International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 12
ISSN: 2454-4248
443 – 447

Role of Association Rule Mining in DNA Microarray Data - A Research

T. Arundhathi
Asst. Professor Department of CSIT
MANUU, Hyderabad
Research Scholar
Osmania University, Hyderabad

Prof. T. Adilakshmi
Head, Department of CSE
Vasavi Engg. college (A)
Osmania University, Hyderabad

***Abstract*:-** Gene expression data analysis with DNA microarray technology has become a fundamental tool in genomic research. DNA Microarray technology monitors the gene expression levels of thousands of genes at a time under a certain condition makes them suit-able for a quite lot of biological applications. As the Association rules are widely using in market-basket analysis, these rules can also be applied in biological problems. Association rules play an important role in the computational biology. In this paper, how the Association rule mining is important for micro array gene association analysis have been discussed. Though the clustering offers a natural solution to these bio-logical problems by discovering frequent item sets on microarray data and reviewed the different methodologies discussed by the authors on micro array gene expressed data.

Keywords- *Gene expression data, DNA Micro array, Association rule mining.*

*****

I. INTRODUCTION

A. Data Mining is a process of extracting potential and novel Information from large datasets. There are many techniques that have been used to discover such kind of knowledge. These techniques are classification, dependence modeling clustering regression, prediction and Association. One of the most important data mining is mining association rules. Association rules rst introduced in 1993 are used to identify relationships among a set of items in databases.

B. Gene Expression Microarray Data technology allows the simultaneously monitor of expression levels for thousands of genes or entire genomes. Micro array gene expression is the conversion of the DNA sequences into mRNA sequences by transcription then translated into amino acid sequences called proteins. Micro array technologies provide the opportunities to compute the expression level of tens of thousands of genes in cells simultaneously. The expression level is associated with the corresponding protein made under different conditions. The data collected from micro array experiments is commonly in the form of MXN matrix of expression level when represents columns. The total gene expression data can be valuable in understanding of genes, cellular states and biological net-works. Analysis of this genomic data has two important goals. First is to determine how expression of any particular gene might affect the expression of other genes. Second is to determine what genes are ex-pressed as result of certain cellular conditions. What genes are expressed in diseased cells that are not expressed in healthy cells. A Micro array data base is repository containing microarray gene expression data. The Key uses of micro array data base are to store the measurement data, manage searchable index and make the data available to other applications for analysis and interpretation.

C. Association Rule Mining (ARM) is the one of the widely using best researched methods of data mining [1].Association Rule Mining is an unsupervised data mining technique which produces understand-able rules. An Association rule is an implication expression of the form X-> Y where X and Y are the two disjoint item sets i.e X Y =. The strength of an association rule can be measured in terms of its support and con dence[2].In data mining , the association rule mining is introduced to detect hidden facts in large data sets and drawing inferences on how a sub-set of items inferences the presences of another sub-set[3]. Let I= i1, i2, i3,..id be the set of all items and T=t1,t2,t3,..tn be the set of all transactions. Each transaction Ti contains a subset of items chosen from I, item set contain 0 or more items. Support is how of-ten a rule is applicable to dataset. Confidence is how frequently items in Y appear in transaction that con-tain X. Support s(X- > Y) = (XUY)/N = (support count for X and Y)/N. Confidence = (genes) and N represents rows (samples) (support count for X and Y) / (support count for X) = (XUY)/(X). Support count is the item set property that is the no. of transactions that contain a particular item set. (X) = ti X ti,ti . If support value is very low implies a rule occurs simply by chance. Low support rule is uninteresting in any domain. Therefore sup-port is used to eliminate the uninteresting rules and to exploit for the efficiency of association rule. Confidence measures the reliability of the inference made by the rule (X-> Y).The higher the confidence the more likely it is for Y to be present in transactions that contain X.

