Journal of Computer Science 7 (4): 519-525, 2011
ISSN 1549-3636

A Descriptive Framework for the Multidimensional Medical Data Mining and Representation

Veeramalai Sankaradass and Kannan Arputharaj
Department of Information Science and Technology,
College of Engineering, Anna University, Chennai-25, Tamil Nadu, India

Abstract: Problem statement: Association rule mining with fuzzy logic was explored by research for effective datamining and classification. **Approach:** It was used to find all the rules existing in the transactional database that satisfy some minimum support and minimum confidence constraints. **Results:** In this study, we propose new rule mining technique using fuzzy logic for mining medical data in order to understand and better serve the needs of Multidimensional Breast cancer Data applications. **Conclusion:** The main objective of multidimensional Medical data mining is to provide the end user with more useful and interesting patterns. Therefore, the main contribution of this study is the proposed and implementation of fuzzy temporal association rule mining algorithm to classify and detect breast cancer from the dataset.

Key words: Data discretization, fuzzy logic, Association Rule Mining (ARM), Minimum Description Length (MDL), medical data mining, multidimensional data

INTRODUCTION

A temporal Association rule is a well established data mining technique used to discover co-occurrences of items mainly in temporal sequence data where the data items in the database are usually recorded as binary data (present or not present). Many techniques are available in the literature aims to find association rules (with strong support and high confidence) in large datasets. For example, Classical Association Rule Mining (ARM) (Mahafzah *et al*., 2009) deals with the relationships among the data items present in Multidimensional databases.

Similarly, there are a few works that focus on temporal data mining. Temporal rule mining is concerned with data mining of large sequential data sets. Sequential data (Vijayalakshmi and Mohan, 2010), mining deals with the mining of data that is ordered with respect to some index. The scope of temporal data mining (Alcalá-Fdez *et al*., 2009) extends beyond the standard forecast or control applications of time series analysis. Often temporal data mining methods must be capable of analyzing data sets that are prohibitively large for conventional time series modeling techniques to handle efficiently.

Problem statement: Due to tremendous advances and achievement in biomedical, bioinformatics, biological and clinical data is being mined at tremendous speed. Thus the biological sequence data (Hu *et al*., 2009) stored at data warehouse in the format of ultidimensional temporal sequential data can be used for finding temporal pattern (Intan and Yenty, 2008). Moreover, due to the highly distributed uncontrolled mining and use of a wide variety of bio medical data, data collection, data analysis and semantic integration of such heterogeneous and widely distributed temporal sequence data has become an important task for systematic and coordinated analysis of medical dataset (Khan *et al*., 2010). Bio medical data analysis and integration becomes very difficult due to data complexity, distribution, volume of data. Most commercial data mining products provide large number of modules and tools for performing various data mining tasks but few provide intelligent assistance for addressing many important decisions that must be considered during the mining process. Multi objective association rule mining with minimum support and minimum confidence is suitable for datamining analysis (Hamid Reza Qodmanan, *et al*., 2011).

Traditional data analysis techniques cannot support huge and complex medical data set. New data analysis technique such as data mining can be helpful in analysis large and complex medical sequence data set. Researchers may need data and knowledge which was discovered by other researchers for their research that is distributes multidimensional sequential data format. New

Corresponding Author: Veeramalai Sankaradass, Department of Information Science and Technology, College of Engineering, Anna University, Chennai-25, Tamil Nadu, India

systems are needed to manage, Integrate and analyze large &complex medical data from data warehouses. Not only the evaluation estimation and analysis of data is important but providing the intelligent assistance in equally important. Mostly analysis products do not provide the intelligent assistant in decision making process (Papageorgiou, 2011). Fuzzy association rule will be suitable for multidimensional data analysis (Hong *et al*., 2009; Weng and Chen, 2010; Wu *et al*., 2010).

In this study, we propose a fuzzy temporal (Ch. S. Reddy and KVSVN Raju, 2009) association rule mining algorithm for effective temporal datamining. These rules are further used for the classification of breast cancer data and it has been found that accurate prediction of breast cancer is possible with the proposed fuzzy temporal association rule mining algorithm.

