International Journal of Soft Computing 8 (4): 239-246, 2013
ISSN: 1816-9503

An Efficient Entropy Weighted Deviation Approach to Simplify Weighted Association Rules in Bio Medical Applications

[1]B. Gomathy, [2]A. Shanmugam and [2]S.M. Ramesh
[1]Anna University, Coimbatore, India
[2]Department of ECE, Bannari Amman Institute of Technology, Erode, India

Abstract: In today's era prediction of diseases is made in the way of data mining that is providing big deal of finding hidden patterns in large data bases. This prediction is made on the basis of association rules which are large in number and consists of many looping of rules. This increases the complexity of the whole system to arrive a simple result. This association rules compares the parameters of tested data of a patient such as risk factors and arrives results of presence of disease by following some predictive rules which consists of branches in decision tree structure. So, researchers generate system that works on entropy weighted deviation approach having effective resource allocation and utilization of data in minimum cost. In future, the system provides an adept methods disease prediction in various bio-medical applications.

Key words: Prediction, looping of rules, association rules, Entropy Weighted Deviation (EWD), bio-medical

INTRODUCTION

Emerging biomedical applications are characterized by the need to process, analyze and potentially integrate disparate datasets, published by the wider community. In the concept of data mining, association rules is a popular and well researched method for discovering interesting relations between variables in large database. Moreover, association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps. The First step is to provide minimum support is applied to find all frequent item sets in a database. And the second, these frequent item sets and the minimum confidence constraint are used to form rules. While the second step is straightforward, the first step needs more attention. Finding all frequent item sets in a database is difficult since it involves searching all possible item sets. The set of possible item sets is the power set over I and has size 2^{n-1}. Although, the size of the power set growsexponentially in the number of items n in I, efficient search is possible using the downward-closure property of support (Kotsiantis and Kanellopoulos, 2006) which guarantees that for a frequent item set, all its subsets are also frequent and thus for an infrequent item set, all its supersets must also be infrequent. In the system, the concept of pruning plays a vital role in data simplification which involves in trimming up of data and eliminating the insufficient combinations provided by the association rule.

The basics of association rule are framed as follows: Let I = I1, I2, …, In be a set of m distinct attributes, T be transaction that contains a set of items such that T⊆I, D be a database with different transaction records Ts. An association rule is an implication in the form of X⇒Y where X, Y⊂I are sets of items called item sets and X∩Y = ϕ. X is called antecedent while Y is called consequent, the rule means X implies Y. There are two important basic measures for association rules, support (s) and confidence (c). Since, the database is large and users concern about only those frequently purchased items usually thresholds of support and confidence are predefined by users to drop those rules that are not so interesting or useful. The two thresholds are called minimal support and minimal confidence, respectively. Support (s) of an association rule is defined as the percentage/fraction of records that contain X×Y to the total number of records in the database. Suppose the support of an item is 0.1%, it means only 0.1% of the transaction contain purchasing of this item. Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain X×Y to the total number of records that contain X. Confidence is a measure of strength of the association rules, suppose the confidence of the association rule X×Y is 80%, it means that 80% of the transactions that contain X also contain Y together. In general, a set of items (such as the antecedent or the consequent of a rule) is called an item set. The number of items in an item set is called the length of an item set. Item sets of some length k are referred to as

Corresponding Author: B. Gomathy, Anna University, Coimbatore, India

k-item sets. Generally, an association rules mining algorithm contains the following steps. The set of candidate k-item sets is generated by 1-extensions of the large (k-1) item sets generated in the earlier iteration. Supports for the candidate k-item sets are generated by a pass over the database. Item sets that do not have the minimum support are discarded and the remaining item sets are called large k-item sets. This process is repeated until no larger item sets are found. The AIS algorithm was the first algorithm proposed for mining association rule (association rules mining). In this algorithm only one item consequent association rules are generated which means that the consequent of those rules only contain one item for example researchers only generate rules like $X \cap Y \Rightarrow Z$ but not those rules as $X \Rightarrow Y \cap Z$. Although, data is separated into different and more complex tables during normalization, the process of normalizing a database can help to organize data more efficiently by minimizing redundancy and providing more accurate records. During the process, column and field names are consolidated into more specific ones to avoid repetition of data. In many cases, tables are divided into two or more tables and linked via a relationship using their primary keys and/or foreign keys. The main goal is to allow each table to be updated individually without directly affecting the information contain within the fields of other tables.

