

# A Rough Association Rule based Approach for Class Prediction with Missing Attribute Values

D.V.Janardhan Rao and Pabitra Mitra

Department of Computer Science and Engineering  
Indian Institute of Technology Kanpur  
Kanpur-208016, India  
email: {dvjrao,pmitra}@iitk.ac.in

**Abstract.** In many real-life classification tasks one often encounters instances where values of certain attributes are missing. Rough association rules, extracted from a training data, may be used to predict the class labels of such instances with missing attribute values. The association rules capture the statistical dependency between the feature values. We propose a methodology based on rough set feature selection, association rule mining, and rule matching for prediction of class labels. Experiments are performed on two benchmark data sets and a real-life medical diagnosis task. The algorithm is found to provide high class prediction accuracy, even with large number of features missing.

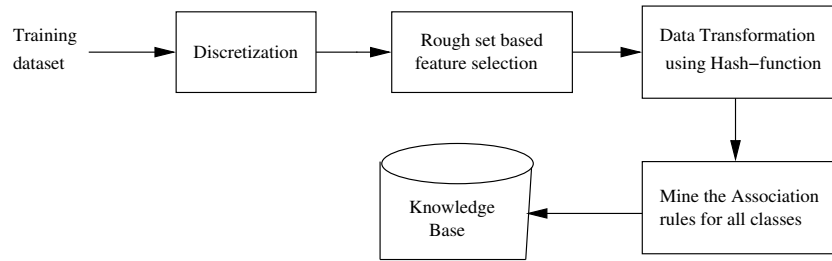
## 1 Introduction

In many real-life applications, the available data may be missing values for some attributes. For example, in a medical diagnostic system where we wish to predict a patient's disease based on various laboratory tests, it may be that some of the lab test results are not available for all the patients. In such cases, it is common to estimate the missing attribute value based on other examples for which this attribute has a known value. There are several approaches of handling missing attribute values in the literature which are surveyed and compared in [7]. There are a few approaches which use rough set theory to deal with missing attribute values [3, 2, 4, 5, 6, 7]. Some of the rough set based approaches use the concept of attribute-value blocks by computing characteristic sets and characteristic relations, which are a generalization of the indiscernibility relations. But most of these approaches deal with missing attribute values in the context of learning or knowledge acquisition or rule induction from training examples with missing values.

In this paper we propose a method for predicting the class of an unseen instance with missing attribute values, without estimating the missing values, based on known feature values using rough association rules.

The proposed approach for predicting the class of a unseen instance with missing feature values based on Rough association rules comprises of the following steps. First we reduce the feature space using rough set based feature selection to select a minimal subset of features called *reduct* which has the same

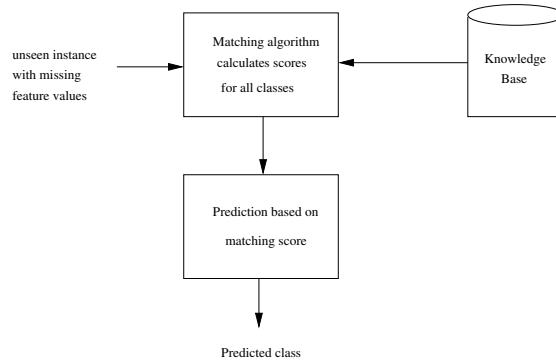
classification power as the original feature set and with irrelevant features removed. As reduct computation can be performed on discrete-valued attributes, continuous-valued attributes must therefore be discretized prior to its use. We then transform each feature value in the reduced decision table using a one-to-one hash function which maps each distinct feature value to a distinct number (say *itemno*) based on the feature number and its discretized value. A reduced (containing only reduct attributes) decision table is constructed using these transformed attributes. Suppose that each tuple in the reduced decision table represents a transaction of items (represented by *itemno*). Apply association rule mining to find the association rules with minimum support  $s\%$  and confidence  $c\%$ . Note that the support and confidence values are related to the minimum confidence probability with which the user wants to predict the class of the unseen instance with missing feature values. We then segregate the association rules for each class and store them in the rulebase. This forms our knowledge base which would be used later to predict the class labels. To predict the class of an unseen instance with missing feature values, we match it with association rules for each class and compute a matching score. The class for which the matching score is highest will be the predicted class. The above steps are illustrated in Figures 1 and 2. The details of the individual steps are discussed in the subsequent sections.