D. Association Rule Discovery (ARD) Given a set of transactions T, find all the rules having support = min.sup and confidence =min. conf When min.sup, min.conf are the corresponding support and confidence thresholds. In general

 ISSN: 2454-4248
443 – 447

every association Rule must satisfy both support and confidence values. So the target is to generate all association rules that Satisfy user threshold minimum support and confidence values [9]. A conventional method, Brute -force approach for mining association rules is to compute the support and confidence for every possible rule .How-ever this method is very expensive and that for every small data set it is extracting more no. of association rules and among them approximately 80To generate frequent item set To generate rules, which are depend on frequent item set. E. Discretization Most of the application of association rule mining on microar-ray gene expression still relies on discretization tasks before applying any data mining technique [2]. The normalized microarray dataset is usually represented as a series of continuous numbers [1]. Discretization is the process of transformation from continuous data into discrete data. The threshold method used to discretize the data. This method is suitable for microarray analysis [6, 7].Genes with log expression values greater than a particular value is considered as over expressed, otherwise as under expressed. Using threshold method each gene expression is converted into one of the two discrete values 1, 0 for over ex-pressed and under expressed.

II. RELATED WORK

The basic task of mining association rules technique extract interesting relationships among set of items. So the rule must satisfy both support and confidence values. So the target is to generate all association rules that satisfy user threshold minimum support and confidence values.

SECTION 1: Deriving Association Rules on Microarray data set. Methodology- 1: For ex-tracting the frequent items sets using Aprio Algorithm INPUT: D- A data base of transactions, Min-sup- the minimum support count threshold OUTPUT : L-frequent item-sets in D. DETAILS: The frequent item set is an item set that should satisfy the minsup threshold. In this method the Apriori principle is an effective way to eliminate some of the items without counting their support values. The principle of Apriori is that if an item set is frequent, then all of its subsets must also be frequent .The principle has a property called antimonotone. The principle prune the item sets which are not infrequent called the support based prunning and this based on the support measure. The antimonotone property says that the support for an item set never exceeds the support count for its subsets [2].

a) The rule extraction procedure is b) Read the gene expression data form source c) Convert continuous values into discretized values

d) Discover the frequent items by using Apriori algorithm. e) Generate the association rules from frequent items. f) Discriminant association rules. The Algorithm is as follows Step 1: determine the support of each item .after completion of this step the algorithm produces the set of all frequent 1-item sets called F1. Step 2: iteratively generate new candidate k-item sets using the frequent (k-1) item-sets found in the previous iteration and count the support of the candidates. Step3: Eliminates some of the candidate k-item sets using support-based prunning strategy. Step4: algorithm terminates when there are no new frequent item sets generated Fk=0. An example [8] on microarray data.

Table 1. Microarray Dataset

	S1	S2	S3	S4
gene1	0.68	-0.5	0.78	-0.34
gene2	-0.23	1.2	0.61	0.89
gene3	0.66	0.84	0.99	-0.10
gene4	0.87	-1.0	-0.67	-0.44
gene5	-1.0	0.83	0.65	0.61

Table 2: Discretized Microarray data

	S1	S2	S3	S4
gene1	1	0	1	0
gene2	0	1	1	1
gene3	1	1	1	0
gene4	1	0	0	0
gene5	0	1	1	1

Tid	Items
sample 1	gene1,gene3,gene4
sample 2	gene2,gene3,gene5
sample 3	gene1,gene2,gene3,gene5
sample 4	gene2,gene5

Assume min.sup. is 50Support count =50/100 * no. of samples =1/2 * 4=2 Iteration 1:

Itemset	support count
gene1	2
gene2	3
gene3	3
gene4	1
gene5	3

Gene 4 has been discarded since its support count is less than the minimum support count.

 ISSN: 2454-4248
443 – 447

Itemset
gene1
gene2
gene3
gene5

Iteration 2:

Itemset	support count
gene1,gene2	1
gene1,gene3	2
gene1,gene5	1
gene2,gene3	2
gene2,gene5	3
gene3,gene5	2

Gene1, gene2 and gene1,gene5 have been discarded since its support count is less than the minimum Sup-port count.