In addition, the normalization process ensures that rows are not identical, despite their order and they have precise data about an entity. All existing columns are unique and they can be in any other. Each column has entities of the same type and their attributes. Moreover, the tables clearly state what kind of information is required and their cells only (Obuchi and Stern, 2003). The ability to identify gene mentions in text and normalize them to the proper unique identifiers is crucial for down-stream text mining applications in bioinformatics. Researchers have developed a rule-based algorithm that divides the normalization task into two steps. The first step includes pattern matching for gene symbols and an approximate term searching technique for gene names. Next, the algorithm measures several features based on morphological, statistical and contextual information to estimate the level of confidence that the correct identifier is selected for a potential mention (Lau *et al*., 2007; Koh and Pears, 2008). The main drawback of the AIS algorithm is too many candidate item sets that finally turned out to be small are generated which requires more space and wastes much effort that turned out to be useless. At the same time this algorithm requires too many passes over the whole database. Apriori is more efficient during the candidate generation process (Agrawal and Srikant, 1994).

Apriori uses pruning techniques to avoid measuring certain item sets while guaranteeing completeness. These are the item sets that the algorithm can prove will not turn out to be large. However, there are two bottlenecks of the apriori algorithm. One is the complex candidate generation process that uses most of the time, space and memory in bio medical applications and all the derivation are discussed.

In the proposed methodology, the two phases of entropy comes into effect that combined to give the pruning rules. The pruning process eliminates the insufficient combinations of data to predict the results and shorten the combined data set to conclude the process firmly. Hence, the system provides us with accurate and optimised result with minimum time and cost.

LITERATURE REVIEW

Association rules are statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. Association rules are created by analyzing data for frequent patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the statements have been found to be true (Agrawal *et al*., 1993).

As researchers say support deals with the significance of the dataset whereas confidence deals with the measure of strength of the association rules. The traditional method of association rule mining is enhanced with the items having weight to frame the weighted association rule. The ultimate goal of the weighted association rule is to focus the mining process on the basis of significant relationships involving items with significant weights which in turn reduces the flooding combinational explosion of insignificant relationships.

Another method described by Zhou *et al*. (2007) deals with the lucene index. Lucene is a high performance, full featured, text search engine library developed using java. It is a scalable indexing technique that creates index for transactional database. This process involves in increasing the speed of ARM whereas the efficiency will be not sufficient (Zhou *et al*., 2007).

Transaction clustering is used to generate rare association rule in the study research (Koh and Pears, 2008). This is concept of pre-processing the dataset by clustering to express their own combinations without interference (Koh and Pears, 2008). This process is the combination of clustering and Association Rule Mining (ARM). The approach involves homogeneous cluster

formation and generates rules from each. The limitation of this approach is the weight consideration of itemset. For further efficiency the of the item set to be calculated by assigning low frequency item set as higher weight.

The accuracy of classification system can be improved by the application of ARM with it. The WARM combined with classifiers to improve the prediction accuracy (Soni *et al*., 2009). Pruning is the process to improve accuracy and to reduce over fitting in the mining process. There are two types of pruning methods, first is pre-pruning and the next is post pruning (Patil *et al*., 2010).

Pre-pruning is the process of building the decision tree and simultaneously checking g whether tree is over fitting based on different measures like Laplace error. Post-pruning is the concept in which the tree is built first and then the reduction of branches and level of decision tree is done. Tree pruning methods convert a large tree into a small tree, making it easier to understand. Such methods typically use statistical measures to remove the least reliable branches, generally resulting in faster classification and an improvement in the ability of the tree to correctly classify independent test data.

DAST is the technique of processing the dataset towards disease prediction. The process of association rule mining tends us to meet certain complications like discovery of large number of rules which comprises some irrelevant information makes the process complex and that makes the research slow. DAST is a simple measure of algorithm to find the strength of association among attributes and to find the occurrence of attribute association based on various diseases. The main motive of DAST is to calculate the association strength and eliminate insufficient combinations (Srinivas *et al*., 2012). In the proposed research, the implementation of new concept called Entropy Weighted Deviation (EWD) to make the rule generating process in more efficient and optimized manner in resource allocation and data utilization.

