International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 09 | Dec-2015 www.irjet.net p-ISSN: 2395-0072

Efficiently Generating The Rank Based Weighted Association Rule Mining Using Apriori Algorithm In High Biological Database

Premalatha S .[1], Usha Nandhini C.[2]

[1] *Research Scholar, Department of Computer Science, Vellalar College for Women, Tamilnadu, India*
[2] *Assistant Professor, Department of Computer Applications, Vellalar College for Women, Tamilnadu, India*

Abstract - ***Association rule mining algorithm generally used for discovering the relationship among high dimensional gene dataset. Knowledge Discovery and Data mining (KDD) is an interdisciplinary field that mainly focuses on the systematic ways of acquiring interesting rules and patterns which are estimated by interestingness measures include association rules. The large number of evolved rules of items (genes) by Association Rule Mining Algorithms makes confusion to the decision maker. The Rank Based Weighted Association Rule Mining (RANWAR) to rank the rules using two novel rule-interestingness measures. They are rank based weighted condensed support and weight condensed confidence measures to overcome the problem. Such kinds of measures are usually based on the rank of genes. Assign the weight to each item based on the rank which generates less number of frequent item sets than the State-of-the-art using rule mining algorithms. However this scenario takes lot of time to generate frequent item set. In the proposed scenario Temporal Apriori algorithm is introduced to rank the items using weighted condensed support and weight condensed confidence measures. Based on the rank values, the weight values are calculated for each item set. Thus it saves time of execution of the algorithm. Finally concluded the proposed scenario yields superior performance rather than existing scenario through Temporal Apriori algorithm.***

Key Words: *Clustering, Gene item sets, Apriori Algorithm.*

1. DATA MINING

Data mining is a recently promising field, connecting the large Databases, Artificial Intelligence and Statistics. Data mining is also known as knowledge discovery. In contrast to standard statistical methods and data mining techniques for search exciting information without challenging a priori hypotheses. As a field, it has introduced new concept is association rule learning. It has also applied to machine-learning algorithms such as inductive-rule learning (e.g., by decision trees) to very large databases. Data mining techniques are used in business and research that are becoming more and more popular with time. Data mining is an essential step in the process of Knowledge Discovery in Databases (KDD). Data Mining is to extracting or mining knowledge from large amounts of data stored either in databases, or other information repositories.

There are many different meanings of data mining like as knowledge mining from databases, Data/pattern analysis, Data archaeology knowledge extraction, and Data dredging.

The data mining is a non-trivial extraction of implicit, formerly unknown, and useful information from data. Another definition is that data mining has a different type of techniques used to identifying information or decision-making knowledge in bodies of data, prediction, forecasting and estimation.

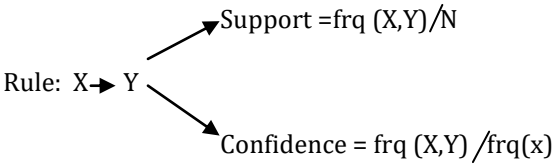
Data mining is the computer-assisted process of digging through analyzing the enormous sets of data then extracting that data. Data mining tools predict behaviors and future trends, allowing business for making proactive, Knowledge-driven decisions. The hunt databases for hidden patterns, finding predictive information that expert may neglect because it lies outside their expectations.

The data mining has several terms, including knowledge extraction, data archaeology, information harvesting, software and data dredging, that actually describe the concept of knowledge discovery in databases[7].

1.1 Association Rule Mining (ARM)

Association rule mining is mainly focused on finding frequent co-occurring associations among a collection of items. It is also called as Market Basket Analysis, since that was the novel application area of association mining. The aim is to find associations of stuff that occur together more often than expect from a random sampling of all possibilities. Association rules are if/then statements that help to uncover relationships between unrelated data in a database and relational database or other information warehouse. Association rules are used to find the relationships between the objects which are regularly used together. Applications of association rules are basket data analysis, classification, cross-marketing, clustering, catalog design etc. For example, if the customer buys milk then he may also buy butter. If the customer buys mobile then he may also buy memory card. There are two basic criteria that support and confidence. It identifies the relationships and rules generated by analyzing the

data. Association rules are usually needed to satisfy a user-specified minimum support and a user -specified minimum confidence at the same time



2. RELATED WORK

M.Anandhavalli et al [2010], analyzed the association rule mining techniques that have been recently developed and used for genomic data analysis is reviewed and discussed. Basically, data mining is an application-dependent issue and different applications may have need of different mining techniques to copy up with. To apply mining association rules in gene expression analysis, the properties of gene expression data efficiently. Basically, data mining is an application-dependent and may require different mining techniques [1].