Itemset
gene1,gene3
gene2,gene3
gene2,gene5
gene3,gene5

Iteration 3:

Itemset	support count
gene1,gene2,gene3	1
gene1,gene2,gene3,gene5	1
gene1,gene3,gene5	1
gene2,gene3,gene5	2

the follwoing item sets have been discarded. since its support count is less than the minimum sup-port count. gene1,gene2 ,gene3 and gene1,gene2, gene3,gene5 and gene1,gene3,gene5

Ultimate frequent item set

Itemset
gene2, gene3, gene5

Methodology 2: Extracting Association rules from the frequent item Sets. Rule By using the frequent item set we can extract association rules very efficiently. Every frequent k-itemset Y can produce (2k-2) association rules. An association rule that can be extracted by partitioning the item set Y into two non-empty subsets, X and Y-X satisfy the confidence threshold. The association rules are generated as follows.

Association Rule
gene2!gene3,gene5
gene3!gene2,gene5
gene5!gene2,gene3
gene3,gene5!gene2
gene2,gene3!gene5
gene2,gene5!gene3

Cosider the rule gene2-> gene3, gene5 Confidence of the rule = support of gene2,gene3,gene5 / support of gene2 = 2/3 =66

Association Rule	Confidence	Support
gene2!gene3,gene5	66%	50%
gene3!gene2,gene5	66%	50%
gene5!gene2,gene3	66 %	50%
gene3,gene5!gene2	100%	50%
gene2,gene3!gene5	100%	50%
gene2,gene5!gene3	66%	50%

Confidence Based pruning method to lter the association rules. Note that the confidence does not have any monotone property. If we compare the rules gen-erated from the same frequent item set Y, the follow-ing theorem holds the confidence measure.

3 Theorem: If a rule X-Y-X does not satisfy the confidence threshold, then any rule X1-Y- X1 where X1 is a subset of X, must not satisfy the confidence threshold as well.

 ISSN: 2454-4248
443 – 447

Proof: - take any two rules X1-Y- X1 and X-Y-X where X1 contained X. The confidence of the rules re (Y) / (X1) and (Y)/(X) respectively. (X1) = (X) since X1 is a subset of X. therefore the former rule cannot have a higher confidence than the latter rule. Rule Generation in Apriori Algorithm uses the level wise approach for generating the association rules. Initially all the high confidence rules that have only one item in the rule consequent are extracted. These rules are then used to generate new candidate rules, the fol-lowing is the lattice structure for the association rules generated from the frequent item set gene2, gene3, gene5. Any node in the lattice has low confidence then according to the above theorem the entire sub graph spanned by the node can be pruned immediately. Example is g2, g5 ->g3 is low confidence then all the rules containing item gene3 in its consequent also can be discarded. Computational complexity of the Apriori Algorithm The computational complexity of the Apriori algorithm can be affected by the following factors. 1. Support threshold:- If the sup-port threshold value is low then more items being declared as frequent. This effect on computational complexity of the algorithms the maximum size of the frequent itemsets increases, the algorithm will need to make more passes over the data set. 2. No. of items in the data set: As the no. of items increases, more space will be needed to store the support counts of items. 3. No. of transactions:-Since the Aprior algorithm makes repeated passes over the dataset, its run time increases with a large number of trans-actions. 4. Average transaction width: Data which has the property that the no. of items (genes) in the dataset is higher than the number of transaction. For these dense data sets, the average transaction width can be very large. The effect of this is in two ways. a) The maximum size of frequent item sets trends to in-crease as the average transaction width increase. b) As the transaction width increases, more item sets are contained in the transaction. This will increase the number of hash tree traversals performed during support counting.