PROPOSED SYSTEM

The proposed system is progressed initially with preprocessing the dataset. The preprocessing is accomplished on applying normalization methodologies. Subsequently the intrinsic features are extracted from the dataset. The feature extraction is carried based on deep analyzing the dataset. The feature extraction varies with respect to the dataset.

For instance, in a heart diseases dataset, entirely it is consisting of 76 attributes but researchers have considered only 14 significant attributes. Subsequently

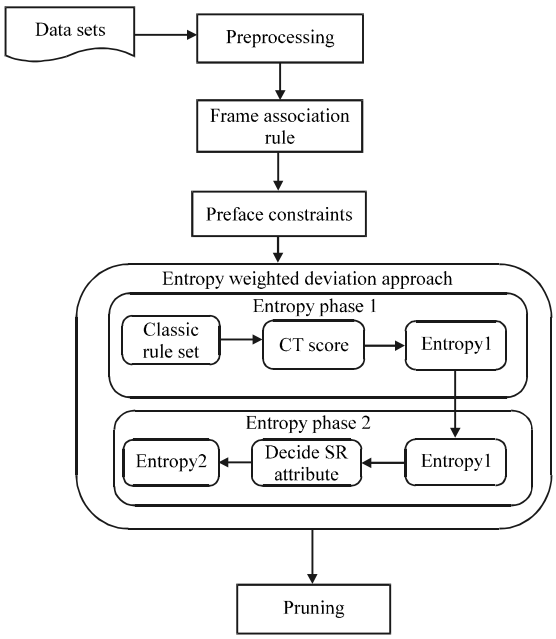

Fig. 1: System architecture

proceeds with the preface constraint in this study the preface coefficients are assigned for each intrinsic attribute that is for consideration. The preface constraint is the judgment factor for which the assignment is made for each rule (Fig. 1).

Entropy weighted deviation approach (phase 1): Association rules are framed considering each and every intrinsic attribute. There will be some set of rules framed as per the number of attributes. Let R be a rule such that it consist of combination of n attributes.

The rules are formed on this basis. Consider the feature set {f1, f2, f3, ..., fn} and Transaction set {T1, T2, T3, ..., T4}. Every transaction represents the uniqueness by providing a unique ID or any primary constraints. A rule is an inference of set of features says $X \Rightarrow Y$ which is mutually disjoint such that $\{f1||f2\} \Rightarrow \{f3\}$ where $\{f1, f2\} \in A$ and $\{f3\} \in B$.

Let F be preface constraint for the corresponding features such as {CP = 3, restbps = 150, chol = 220, thalack = 150, exang = 1, oldpeak = 2, slope = 2, ca = 1, thal = 2}. Any set of F represents the sub feature set denoting certain peculiar attributes for confirming the diseases. But for the sake of conformity minimum five attributes are considered for disease prediction. Let J be the disease set under examination. The support value for J defined by support (J) is exemplified as the proportion of records in dataset satisfying J.

For instance there can exist a subset such as {thalack = 150 $\Rightarrow$ exang = 1, oldpeak = 2} here the

support is encountered on separating the rule with LHS and RHS. There are set of features in A and B such that the count of either A or B is considered as A∪B. Consider a rule R ⇒ {restbps, chol} ⇒ thalach. The preface constraint is prefixed as {restbps = 150, chol = 220} ⇒ {thalach = 150}.

The rules are pruned after framing rules based on number of features because the prediction of diseases is based on at least n number of features. Moreover, there are surplus rules which cause large number of search to identify the possibility for an apt rule. Hence, initial pruning process is carried out. Next encounters the entropy weighted deviation approach phase 1. In this phase, researchers are going to estimate the Entropy1 score (E1). The algorithmic phase is given below. Each rule is considered and the attribute sets are extracted individually and its corresponding preface constraint is assigned to each attribute in the rule set (Fig. 2).