A Bayesian network consists of a structural model and a set of conditional probabilities. The structural model is a directed graph in which nodes represent attributes and arcs represent attribute dependencies. Attribute dependencies are quantified by conditional probabilities for each node of given its parents. Bayesian networks (Khan *et al*., 2010) are often used for classification problems, in which a learner attempts to construct a classifier from a given set of training examples with class labels. Assume that A1;A2; . . .An are n attributes (corresponding to attribute nodes in a Bayesian network). An example E is represented by a vector <a1; a2; . . . ; an>, where ai is the value of Ai. Let C represent the class variable (corresponding to the class node in a Bayesian network). We use c to represent the value that C takes and c to denote the class of E. The classifier represented by a general Bayesian network is defined in (1):

$$c(E)=\arg\max_{c\in C} P(c)P(a_1,a_2,...,a_n|c)$$

Assume that all attributes are independent given the class (conditional independence assumption), the resulting classifier is called naive Bayes (Khan *et al*., 2010):

$$c(E)=\arg\max_{c\in C} P(c)\prod_{i=1}^{n} P(a_i|c)$$

In naive Bayes, each attribute node has the class node as its parent, but does not have any parent from attribute nodes. Because the values of p and P can be easily estimated from training examples, naive Bayes is easy to construct. Naive Bayes is the simplest form of Bayesian networks (Khan *et al*., 2010). It is obvious that the conditional independence assumption in naive Bayes is rarely true in reality, which would harm its performance in the applications with complex attribute dependencies. Based on the theory of Bayesian networks, Naive Bayes is a simple yet consistently performing probabilistic model. Data classification with naive Bayes (Khan *et al*., 2010) is the task of predicting the class of an instance from a set of attributes describing that instance and assumes that all the attributes are conditionally independent given the class.

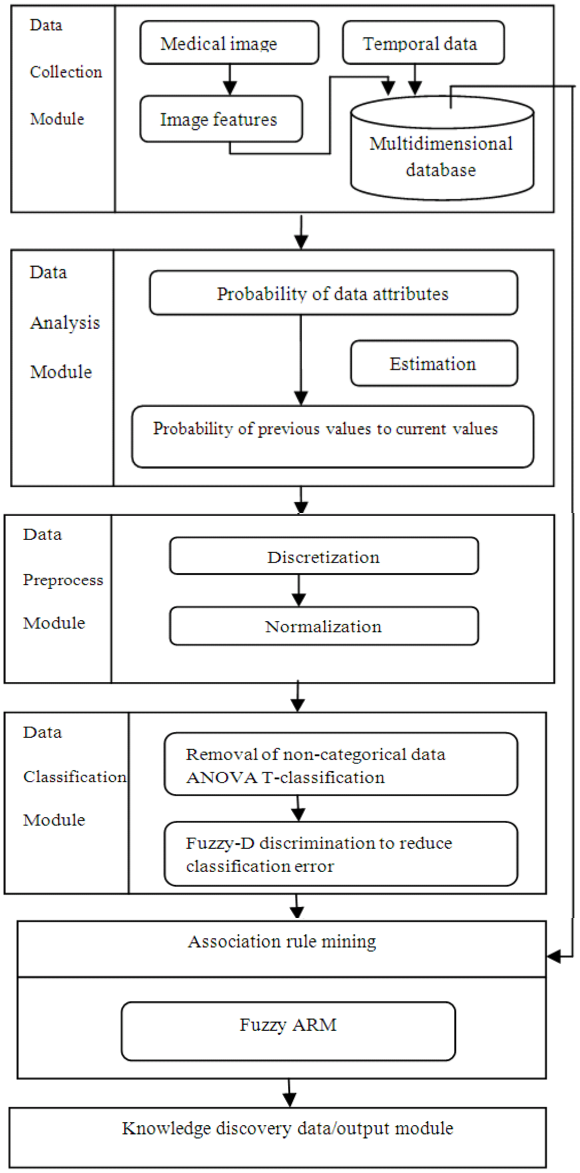
Data
Collection
Module
Medical image
Temporal data
Image features
Multidimensional database
Data
Analysis
Module
Probability of data attributes
Estimation
Probability of previous values to current values
Data
Preprocess
Module
Discretization
Normalization
Data
Classification
Module
Removal of non-categorical data ANOVA T-classification
Fuzzy-D discrimination to reduce classification error
Association rule mining
Fuzzy ARM
Knowledge discovery data/output module

Fig. 1: Architecture of intelligent data modeling.