**Fig. 1.** Block diagram for building the knowledge base of rough association rules from the training data.

## 2 Rough Set Based Feature Selection

The information content of the features in terms of class prediction determines the accuracy of the classification model developed using the features. From a theoretical view point having more features should give us more classification power. However, in the real-world this is not generally the case due to several reasons. First, the running time of an induction algorithm often increases dramatically with the number of features making it impractical for problems with a large no of features. Secondly, irrelevant and redundant features also cause problems in this context as they may confuse the learning algorithm. Thus, by



**Fig. 2.** Block diagram for class prediction methodology.

reducing the feature space, we can considerably decrease the running time of the algorithm, and we can increase the accuracy of the resulting model.

Rough set theory provides a convenient framework for feature selection problems. The success of rough sets theory in feature selection is owing to the fact that rough sets can tell whether the data is complete or not based on the data itself. If the data is incomplete, it suggests more information about the objects need to be calculated in order to build a good classification model. On the other hand, if the data is complete, rough sets theory can also determine whether there are redundant attributes in the data and identifies the minimal attribute set required to build a good classification model without sacrificing the accuracy of the classification model.

Rough sets theory classifies all the attributes into three categories: core attributes, reduct attributes and superfluous attributes. The attributes whose removal makes the decision table inconsistent and hence reduces the classification power are called core attributes and should be retained in the dataset. The minimal subset of attributes which has the same classification power as the total condition attributes is called a reduct. A decision table may have more than one possible reduct. Any of them can be used to replace the original table. Here we use the rough set based feature selection algorithm, proposed by Lin et al [11], which is independent of the induction algorithm. The algorithm uses efficient set oriented relational algebraic operations and is scalable to large number of features. It is also tolerant to inconsistencies in the data table. We ignore the instances with missing attribute values from training data while calculating the reduct.

## 2.1 Data Transformation

Transforming the data using the one-to-one Hash-function described below will replace each attribute value pair with a distinct identifier (say *itemno*). Now each instance of the dataset is transformed into a transaction of items. The main advantages of this transformation are,

- It will reduce the size of the knowledge base drastically by reducing the memory required to store each association rule.
- It facilitates efficient matching of association rules for class prediction by reducing the time taken for matching at a later stage.

**One-to-One Hash-function** Let  $n$  be the total number of features or condition attributes. Let  $f_1, f_2, \dots, f_n$  represents the  $n$  features. Let each feature  $f_i$  takes the values  $\{1, 2, \dots, n(f_i)\}$  where,  $n(f_i)$  represents the number of distinct values the feature  $f_i$  can take. If the feature is continuous, it should be discretized first. Let  $M = \text{Max}\{n(f_i)\} + 1, i = 1, 2, \dots, n$ . Let  $F$  be the set representing the feature indices or positions  $1, 2, \dots, n$ , and  $V$  be the set representing the values that any feature can take  $\{1, 2, \dots, M-1\}$ . Also, let  $H : F \times V \rightarrow N$  be the Hash-function defined as

$$H(f, v) = f \cdot M + v \% M,$$

where,  $N$  is the set of natural numbers and  $\%$  stands for usual modulo operator.

**Theorem 1.** *The Hash-function  $H : F \times V \rightarrow N$  defined as,  $H(f, v) = f \cdot M + v \% M$  is one-to-one.*

*Proof.* Let  $f_1, f_2 \in F$  and  $v_1, v_2 \in V$ . Suppose  $(f_1, v_1) \neq (f_2, v_2)$ . In order to prove that  $H$  is one-to-one, we have to prove that  $H(f_1, v_1) \neq H(f_2, v_2)$ .

**case 1:**  $f_1 \neq f_2, v_1 = v_2 = v$  (say)

$$\begin{aligned} & f_1 \neq f_2 \\ & \Rightarrow f_1 \cdot M \neq f_2 \cdot M \\ & \Rightarrow f_1 \cdot M + v \% M \neq f_2 \cdot M + v \% M \\ & \Rightarrow H(f_1, v_1) \neq H(f_2, v_2) \end{aligned}$$

**case 2:**  $f_1 = f_2 = f$  (say),  $v_1 \neq v_2$

$$\begin{aligned} & \Rightarrow v_1 \% M \neq v_2 \% M \text{ (since } v_1, v_2 < M) \\ & \Rightarrow f \cdot M + v_1 \% M \neq f \cdot M + v_2 \% M \\ & \Rightarrow H(f_1, v_1) \neq H(f_2, v_2) \end{aligned}$$

**case 3:**  $f_1 \neq f_2, v_1 \neq v_2$

$$1 \leq |f_1 - f_2| \leq n, \quad 1 \leq |v_1 - v_2| \leq M - 1, \quad f_1 \neq f_2$$

$$\Rightarrow f_1 \cdot M \neq f_2 \cdot M$$

(where  $|x|$  gives the absolute value of  $x$ .)