ZHAI Lianga et al [2007], discussed a new algorithm "T-Apriori" based on the time constraint is designed and implemented on the basis of analyzing the related definitions and wide-ranging of temporal association rule mining. The concepts of ecological event and sequence of biological events are proposed and their problems of temporal association rule mining based on a sequence of biological events are occurred [4].

Overview of Existing system:

In existing scenario, the RANWAR algorithm is used to avoid the more numbers of frequent item sets. Association rule mining algorithm is a prominent approach to discover the strong relationship among genes or items from the dataset. In this scenario, they added the concept of weighted rule mining approach for two measures such as support and confidence factors. The main aim of the RANWAR algorithm is generates the less number of frequent items based on the rank value. Association rule mining performs better than preceding algorithms in terms of higher efficiency and accurate results. In this scenario, consider the two novel rule measures such as weighted condensed confidence (wcc) and weighted condensed support (wcs) based on the independent of genes of microarray dataset.

Consider in this scenario that m and n where m identifies the sample and n identifies the genes in the specified dataset. The allocated weights to genes are considered to be w and this weight is attached to every gene. The pair of genes and weights represented by {g, w} which is named as weighted gene. Initially they have to discover the differentially expressed genes based on the ranking method. It performs the normalization process such as preprocessing and smoothing on the specified dataset. Then as per RANWAR rule they have to allocate the weight for each gene and perform the data discretization. Then discover the minimum support count value as well as confidence value for each gene items by using weighted rank values. According to this value, they can identify the most top ranked frequent item sets and filter the least frequent items. [6]

Challenges in the Existing Scenario

- Outlier data items are not eliminated.
- It consumes more time.
- It reduces the overall system performance significantly.

3. PROPOSED SCENARIO

In this proposed scenario, Temporal Apriori algorithm is implemented to handle the huge gene database efficiently. In this research, the gene database which includes Sequence Name, mcg, gvh, lip, chg, aac, alm1, alm2are considered. Analyze the temporal database using time threshold. Time range can be expressed as min_t_s and min_t_c for support and confidence respectively. Then apply the T-Apriori algorithm to generate frequent item sets and corresponding temporal association rules. In many real time applications, the time information exists in their data that is to say their data has temporal relativity. This algorithm is also includes the pre-processing which associates numeric value along with discrete values to increase the performance of the scenario. This algorithm is focused on the identifying the more number of dangerous gene among the total genes in the dataset. It discovers the minimum support count value and based on this values they can consider further gene which satisfies the support count and eliminates which gene support count is lesser than the support count. This proposed Apriori algorithm is retrieved the most frequent genes as well as influential one in the ranked association rule mining scenario. Hence this scenario is able to recognize the time datasets more superior with less computation time.

Features of Proposed Scenario

- Outliers are eliminated by preprocessing.
- It reduces the execution time.
- It generates rules that are not in existing but are highly biologically significant to related diseases.

4. METHODOLOGY

The modules in the current work are listed below:

- Preprocessing
- Ranking Process
- Conveying Weight to Each Gene

- Identification of Frequent Item set
- Data clustering

4.1. Preprocessing

The outliers are identified and removed during preprocessing operation. Thus the size of the database is significantly reduced. Without preprocessing the dataset takes long time for computation and it does not provide efficient frequent item set.

4.2. Ranking Process

This scenario, the rank is computed for each gene and allocates the top ranked genes in the dataset. Hence the ranking of rules from the medical dataset is very prominent to provide the most relevant and useful results. To accomplish this process, the two novel measures are introduced such as support and confidence. Yet, the problem is generating more number of frequent items. It leads more amount of time for computation process. So overcome this issue to introduce in this module, named ranked or weighted rule mining from the gene dataset. This is called as rank based weighted association rules. This RANWAR rules are extracting the more frequent genes with static dataset. For the best and worst cases it will predict the genes only depend on the ranking method. There are different expression values computed through p-values in any kind of statistical test data. If all the genes are dependent, then p-values of the genes in the test will be invalid. Therefore, on the basis of the independency of the genes of a microarray dataset, have to determine item set and transaction weight. The weight of a gene is calculated through p-value ranking of the gene. Thus, item sets-transaction weight can be defined as multiplication of weights of all the genes(items) of the item set in a transaction for a microarray dataset. It is estimated as:

$$W_k(Z)=\prod_{i=1}^{Q} \forall g_i \in Z, Q = |Z|\, wk_i$$

Where $W_k(Z)$ denotes item set-transaction weight of item set for k-th transaction.