Predicting the class of an instance are done through utility independent privacy preserving data mining by vertically partitioned data (Poovammal and Ponnavaikko, 2009). It has been shown that naïve Bayesian classifier is extremely effective in practice and difficult to improve upon.

System architecture: The complete implementation system architecture is given in Fig. 1 which include data collection modules, data analysis modules, data preprocess modules, data classification module and Association rule mining Knowledge discovered data output module with all the internal component of proposed work in details. The details of all the component of the proposed model is given in details The proposed algorithm ANOVA T classification and fuzzy D discretization is also given in details.

MATERIAL AND METHODS

Data analysis: Considering that an attribute X has a large number of values, the probability of the value P(X=xi |C=c) from Eq. 2 can be infinitely small. Hence the probability density estimation is used assuming that X within the class c are drawn from a normal (Gaussian) distribution:

$$\frac{1}{\sqrt{2\pi\sigma_c}} e \frac{(x_i-\mu_c)^2}{2\sigma_c^2}$$

Where:

σ_c = The standard deviation

μ_c = The mean of the attribute values from the training set

The major problem with this approach is that if the attribute data does not follow a normal distribution, as often is the case with real-world data, the estimation could be unreliable. Other methods suggested include the kernel density estimation approach. But since this approach causes very high computational memory and time it does not suit the simplicity of naive Bayes classification). When there are no values for a class label as well as an attribute value, then the conditional probability P(x|c) will be also zero if frequency counts are considered. To circumvent this problem, a typical approach is to use the Laplace-m estimate. Accordingly:

$$P(C=c)=\frac{n_c+K}{N+n\times K}$$

Where:

n_c = Number of instances satisfying C=c

N = Number of training instances

n = Number of classes and k=1(Predefined):

$$P(X=x_i \,|\, C=c)=\frac{n_{ci}+m\times P(X=x_i)}{n_c+m}$$

Where:

n_{ci} = Number of instances satisfying both X=xi and C=c m=2 (a constant)

P(X=xi) = Estimated similarly as P(C=c) given above

Data discretization for preprocessing: Discretization is the process of transforming data containing a quantitative attribute so that the attribute in question is replaced by a qualitative attribute (Pedreschi *et al*., 2008). Data attributes are either numeric or categorical. While categorical attributes are discrete, numerical attributes are either discrete or continuous. Research study shows that naive Bayes classification works best for discretized attributes and discretization effectively approximates a continuous variable.

The Minimum Description Length (MDL) discretization is Entropy based heuristic given by Fayyad and Irani . The technique evaluates a candidate cut point between each successive pair of sorted values. For each candidate cut point, the data are discretized into two intervals and the class information entropy is calculated. The candidate cut point, which provides the minimum entropy is chosen as the cut point. The technique is applied recursively to the two subintervals until the criteria of the Minimum candidate cut point, the data are discretized into two intervals and the class information entropy is Description Length . For a set of instances S, a feature A and a partition boundary T, the class information entropy of the partition induced by T is given by:

$$E(A,T,S)=\frac{|S_1|\,Ent(Sl)}{|s|}+\frac{|S_2|\,Ent(S_2)}{|s|}$$

And:

$$Ent(S)=-\sum_{i=1}^{c} P(C_{i,}S)\log 2(C_{i,}S)$$

For the given feature the boundary T_{min} that minimizes the class information entropy over the possible partitions is selected as the binary discretization boundary. The method is then applied recursively to both partitions induced by T_{min} until the stopping criteria known as the Minimum Description Length (MDL) is met. The MDL principle ascertains that for accepting a partition T, the cost of encoding the partition and classes of the instances in the intervals induced by T should be less than the cost of encoding the instances before the splitting. The partition is accepted only when:

$$\mathrm{Gain}(A,T,S) > \frac{\log_2(N-1)}{N} + \frac{\Delta(A,T,S)}{N}$$

Where:

$$\Delta(A,T,S) = \log_2(3^c - 2) - c\mathrm{Ent}(S) - c_1\mathrm{Ent}(S_1) - c_2\mathrm{Ent}(S_2)$$