The new rule set is formed comprising the individual attributes which is called as Classic Rule set (CR set). The system architecture represents the overview of the project proposal and its clear flow of the system. As researchers mentioned before, the entropy1 combines with entropy2 and given for pruning to get the efficacy and accurate results, e.g., the individual attributes are restbps, chol, exang, chol. The classic rule set is casing on its corresponding preface constraint {restbps = 150, chol = 220} ⇒ {exang = 1, oldpeak = 2} following for each attribute in the rule set on considering each CR set, the rows supporting the particular rule are extracted and its corresponding Weight (W) is gathered individually for each attribute. The Support Count (SC) is also found for each attribute. The resultant weights are added independently. Finally, the Contribution (CT) score is estimated:

$$d = \frac{W_k}{SC_k} \tag{1}$$

where, k indicates the rows in the dataset.

$$M = \sum_{i=1}^{F} d_i \tag{2}$$

Where:

F = Corresponding features

d = The weighted mean score for an attribute in the rule

This is computed in order to estimate the individual contribution of each attribute in the rule comprising the CR set. A maximum entropy probability distribution is a probability distribution has its wide range of its application in measuring the statistical inference. The possibility of particular outcomes for that rule is to be measured. Hence, researchers apply the combined CT score for predicting the likelihood of the rule. Besides that this distribution is appropriate for probability density estimation of some fussy constraints such as average score. For the sake of accuracy the distribution is afforded with the criteria:

$$E1_k = \log\left(\sum_{i=1}^{F} d_i\right) \tag{3}$$

Here, k denotes the process is repeated for each rule.

Entropy weighted deviation approach (phase 2): In the EWD approach the term weighted deviation arrives because the mean square weighted deviation is applied. Researchers are using the bio medical data; the rules must predict the features combination in which researchers can predict their involvement with their corresponding features with that rule. Hence, weighted deviation statistics is pertinent. There is another deterministic criterion to be found that is the S attribute called as significant attribute. The weighted deviation is applied to that significant attribute. This is to measure its contribution with the other attributes in the rule against the other features (Fig. 3).

S attribute selection based on linear regression: This S attribute selection is based on the importance of its role in

Algorithm 1: Entropy weighted deviation approach: phase I

Input: Rule set along with the preface constraint.

Output: Entropy1

Begin: For each rule, set its preface constraint

- Frame classic rule set/{A, B}⇒{C} such that {A = a|| B = b1}⇒{C = c1}
- Scan the database supporting the classic rule set.
- For each item in CS set
  - Accumulate the weight of items individually
  - Compute the CT score (M) for each item in the rule set.
  - Evaluate the combined CT score for all items in rule

  End for
- Examine the entropy1

End for

End.

Fig. 2: Algorithm 1 for entropy1 calculation

Algorithm 1: Entropy weighted deviation approach: phase II
Input: Rule set along with the preface constraint, entropy1
Output: Entropy2
Begin: Decide the S attribute
For each rule,
Scan the database supporting the classic rule set.
For each item in CS set
- Accumulate the weight of items individually
- Get the S Attribute score for each row
- Examine the deviation score
- Estimate the CT score

End for
Examine the entropy2.
End for
End.

Fig. 3: Algorithm 2 for entropy2 calculation

making right decision with that particular rule. Hence, researchers have used linear regression for judging the correct S attribute.

Other than the intrinsic attributes vales researchers need to originate the S attribute. For this, researchers consider the each attribute with the weighted deviation of any intrinsic attribute weights within the rule. The coefficient of determination is applied to foretell the dependent variable (Y)'s significance with independent variable (x). In other words, the future of Y with to X:

$$R^2 = \frac{\left(\frac{1}{N}\right) \times \sum \left[(x_i - \text{mean}(x)(y_i - \text{mean}(y)\right]}{\left(\sigma_x \times \sigma_y\right)} \quad (4)$$

Where:
σ_x = The standard deviation of x
σ_y = The standard deviation of y

Based on highest confidence threshold R^2 decision is made whether it can be undertaken for rule prediction or not.

Algorithm (phase 2): The algorithm proceeds as follows: Initially E1 score intended for forecasting how the attributes contribute their support constraint to that specific rule. This E1 score is applied in EWD phase II. The weight is extracted from each row. Another crucial parameter is the mean support count. It is the ratio between all attributes in the set supporting the rule with the total number of attributes adapted to the rule.