III. SUMMARY

Using association rule mining approach, we can analyze the expression of one gene leads to the induction of a serial of target gene expressions called the regulation of genes expression. b. The relation-ships between one gene with the other genes can be viewed as an association relation. c. Gene expression may lead to the induction of new biological functions.

SECTION 2: In [9] the author has been focused on Microarray gene association analysis from a frequent pattern mining approach and compared with the Apriori Algorithm Methodology: FP-GROWTH ALGORITHM IN-PUT: FP-Tree, f OUTPUT: All frequent patterns. DETAILS: The method discovered frequent item sets without the generation of candidate item set. The algorithm is divided into two ways a) Build a compact data structure called FP-tree. b) Extracting the frequent item sets directly from the FP-Tree. **ADVANTAGES:** Due to the compact structure, no need to generate candidate itemsets. Second It requires less memory and third Execution time is also very less.

SECTION 3: In [10] the author has been focused on Microarray gene association analysis from a frequent pattern mining approach with gene interval. Methodology: Apriori Algorithm INPUT: D- A data base of transactions Min-sup-the minimum support count threshold OUTPUT: L-frequent itemsets in D. DETAILS: in this method the rule extraction procedure is a. Read the gene expression data form source b. Discritize the data using Equal width interval bin method and substitute the gene values by gene intervals. c. Discover the frequent items by using Apriori algorithm with gene intervals. d. Generate the association rules. e. Discriminant association rules. f. Visualize the biological knowledge.

Conclusion: In this paper it discovered the frequent itemsets on microarray gene expression data using Apriori Algorithm. Reviewed the FP-Growth Algorithm for mining frequent itemsets and also reviewed the different discritization method that the equal width interval bin method on microarray cancer data. In the future as per understanding of the demerits in Apriori algorithm, need to modify the algorithm and need to discuss the alternative method for generating frequent itemsets and their merits and demerits which were discussed in [2].

References

[1] Agrawal, R Imalinski T, Swami A. (1993) Mining association between sets of items in massive database. International Proceedings of the ACM-SIGMOD. International conference on manage-ment of data (PP.207-216).
[2] Pang-Ning Tan- Vipin Kumar Micheal Stein-barch.
[3] Wu,X., Zhang , C., Zhang ,S.efficient Mining of both positive and negative Association rules.ACM Transactions on Information A Systems. 22(3): 381-405(2004).
[4] M.Anandhavalli Member ICSIT, IAENG, M.K.Ghouse,K.Goutam , Association Rule Mining in Genomics.
[5] Alves,R.,Rodriguez-Baena,D,S, and Aguilar-Ruiz,J,S "gene expression data, Brie ngs in Bioinformatics, 2009, vol.2, no.2, pp.210-224.
[6] Zakaria W, Kotb Y and Ghaleb F MCR-Miner: Maximal Confident Association Rules Miner Algorithm for Up/Down Expressed Genes, International Journal of applied Mathematics and Information Sciences, 2014, volume 8 no 2, pp 799-809
[7] Wang, J., Han,J., and Pei,J., CLOSET+ searching for the best strategies for mining frequent closed itemsets. In: proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Wash-ington, DC, USA: ACM, 2003.
[8] Agrwal,J., and Rames,J,C.,h. Analysis of Gene Microarray Data using Association Rule Mining, Journal of computing, 4, 2012.[8] Han,J., Pei,J., and Yin.Y., Mining frequent patterns without candidate Generation, in: Proceeding of ACM SIGMOD

ISSN: 2454-4248
443 – 447

[9] Alagakumar S. Lawrence.R A selective Analy-sis of Microarray Data Using Association Rule Mining .Elsevier ,Procedia Computer Science 47(2015) 3-12

[10] Alagakumar S. Lawrence .R Algorithem for Microaaray Cancer Data Analysis using Frequent Pattern Mining and Gene Intervals. International Journal of computer applications (0975-8887)-National conference on researches issues in Image Analysis and Mining Intelligence (NCRIAMI-2015)