And:

$$\mathrm{Gain}(A,T,S) = \mathrm{Ent}(S) - E(A,T,S)$$

N = number of instances, c,c1,c2 are number of distinct classes present in S, S1 and S2 respectively. MDL discretized datasets show good classification accuracy performance with naive Bayes.

Classification on ANOVA-T data selection: The proposed ANOVA-T statistical algorithm is used for Classification. Feature selection is often an essential data preprocessing step prior to applying a classification algorithm such as variance ANOVA-T.

$$Y_{ij} = M + a_i + e_{ij}$$

Here the M= Mean of Variables, a- standard deviation, e- Residual respectively.

Standard deviation:

$$\frac{a = (b-a)(b-c)}{d}$$

where d = control group.

The term feature: Selection is taken to refer to algorithms that output a subset of the input feature set. One factor that plagues classification algorithms is the quality of the data. If information is irrelevant or redundant or the data is noisy and unreliable then knowledge discovery during training is more difficult .Regardless of whether a learner attempts to select features itself or ignores the issue, feature selection prior to learning can be beneficial. Reducing the dimensionality of the data reduces the size of the data information set and more effectively. In some cases accuracy on classification can be improved. As a learning scheme naive Bayes is simple, very robust with noisy data and easily implementable.

After the proposed statistical analysis with ANOVA-T classification, attribute value represent Fig. 2 has some irrelevant data to be removed.

Fuzzy-d discretization: Fuzzy-D discretization is another proposed method. By using the Fuzzy-D discretization methods, we reduce the classification error as shown in Fig. 2, the estimation of p (ai < Xi · bij | C = c) is obtained. Because space limits, we present here only the version that according to our experiments, best to reduce the classification error. Fuzzy-D initially forms k equal-width intervals (ai; bi) (1 · i · k) using EWD (equal width). Then FD estimates p (ai < Xi · bi jC = c) from all training instances rather than from instances that have value of Xi in (ai; bi). The influence of a training instance with value v of Xi on (ai; bi) is assumed to be normal.

Pseudo code for fuzzy-d discretization:

- Training instance value v of x; on (a_{1i},b_i)
- Normal distributed mean value equal to V∝ to P(u, σ, i):

  $$P(u,\sigma,i) = \int_{ai}^{bi} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-v}{\sigma})^2} dx$$

  Here,
  σ -> parameter

- Equal width discrete divides number of line V_{min}, V_{max} into k -> intervals , w -> width

  $$W = (V_{max} - V_{min}) = K$$

- Cut points are:

  $$V_{min} + W, V_{min} + 2W..........V_{min} + (K-1)W$$

  Here,
  k = 10 ie user defined parameter
- Equal width interval:

  $$(a_i, b_i)(j = 1,........, n_c)P(V_{j,}\sigma, i)$$

  Here,
  n_c-> training instances with known value for X; class c,
- Fuzzy-D Probability estimation

  $P(a_i < x_i \leq b_i | C = c)$ to be obtained by evaluation of:

  $$P\frac{(a_i < x_i \leq b_i \wedge C = c)}{P(C = c)} \approx \sum_{j=1}^{n_c} \frac{P(V_i ,\sigma, i) / n}{P(C = c)}$$

Distributed with the mean value equal to v and is proportional to:

$$P(\upsilon,\sigma,i) = \int_{a_i}^{b_i} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-v}{\sigma})^2} d\chi$$

σ is a parameter to the algorithm and is used to control the 'fuzziness' of the interval bounds. Hence the Equal Width Discretization (EWD) (Yang and Webb, 2002) divides the number line between V_{min} and V_{max} into k intervals of equal width. Thus the intervals have width:

$$w = (V\max - V\min) = k$$

and the cut points are at:

$$V\min + w, V\min + 2w \ldots v\min + (k-1)w$$

Here the k is a user predefined parameter and is set as 10 in our experiments.