The mean support count (MX) is intended for S attribute for measuring its involvement for that rule too. While each attribute supports that rule, the S Attribute Score (SA) is also computed. X_i Implies the S attribute weight for each row supporting the rule its respected:

$$SA = X_i - MX \quad (5)$$

Next, it is needed to determine the deviation score. Deviation score is the score how the rows supporting from the mean score:

$$DS = \frac{W_i \times SA \times SA}{\sum_{i=1}^{n} d_i} \quad (6)$$

Moreover, the total weights are computed for each individual intrinsic attributes. As per Eq. 6, the weight of each item is multiplied with SA and there by substituting Eq. 1 in d_i. Final CT score is computed as the equation:

$$M_k = (\sum_{i=1}^{F} DS_i) + El_k \quad (7)$$

Here:
F = The number of attributes in a rule set
$E1_k$ = The entropy value for each rule

Again applying for calculating entropy2:

$$E2_k = \log(M_k) \quad (8)$$

The above said procedure is repeated for each rule. The procedure is eternally based on maximum entropy probability distribution where the k value has n number of distribution to calculate the value for entropy2.

Rule elimination principle: Pruning is the concept of trimming up of data in an effective manner by eliminating the insufficient combinations. It is more effective to include the concept of pruning in association rule to find the significant combinations of dataset for disease prediction.

Initially the system enters on framing surplus rules on selection of each attribute. Then, assigning preface constraints, subsequently framing the classic rule sets. On scanning the rows, the CT score is computed. Following it is repeated for each feature, finally for each rule.

Along with the entropy1 for each rule, the system enters into the II phase, next it is needed to determinate the significant attribute. Each rule is extracted and its respected S attribute score is evaluated then deviation score. At the end researchers will compute entropy2 followed by the entropy2.

When researchers combine the results of entropy1 and 2 into the pruning algorithm, the acceptable threshold value has to be maintained approximately as 0.4. It is

determined as decisive factor (df). For the pruning methodology, researchers have considered a value limit for decisive factor which is given as $0 \leq df \leq 0.5$. The P factor is the pruning factor which is defined by the following equation:

$$\text{P factor} = \text{Entropy 2} - 0.4 \quad (9)$$

After calculating these values, there is a comparison factor to be referred. If the value of P factor is greater than or equal to the error factor then the result is taken into consideration, else the process gets eliminated. Thus, the process results with an efficient error predicting methodology.

EXPERIMENTAL RESULTS

The theoretical point of views are put forth to prove efficiency standards in pruner is exponent when compared to the predecessors. Researchers have presented more accuracy in terms of efficiency statistically. The consideration is made in order to show the advancement of every parameter like processing time, pruning throughput, entropy and efficiency.

The proposed system initiates with dataset preprocessing. Which is accomplished on applying normalization methodologies? The linear transformation of original data has been performed by the normalization min-max process. Consider min_a and max_a be the minimum and the maximum values of an attribute A. Normalization min-max will map the value v, A into ύ in the range [new_min_a, new_max_a] by using the equation:

$$\acute{\upsilon} = \left[\frac{(v - min_a)}{(max_a - min_a)}\right](new\_max_a - new\_min_a) + new\_min_a$$

The feature extraction is completely proportional to dataset. For instance in a diabetic dataset, entirely it consists of 76 attributes but researchers have considered only 14 significant attributes. Those significant attributes are much enough for the prediction algorithm. Subsequently proceeds with the preface constraint in this session the preface coefficients are assigned for each intrinsic attribute that is for consideration. While performing the EWD approach, there are two phases of entropy will be combined to be given as the input for pruning mechanism. When the pruning rule comes into effect, it has taken over the efficiency and optimization part of the algorithm. The following graphical representation illustrates the concept of the algorithm. Researchers have illustrated the graph for prediction of diseases such as heart diseases, diabetics, breast cancer and obesity so on. The dataset information for the project demo has been acquired from the universal datasets like UCI machine learning repositary. Lets have a look into the performance (Fig. 4).

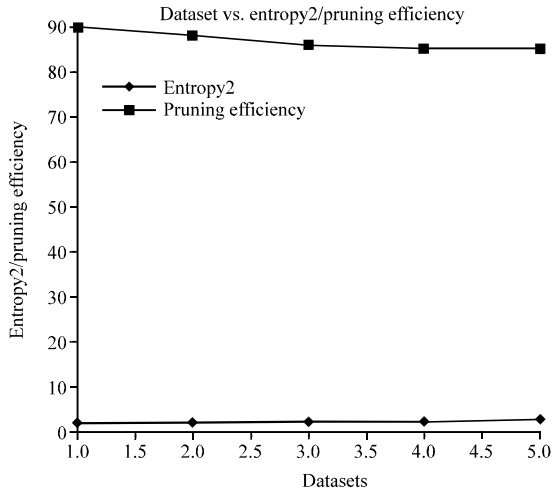

Fig. 4: Portrays pictorial representation between datasets, pruning efficiency and the corresponding entropy2