As maximum value of  $|v_1 \% M - v_2 \% M| = M - 2 (< M)$  and  $|f_1 - f_2| \geq 1$

$$\Rightarrow f_1 \cdot M + v_1 \% M \neq f_2 \cdot M + v_2 \% M$$

$$\Rightarrow H(f_1, v_1) \neq H(f_2, v_2)$$

Hence,  $H$  is a one-to-one function. □

### 3 Association Rule Mining

Association rules, introduced in [10], provide a useful mechanism for discovering correlations among items originally belonging to customer transactions in

a market basket database. Association rules can identify collections of data attributes that are statistically related in the underlying data. Hence, fits well for predicting the class of an unseen instance with missing feature values based on statistical dependences.

In the context of class prediction with missing feature values, this problem amounts to discovering the correlations among the feature values for each class separately and use them for class prediction at a later stage. Each transaction is comprised of a set of feature (condition attribute) values along with the class label. The association rules will be of the form  $X \rightarrow Y$  where  $X$  and  $Y$  are disjoint conjunctions of the attribute-value pairs. The confidence of the rule is the conditional probability of  $Y$  given  $X$ ,  $Pr(Y|X)$  and the support of the rule is the prior probability of  $X$  and  $Y$ ,  $Pr(X)$  and  $Pr(Y)$ . Here the probability is taken to be the observed relative frequency in the dataset.

We are given a set of instances with a class label to indicate the class to which each instance belongs. Note that, class label is called the decision attribute, the rest of the attributes the condition attributes or features. We assume that our dataset is represented in a relational table with the form  $Table(condition - attributes, decision - attributes)$ .  $C$  is used to denote the condition attributes,  $D$  for decision attributes, where  $C \cap D = \phi$ . The association rule generation algorithm is described below.

**Algorithm:** Association rule generation

- Input :
1. Decision table  $T(C, D)$
  2. percentage of minimum confidence  $c$
  3. percentage of minimum support  $s$

*Method:*

Step 1. Discretization of continuous-valued attributes

Step 2. Calculate  $R$ , the reduct attribute set  $R \subseteq C$  using the rough sets model based on database systems [11]. (Reduce the feature space using Rough set based feature selection.)

Step 3. Get the reduced table  $T_{red}$  from the original decision table  $T$   
 $T_{red} = \prod_{R \cup D}(T)$ , where  $\prod$  stands for relational algebraic projection operator.

Step 4. Generate the association rules for each decision class. Let us assume that there are  $k$  distinct classes and their class labels be  $d_1, d_2, \dots, d_k$ .

for (  $i = 1; i \leq k; i++$  ) {

Step 4.1.  $DT_i = \prod_R(\sigma_{D="d_i"}(T_{red}))$  (Separating the tuples with class label  $d_i$ .) where  $\sigma$  stands for relational algebraic selection operator.

Step 4.2. Transform each tuple of the relational table  $DT_i$  using the one-to-one Hash-function  $H$  described earlier into a transaction of items.

Step 4.3. Generate Association rules with minimum support  $s\%$  and confidence  $c\%$  using any standard association rule mining algorithm [10, 9]

}

The percentage of confidence is the minimum confidence probability with which the user wants to predict the class of the unseen instance with missing features. The prediction would be accurate if the support is at least 50%.

### 3.1 Matching an Instance with Association Rules

Matching of the unseen instance with the association rules for each class need not be done using both antecedent and consequent of the rule  $A \Rightarrow B$ . The rule itself says that whenever there  $A$  is there,  $B$  is also present with confidence  $c\%$ . Thus the consequent may be pruned for the purpose of class prediction. We have observed that there is no change in the predicted class when full matching is done on the antecedant part and partial matching is done on both antecedant and consequent of the rule  $A \Rightarrow B$ . So, it is better to do full matching on the antecedant part of the rule, reducing the time taken for matching. We replace the consequent part of the association rule  $A \Rightarrow B$  with the class label of the decision class to which it belongs.