Suppose there are nc training instances with known value for Xi and with class c, each with:

$$P(\upsilon_j, \sigma, i)$$

Influence on:

$$\left(a_{i,} b_i\right](j = 1, \ldots, n_c)$$

Fuzzy-T association rule mining: Association Rule Mining algorithm preserving privacy in data analysis. Our algorithm uses two phases in a partition-approach to generate fuzzy association rules . The dataset is logically divided into p disjoint horizontal partitions P1, P2, …, Pp. Each partition is as large as can fit in available main memory. For ease of exposition, we assume that the partitions are equal-sized, though each partition could be of any arbitrary size as well.

We use the following notations:

- E = fuzzy dataset generated after pre-processing
- Set of partitions P = {P1, P2, …, Pp}
- td[it] = tid list of item set it
- μ = fuzzy membership of any item set

RESULTS

In Table 1, the efficiency of ANOVA T classification algorithm is compared with existing ANOVA classification algorithm. In Table 2, the proposed fuzzy-D discretization algorithm is analyzed by comparing with existing algorithm in calculating classification error rate. In Table 3, the proposed Fuzzy-T ARM algorithm is analyzed by comparing with existing algorithm in analyzing the performance.

Figure 2 shows the performance analysis of the proposed method by comparing with the existing. The performance of the proposed method is 5 percentages higher than the existing method.

Figure 3 shows the performance analysis of the proposed FTARM method by comparing with the existing FARM. The performance of the proposed method is faster than the existing method.

- Count[it] = cumulative μ of item set it over all partitions in which it has been processed
- d = number of partitions (for any particular item set it) that have been processed since the partition in which it was added.

The byte-vector-like data structure will represent the phase code. Each cell of the byte-vector stores μ of the item set corresponding to the cell index of the tid to which the μ pertains. Thus, the i^{th} cell of the byte-vector contains the μ for the i^{th} tid. If a particular transaction does not contain the item set under consideration, the cell corresponding to that transaction is assigned a value of 0. When the byte-vector is initialized, each cell by default has 0.

Table 1: ANOVA-T classification data attributes

	Dataset			
Attribute names	Data types	Number of data attributes	ANOVA classification efficiency (%)	ANOVA T classification efficiency (%)
Non-Relapse contig45645_RC	0.566	12	43.00	93.00
Continuous contig44916_rc	0.093	14	90.00	93.80
Continuous D25272	0.136	14	70.08	78.00
Continuous J00129	0.039	10	78.00	85.70
Continuous Contig29982_RC	0.066	12	72.00	77.00
Continuous Contig26811	0.037	10	80.00	78.00
Continuous D25274	0.146	10	78.00	79.00
Continuous Contig36292	0.389	12	78.50	84.00
Continuous Contig42854	0.270	11	77.00	84.00
Continuous Contig3396_RC	0.363	11	94.00	91.70
Continuous Contig1938_RC	0.008	8	90.05	94.05
Continuous Contig34851	0.169	8	40.00	89.00
Continuous AB033050	0.029	10	67.00	72.00

Table 2: Fuzzy-D discretization-Reduced classification error report

	Dataset			
Attribute names	Data types	Number of data attributes	Before discretization (error rate in %)	After fuzzy-D discretization (error rate in %)
Non-Relapse contig45645_RC	0.566	12	93.00	81.00
Continuous contig26811	0.037	10	78.00	73.00
Continuous contig3396_RC	0.363	11	91.70	94.70
Continuous contig34851	0.169	8	87.00	79.20

Table.3.Performance analysis of Fuzzy-T ARM

Attribute Names	Fuzzy ARM attribute values	Time (in sec)	Fuzzy -T ARM attribute values	Time (in sec)
continuous Contig44916_RC	90.00	0.45	96.98	0.40
continuous D25272	70.08	0.50	88.00	0.41
continuous J00129	78.00	0.75	85.70	0.61
continuous Contig29982_RC	72.00	0.85	99.50	0.77
continuous D25274	78.00	1.50	79.00	1.00
continuous Contig36292	78.50	2.00	98.00	1.70
continuous Contig42854	77.00	2.50	97.00	1.90
continuous Contig1938_RC	90.05	3.50	97.05	3.00
continuous AB033050	67.00	4.50	80.00	3.60

