretization.
9) Initialize k =1.
10) Discover frequent 1-itemsets,
    $FI_k = \{i \mid i \in A1 \wedge wcs(i) \geq \min\_wsup\, p\}$
11) repeat
12) Increment k by 1.
13) Produce itemsets that are candidate, $CI_k$from $FI_{k-1}$itemsets.
14) for all itemsets that are candidate c $\in CI_k$,
    do
15) Evaluate wcs(c) for every c.
16) ifwcs(c) $\geq$min_wsupp then
17) $FI_k \leftarrow [FI_k; c]$
18) Produce rules, rule(:) from the frequentitemset, c.
19) Find out wcc(:) for every rule(:).
20) for all the evolved rules, r $\in$ rule(:) do
21) ifwcc(r) >= min_wconf then
22) Accumulate the r in the resultant rule-listalong with its wcs as well as wcc; $Rules \leftarrow r$ $RuleSupp \leftarrow wcs(r)$ and $RuleConf \leftarrow wcc(r)$ .
23) end of if
24) end of for
25) end of if
26) end of for
27) until
28) end of procedure
```

B. Mathematical Model

Normalization:

The zero-mean normalization can be shown as:

$$X_{ij}^{norm} = (X_{ij} - \mu)/\sigma$$

where, X_{ij} and X_{ij}^{norm} represents the gene-value at ith position and jth sample before and after being normalized. μ is the arithmetic mean and σ is the standard deviation.

Limma Test:

The tempered t-statistic in Limma is shown as:

$$\bar{t}_g (1/\sqrt{(1/n_1)+(1/n_2)}) * (\hat{\beta}_g / \bar{s}_g)$$

Assigning Weight:

The formula for calculation of weight is:

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

where w_i is the calculated weight of ith gene and r_i is the rank of ith gene in the system.

C. Experimental Setup

This system developed on Java framework(version jdk 6) and Netbeans (version 6.9)is used as a development tool on Windowsplatform. System is able to run on commonmachine also it does not require any specifichardware.

IV. EXPERIMENTS AND RESULTS

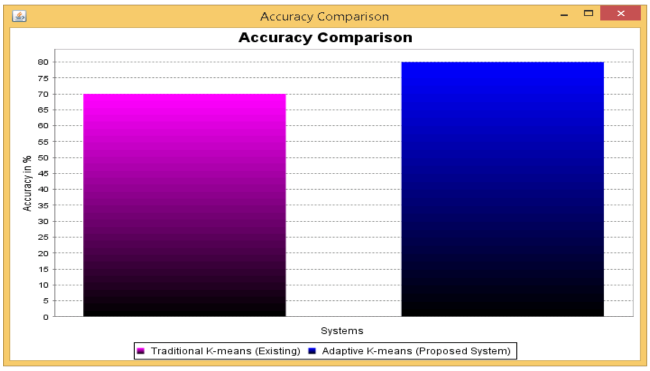

Fig. 2.Comparison of accuracy of k-means and adaptive k-means.

As the quantity of genes in every dataset is extensive thus, Limma test is performed on the dataset which is normalized and after that the system is introduced to construct a rank-wise list of genes from best to most pessimistic scenario in accordance with their p-values inside of the test. It is determined as that intensively (statistically) essential genes will be vital to the comparing disease for medical science data. Along these lines, expecting all genes of an outsized dataset may expand time period greatly and that it can take an outsized collection of rules inside of which the majority of them are excess/irrelevant. In this manner, it is proposed to consider high genes and create exclusively most indispensable rules instead of an outsized set of repetitive rules. Here, this system has exhibited to set entirely various accomplished diverse shorts of p-value for different datasets for classifying a few numbers of high reviewed genes.