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

g_i - Every gene in gene dataset

w_i, - Weighted gene

wk_i - Weight of gene gi for the kth transaction.

4.3 Conveying Weight to Each Gene

In this process, weight value is assigned for each gene in the gene expression dataset. At first, some pre-filtering process is applied on the data (viz., removal of genes having low variance). In fact, due to the low discrepancy of the gene, sometime lower p-value is formed 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 have very short variance. The drinkable data should be normalized gene-wise as normalization converts the data from different scales into a common scale [1].

In this approach, all the genes have not same importance. Hence some weight is assigned to each gene with respect to their p-value ranking mentioned earlier. Here, the weights of the genes are calculated in that difference between the weights of any two repeated ranked genes are same, and the weighted of the first ranked gene is always 1. The ranges of weight lie in between 0 and 1. Suppose, is number of genes. Thus, the weight of each gene estimated from a function of the above rank and number of genes as described below

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

4.4 Data clustering

Suppose the input data matrix. Here, denotes genes, and denotes samples. First of all, the matrix is transposed. At this time discretization of the input data matrix is mandatory for applying association rule mining. In clustering purpose utilized standard k-means clustering algorithm. But, before using k-means, by choosing initial seed values for k-means. For doing this at first, choose the first cluster center uniformly at random from all the data points (X). Thereafter, for each data-point, have to calculate the distance between the data-point and nearest center which is already chosen. To discover the probability function based on the following formula:

$$D(y')^2/(\sum_{y \in X} D(y)^2$$

Repeat the last step until to select the number of cluster centers. In this way, the initial centers are chosen that are used for the standard k-means clustering. The cluster have a higher centroid value is the cluster of up-regulated genes and the other cluster is down-regulated genes.

4.5 Identification of Frequent Item set using RANWAR

Consider the input data matrix as genes, samples, original gene dataset, and ranked gene list, minimum support count as well as confidence value

1. Procedure RANWAR
2. Normalize the data matrix D using zero mean normalization
3. Compute the gene rank
4. Allocate the weights for all genes in a dataset
5. Transpose the normalized data matrix
6. Select the initial seed values using k means clustering'
7. Discretized the transposed matrix applying standard k-means clustering sample-wise.
8. Apply post discretization method
9. Initialize k=1
10. Discover frequent 1-itemset
11. Repeat the process
12. K=k+1
13. Generate the candidate itemsets
14. For each candidate itemsets, c$\in CI_k$
15. Compute the ecs value
16. If wcs>=min_wsupp then
17. $FI_k \leftarrow [FI_k; c]$
18. Generate the rule from frequent itemset
19. Discover the wcc
20. For each rule do
21. If wcc(r)>= min_wconf then
22. Save the results
23. Rulesupp←wcs(r) and Ruleconf←wcc(r)
24. End if
25. End for
26. End if
27. End for
28. Until(FIk=null)
29. End procedure

Temporal Apriori Algorithm

This research, considers the gene database which includes Sequence Name, mcg, gvh, lip, chg, aac, alm1, alm2. Appling the T-Apriori algorithm to the generate frequent itemsets and corresponding temporal association rules.

T-Apriori Algorithm

1. Consider the input as min_s (minimum support threshold), and T (temporal database)
2. For all Record Sets do which means all Record Sets belongs to a temporal database (T).
3. Get a temporal record sets for satisfying the time threshold.
4. End
5. Calculate if Itemsets = threshold Sets and delete time information.
6. C_1={Candidate 1-itemsets},$L_1 = c \in C_1 | c.count \geq \min\_s\}$
7. For(k=2, $L_{k-1} \neq \emptyset$; k++) do begin and this process is until no more frequent itemsets generate
8. C_k=Apriori_Gen(L_{k-1}) generate k-itemsets candidate frequent itemsets.
9. For all transaction t belongs to Itemsets do begin
10. C_t = subset (C_k,t)
11. For all candidates c$\in C_t$ do
12. Obtain the support count for all candidate frequent item set. It implies c.count++.
13. End
14. L_k= c$\in C_k$|c.count>=min_s;
15. End
16. Finally can obtain the result as = $\bigcup_k L_k$

5. Results and Discussion

In computation, the algorithms are estimated to reduce the accuracy. For number of files the existing and proposed algorithms are executed in various accuracy values. The less time execution values called higher performance in the scenario which is provided by using proposed algorithm.

Accuracy

The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. An accuracy of 100% means that the measured values are exactly the same as the given values.