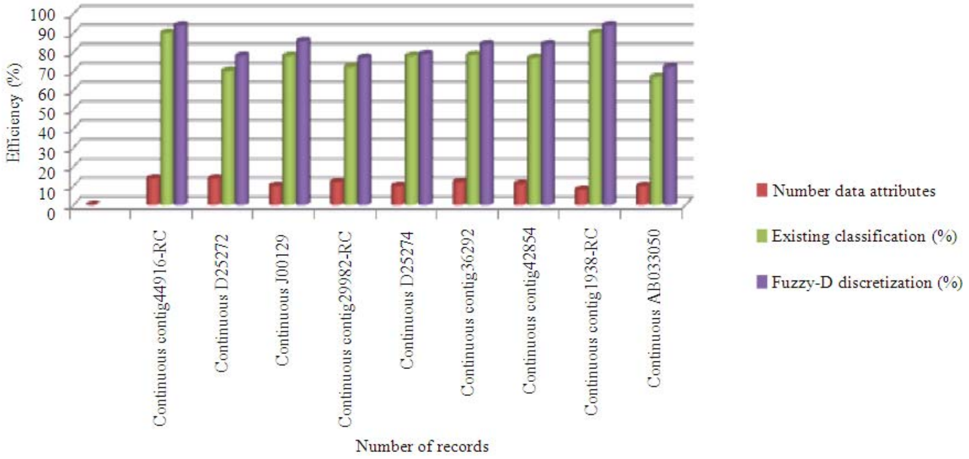

Fig. 2: Performance analysis of fuzzy-D discretization

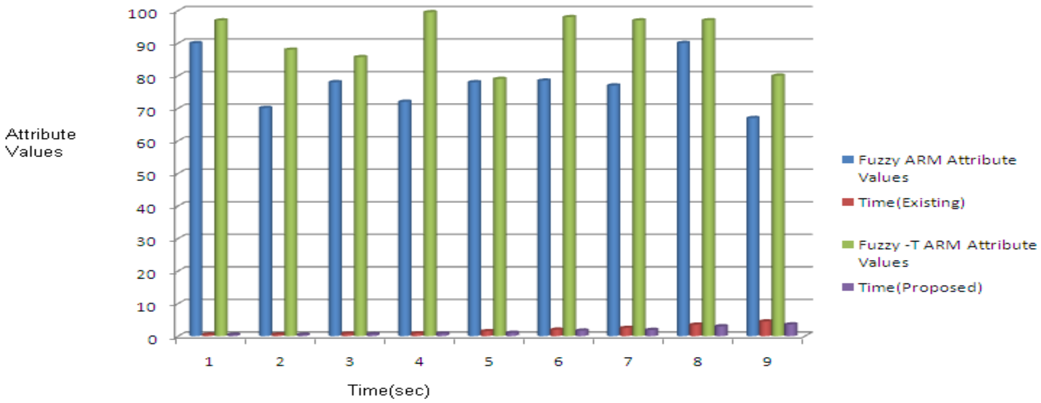

Fig. 3: Performance comparison of fuzzy ARM and fuzzy T ARM

DISCUSSION

In this study, we assess the performance of our algorithm with respect to fuzzy-T ARM is the most popular and widely used online fuzzy mining algorithm. We have used the breast cancer dataset for our experimental analysis. The crisp dataset has around 2.5M transactions and the fuzzy dataset has around 10M transactions. Thus, the dataset size is significantly larger than the available main memory.

It can be clearly observed that Fuzzy-T ARM performs 8-19 times faster than fuzzy ARM, depending on the support used. Please note that the execution times for fuzzy ARM for support values 0.075 and 0.1 have not been calculated as the time exceeded 50K seconds. From that, it is very clear that our algorithm has speeds nearly 8-13 times that of fuzzy-T ARM.