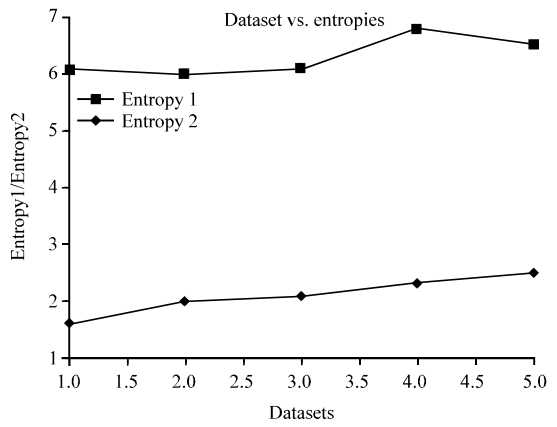

Fig. 5: Pictorial representation between entropy1 and 2

Figure 4 demonstrates the dataset vs. entopy2 pruning efficiency in which the pruning efficiency corresponds to the dataset values. In the data set when the values of dataset increases the corresponding values of entropy2 increases and the pruning sufficiency gets decreased. In the dataset for the dataset value 1, the pruning efficiency will be 90% and for 2, the pruning efficiency will be 88%. Likewise, the algorithm moves on. Thus, the pruning efficiency is inversely proportional to the dataset values. Pruning is an adept way to reduce space between the data in the large database (Fig. 5).

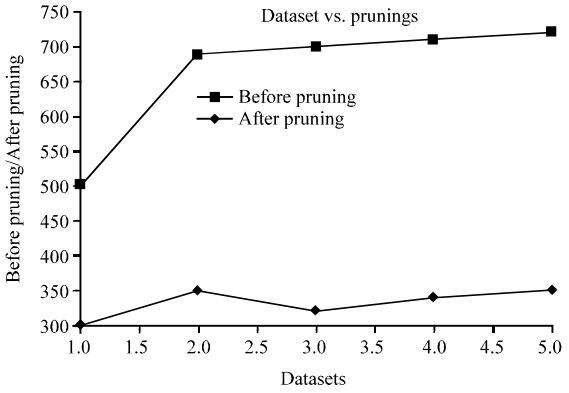

Fig. 6: The comparison between the significant data before and after pruning

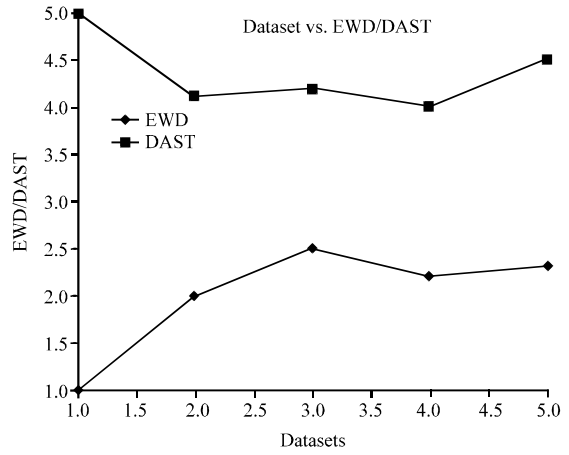

Fig. 7: Pictorial representation between existing approach DAST-Proposed EWD

With the above representation, researchers have made the comparison between the values of entropy1 and 2. Figure 5 researchers can visualize that the value of entropy2 will get decreased with the corresponding increasing value of entropy1 and dataset (Fig. 6).

The earlier graph shows the comparison between the pruned data and the no pruned data. From the graph (Fig. 6), it is legible that the pruned data has significant combination of prediction algorithm and it is more efficient. Pruning is an efficacious attempt that offers us an optimized system for the disease prediction mechanism by reducing the memory space and with the minimum cost framework.