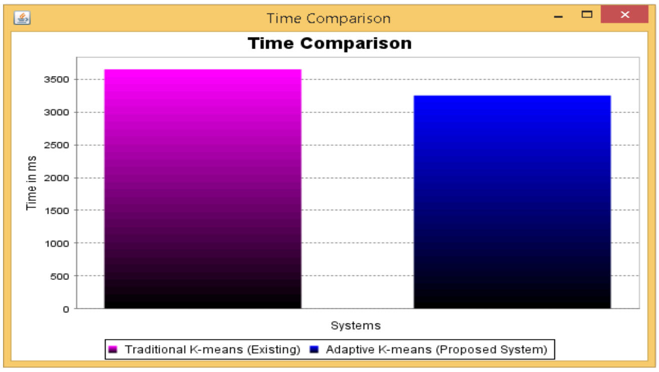

Fig. 3. Comparison of time of k-means and adaptive k-means.

V. CONCLUSION

Top record is gained by using rule generation operation for gene expression in proposed system. In proposed system we deploy some of helpful strategies like Limma test, p-value, gene ranking etc.As a contribution adaptive k-means clustering is used rather than k-means clustering operation to solve some issues of previous system. System performance is improved by using Adaptive k-means clustering and proper outcomes are generated by using RANWAR algorithm. RANWAR algorithm performs more effectively as compared with Apriori rule generation algorithm. Systems efficiency and difficulty is calculated by providing two gene expression datasets.This system has tendency to collect top rules which are extracted from the deployed system to make accessible for research and analysis.

REFERENCES

[1]SauravMallik, Anirban Mukhopadhyay, Ujjwal Maulik ,“RANWAR: Rank-Based Weighted Association RuleMining From Gene Expression and Methylation Data”,IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL.14, NO. 1, JANUARY 2015.

[2] R. Agrawal, T. Imielinski, and A. Swami, ”Mining AssociationRules between Sets of Items in large Databases”, inProc. ACM SIGMOD ACM, New York, vol. 216, pp. 207216.

[3] G. Smyth, ”Linear Models and Empirical Bayes Methodsfor Assessing Differential Expression in Microarray Experiments”,Stat. Appl. Genet.Mol.Biol., vol. 3, no. 1, p. 3, 2004.

[4] S. Malliket al., ”Integrated analysis of gene expression andgenomewide DNA methylation for tumor prediction: Anassociation rule miningbased approach”, in Proc. 2013 IEEESymp. Comput.Intell.Bioinformat.Comput. Biol. (CIBCB),Singapore, pp. 120127.

[5] S. Bandyopadhyay, U. Maulik, and J. T. L. Wang, ”Analysisof Biological Data: A Soft Computing Approach”. Singapore:World Scientific, 2007.

[6] M. Anandhavalli, M. K. Ghose, and K. Gauthaman, ”AssociationRule Mining in Genomics”, Int. J. Comput.TheoryEng., vol. 2, no. 2, pp. 17938201, 2010.

[7] D. Arthur and S. Vassilvitskii, ” the advantages of carefulseeding”, in Proc. ACM-SIAM SODA 2007, Soc. Ind.Appl.Math., Philadelphia, PA, USA, 2007, pp. 10271035.

[8] S. Bandyopadhyayet al., ”A survey and comparative studyof statistical tests for identifying differential expressionfrom microarray data”, IEEE/ACM Trans. Comput. Biol.Bioinformat., vol. 11, no. 1, pp. 95115, 2014.

[9] A. Thomas et al., ”Expression profiling of cervical cancersin Indian women at different stages to identify gene signaturesduring progression of the disease”, Cancer Med.,vol. 2, no. 6, pp. 836848, Dec. 2013.

[10] J. Liu et al., ”Identifying differentially expressed genesand pathways in two types of non-small cell lung cancer:Adenocarcinoma and squamous cell carcinoma”, Genet.Mol. Res., vol. 13, pp. 95102, 2014.

[11] W. Wei et al., ”The potassium-chloride cotransporter 2promotes cervical cancer cell migration and invasion byan ion transport-independent mechanism”, J Physiol., vol.589, pp. 53495359, 2011.