More importantly, for any dataset there is a particular support value for which optimal number of item sets is generated and for supports less than this value, we get a flood of item sets which are of no practical use. From our experiments, we have observed that our algorithm performs most efficiently and speedily at this optimal support value, which occurs in the range of 0.15 - 0.2 for this dataset.

CONCLUSION

In this study, we have presented a novel fuzzy-TARM algorithm, for very huge datasets, as an alternative to fuzzy ARM, which is the most widely used algorithm for fuzzy ARM. Through our experiments, we have shown that our algorithm is faster than fuzzy ARM. This considerable speed up has been achieved because novel properties like two-phased tid list-style processing using partitions, tid lists represented in the form of byte-vectors, effective compression of tid lists and a tauter and quicker second phase of processing.

REFERENCES

Pedreschi, D., S. Ruggieri and F. Turini, 2008. Discrimination-aware data mining. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD'08), ACM New York, NY, USA., pp: 560-568. DOI: 10.1145/1401890.1401959

Alcalá-Fdez, J., R. Alcalá, M.J. Gacto and F. Herrera, 2009. Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms. Fuzzy Sets Syst., 160: 905-921. DOI: 10.1016/j.fss.2008.05.012

Hong, T.P., Y.F. Tung, S.L. Wang, M.T. Wu and Y.L. Wu, 2009. An ACS-based framework for fuzzy data mining. Expert Syst. Appl. 36: 11844-11852. DOI: 10.1016/j.eswa.2009.04.016

Hu, Y.H., Y.L. Chen and K. Tang, 2009. Mining sequential patterns in the B2B environment J. Inform. Sci., 35: 677-694. DOI: 10.1177/0165551509103600

Intan, R. and O. Yenty, 2008. Mining multidimensional fuzzy association rules from a normalized database. Proceedings of the IEEE International Conference on Convergence and Hybrid Information Technology, Aug. 28-30, IEEE Xplore, Daejeon, pp: 425-432. DOI: 10.1109/ICHIT.2008.229

Khan, N.M., C. Jahangeer, D. Goburdhun and M.H. Khan, 2010. Application of a statistical model to biological data analysis: Exclusive breastfeeding. Am. J. Biostat., 1: 42-45. DOI: 10.3844/amjbsp.2010.42.45

Mahafzah, B.A., A.F. Al-Badarneh and M.Z. Zakaria 2009. A new sampling technique for association rule mining. J. Inform. Sci., 35: 358-376. DOI: 10.1177/0165551508100382

Papageorgiou, E.I., 2011. A new methodology for decisions in medical informatics using fuzzy cognitive maps based on fuzzy rule-extraction techniques. Applied Soft Comput., 11: 500-513. DOI: 10.1016/j.asoc.2009.12.010

Poovammal, E. and M. Ponnavaikko, 2009. Utility independent privacy preserving data mining on vertically partitioned data. J. Comput. Sci., 5: 666-673. DOI: 10.3844/jcssp.2009.666.673

Qodmanan, H.R., M. Nasiri and B. Minaei-Bidgoli, 2011. Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. Expert Syst. Appl.: Int. J., 38: 288-298. DOI: 10.1016/j.eswa.2010.06.060

Reddy, C.S. and K. Raju, 2009. Improving the accuracy of effort estimation through fuzzy set representation of size. J. Comput. Sci., 5: 451-455. DOI: 10.3844/jcssp.2009.451.455

Vijayalakshmi, S. and V. Mohan, 2010. Mining sequential access pattern with low support from large pre-processed web logs. J. Comput. Sci., 6: 1293-1300. DOI: 10.3844/jcssp.2010.1293.1300

Weng, C.H. and Y.L. Chen, 2010. Mining fuzzy association rules from uncertain data. Know. Inform. Syst., 23: 129-152. DOI: 10.1007/s10115-009-0223-1

Wu, M.T., T.P. Hong and C.N. Lee. 2010. An improved ant algorithm for fuzzy data mining. Proceedings of the Second international conference on Computational collective intelligence: technologies and applications, (ICCCI'10), Springer-Verlag Berlin, Heidelberg, pp: 344-351.