Table-1 Existing and Proposed Accuracy values

Dataset	No. of instances	No of frequent item set	Ranwar Accuracy (%)	Apriori Accuracy (%)
Ecoli	336	5	87	92
Glass	214	9	84	90
Spect	267	14	83	87
Digit	756	12	85	89

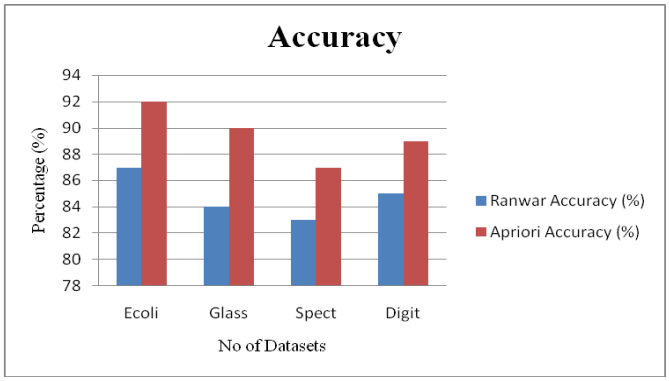


Figure-1 Accuracy

From the above graph, comparison of existing and proposed system in terms of accuracy metric is observed. In the x axis is to plot the Number of datasets and in y axis is to plot the Percentage. In existing scenario, the accuracy values are lower by using RANWAR algorithm. The accuracy value of existing scenario by using Ecoli dataset is 87%. In proposed system, the accuracy value is higher by using the temporal Apriori algorithm is 92%.The accuracy value of existing scenario by using Glass dataset is 84 %. In proposed system, the accuracy value is higher by using the temporal Apriori algorithm is 90%. The accuracy value of existing scenario by using Spect dataset is 83 %. In proposed system, the accuracy value is higher by using the temporal Apriori algorithm is 87%.The accuracy value of existing scenario by using Digit dataset is 85 %. In proposed system, the accuracy value is higher by using the temporal Apriori algorithm is 89%.From the result, concluded that proposed system is superior in performance.

6. CONCLUSION AND FUTURE WORK

The massive number of evolved rules of items (or, genes) by ARM Algorithms makes confusion to the decision maker to choose the top genes. Therefore, in this research proposed two novel rank-based weighted condensed rule-interestingness measures wcs and wcc. A weighted rule mining algorithm Apriori has been developed using the procedures for Gene data. Temporal Apriori algorithm is basically weighted updated form of RANWAR. This research uses gene dataset and compares the act of RANWAR with the state-of-the-art ARM algorithms [6]. Temporal Apriori algorithm generates less number of item set compare with RANWAR. Another advantage of Apriori is that some most biological significant rules stand top here which hold very low rank in RANWAR. Some crest rules extracted from RANWAR that are not present in Apriori, but have high biological significance, are also reported.

FUTURE WORK

Research study is the complete in-depth analysis on a specific area. The Research will have impact on the future and is an on-going activity that never ends. The research work can be enhanced with the following features:

- Extracting attractive patterns and rules from gene alteration datasets can be important in identifying cause of gene tumours and diseases.

- To develop the optimization algorithm to progress the ranked gene dataset as well as item dataset more optimally.

REFERENCES

[1]Anandhavalli.M,Ghose.M.K, and Gauthaman.K, "Association Rule Mining in Genomics," Int. J. Comput. Theory Eng., vol. 2, no. 2, pp.1793–8201, 2010.

[2]Arumalla Nagaraju, Yallamati Prakasarao, A.Veeraswamy "An Implementation of Mining Weighted Association Rules without Pre-assigned Weights" International Journal of Advenced Research in Computer Science and Software Engineering 2(8), August- 2012, pp. 276-283.

[3]Kalpanadevi.M, Usha rani.M "Weighted association rule mining- a review" Vol 04, Special Issue01; 2013.

[4]Liang, Zhai, et al. "Temporal association rule mining based on T-Apriori algorithm and its typical application." Intl. Symposium on Spatial-Temporal Modeling Analysis. Vol. 5. No. 2. 2005.

[5]Ms.Shweta, Dr. KanwalGarg "Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms" International Journal of Advanced Research in Computer Science and Software Engineering 3(6), June - 2013, pp. 306-312.

[6]SauravMallik, AnirbanMukhopadhyay and UjjwalMaulik,"RANWAR: Rank-Based Weighted Association Rule Mining From Gene Expression and Methylation Data"IEEE transactions on Nanobioscience, vol. 14, no. 1, January 2015.

[7]Fayyad. U., Uthursamy.R. , "Data mining and Knowledge discovery in Databases", Communication of the ACM, Pages 24-27, [1996].