Figure 7, it is obvious that the proposed system is more adept than the existing in such a way that it reduces the processing time of creating the significant combination of datasets. EWD is an adept methodology which is accomplished with the highly potential concepts

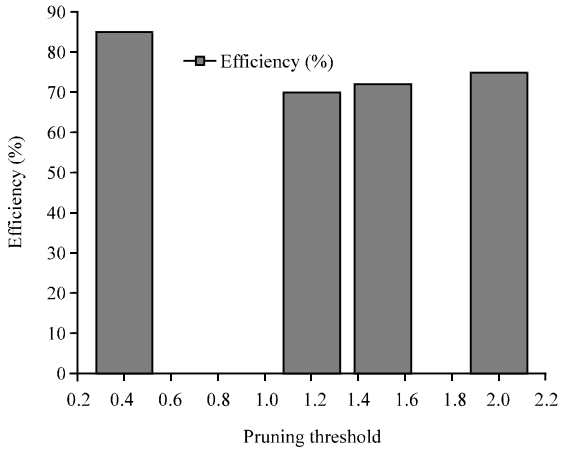

Fig. 8: Depicts pictorial representation between pruning threshold-efficiency

like maximum entropy probability distribution to calculate the two values of entropy and the pruning effect which eliminates the insufficient combinations dataset for prediction algorithm? The combination produces a persuasive methodology for the algorithm.

Figure 8 shows that the is a pruning threshold value had to be maintained which is the least acceptable value supports the algorithm. The efficiency of the system increases by maintaining the low value for pruning threshold. The pruning threshold is calculated in such a way that eliminating the inadequate data from the dataset combinations made with the association rule mechanism. Thus, the proposed system provides a potent mechanism which makes the prediction process more accurate with the efficacy of pruning.

CONCLUSION

Evaluation of resource allocation process and effective data accession policies through EWD rules proved the efficiency in optimal prediction of diseases. Thus, achieves high performance in bio medical application under voluminous data processing system. This study aims to improve the quality of pre-processed data and time consumption of data mining process with Entropy weighted deviation approach. Researchers proposed a model for pruning existing rules that multidimensional data based on simple associative analysis. Pruner is used for data pre-processing using dual entropy that generates the associative pruning rules. All the EWD approaches are well utilized to evaluate, validate and prove the improvement resource allocation and maximum accuracy, respectively. The proposed system with reliable flexibility, data centre updating

process and accuracy is successfully developed a model for high content data delivery mechanism like biomedical applications.

REFERENCES

Agrawal, R. and R. Srikant, 1994. Fast algorithm for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, September, 12-15, 1994, San Francisco, CA., USA., pp: 487-499.

Agrawal, R., T. Imielinski and A. Swami, 1993. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, May 25-28, 1993, Washington, DC., USA., pp: 207-216.

Koh, Y.S. and R. Pears, 2008. Rare association rule mining via transaction clustering. Proceedings of the 7th Australasian Data Mining Conference, November 27-28, 2008, Australia, pp: 87-94.

Kotsiantis, S. and D. Kanellopoulos, 2006. Association rules mining: A recent overview. GESTS Int. Trans. Comput. Sci. Eng., 32: 71-82.

Lau, W.W., C.A. Johnson and K.G. Becker, 2007. Rule-based human gene normalization in biomedical text with confidence estimation. Proceedings of the Computational Systems Bioinformatics Conference, Volume 6, August 13-17, 2007, San Diego, CA., USA., pp: 371-379.

Obuchi, Y. and R.M. Stern, 2003. Normalization of time derivative parameters using histogram equalization. Proceedings of the 8th European Conference on Speech Communication and Technology, September 1-4, 2003, Geneva, Switzerland, pp:189-792.

Patil, D.D., V.M. Wadhai and J.A. Gokhale, 2010. Evaluation of decision tree pruning algorithms for complexity and classification accuracy. Int. J. Comput. Appli., 11: 23-30.

Soni, S., J. Pillai and O.P. Vyas, 2009. An Associative Classifier Using Weighted Association Rule. Proceedungs of the World Congress on Nature and Biologically Inspired Computing, December 9-11, 2009, Coimbatore, pp: 1492-1496.

Srinivas, K., G. Raghavendra Rao and A. Govardhan, 2012. Mining Association Rules from Large Datasets Towards Disease Prediction. Proceedings of the International Conference on Information and Computer Networks, June 19-23, 2012, Springer, UK., pp: 334-343.

Zhou, N., J. Wu, S. Zhang, H. Chen and X. Zhang, 2007. Mining Weighted Association Rules with Lucene Index. Proceedings of the international conference on Wireless Communications, Networking and Mobile Computing, September 21-25, 2007, Shanghai, USA., pp: 3697-3700.