**Lemma 1.** *If  $A_1 \Rightarrow B, A_2 \Rightarrow B$  are two association rules such that  $A_2 \subset A_1$  then  $confidence(A_2 \Rightarrow B) \geq confidence(A_1 \Rightarrow B)$  by Apriori property.*

In the context of class prediction with missing feature values, intuitively the prediction would be accurate if those rules with length of antecedent greater than some threshold say (25%) of the length of the frequent itemset to which it belongs, were only considered for matching. We have introduced an additional factor called specificity (defined in the next section) of the rule in the matching score. This makes the prediction even more accurate.

## 4 Class Prediction

We will use the knowledge-base of rough association rules to predict the class of the unseen instance with missing feature values. The decision to which class an unseen instance belongs is made on the basis of three factors : confidence, specificity, and matching score. They are defined as follows : "confidence" is the confidence of the association rule  $A \Rightarrow d_i$ ; "specificity" is the total number of attribute-value pairs on the left hand side of the rule (length of the antecedent). The matching rules with a larger number of attribute-value pairs are considered more specific. The third factor, "matching score", is defined as the sum of scores of all matching rules from the decision class. The decision class  $d_i$  for which the matching score i.e, the following value

$$\text{Match score} = \sum_{\text{matching rules } R \text{ describing } d_i} confidence(R) \cdot specificity(R), \quad (1)$$

is the largest is a winner and the decision class of the unseen instance with missing features is predicted as  $d_i$ .

**Algorithm:** Class Prediction

Input : 1. Unseen instance with missing feature values ( say X )  
 2. Knowledge base of Rough association rules

Output : Predicted class

*Method:*

Let us assume that there are  $k$  distinct classes and their class labels be  $d_1, d_2, \dots, d_k$

Step 1. Consider only those feature values of X which are part of the reduct attribute set and transform them using the Hash-function  $H$ .

Step 2. Calculate the matching scores for all the decision class.

*for* (  $i = 1; i \leq k; i++$  ) {

$MS_i = \sum_{\text{matching rules } R \text{ describing } d_i} \text{confidence}(R) * \text{Specificity}(R)$

}

Step 3. Find the class for which the matching score is maximum.

$p = \text{arg}_{i=1,2,\dots,k} \text{Max}(MS_i)$

Step 4. Predicted class =  $d_p$

The main advantages of this class prediction algorithm are listed below,

- This method uses association rules which can identify collections of data attributes that are statistically related in the underlying data for class prediction, fits suitably to this problem.
- It predicts the class without estimating the missing attribute values.
- This method is entirely integrated with the relational database and scalable to large datasets due to the following reasons :
  - core and reduct attributes are calculated using the "Rough sets model based on Database systems".
  - There are fast algorithms to generate association rules from a database.
- This method is not restricted to two-class problems.

## 5 Experimental Results

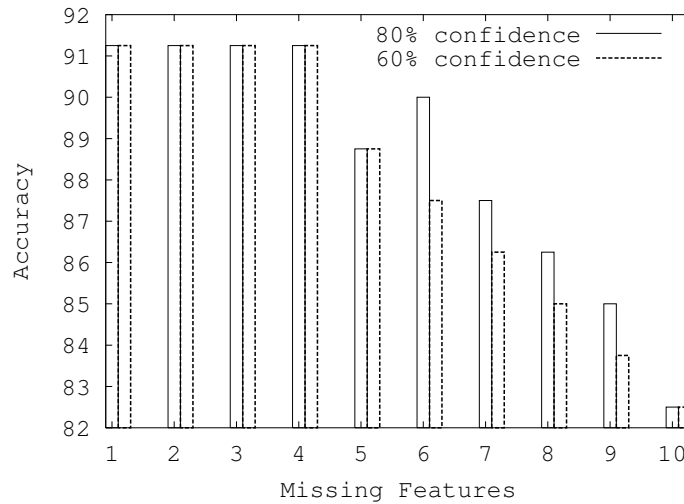
We performed experiments on two UCI machine learning datasets [1] and a real-life Cervical Cancer data [8]. We intentionally dropped some of the feature values from the test set instances. The features to be dropped were selected randomly.  $k$  number of features were dropped for every instance. The accuracy of our class prediction algorithm on this test set with missing values are reported in the Tables 1-2, for different values of  $k$  and the confidence parameter used in association rule mining. For the cervical cancer data some of the feature values were inherently missing, we have considered these only without artificially dropping certain features. The description of the datasets used are as follows,

1. Mushroom dataset : The data consists of 8124 instances with 22 attributes, out of which 2480 have missing attribute values. There are two classes edible and poisonous, each containing 4208 and 3916 respectively. Out of the total

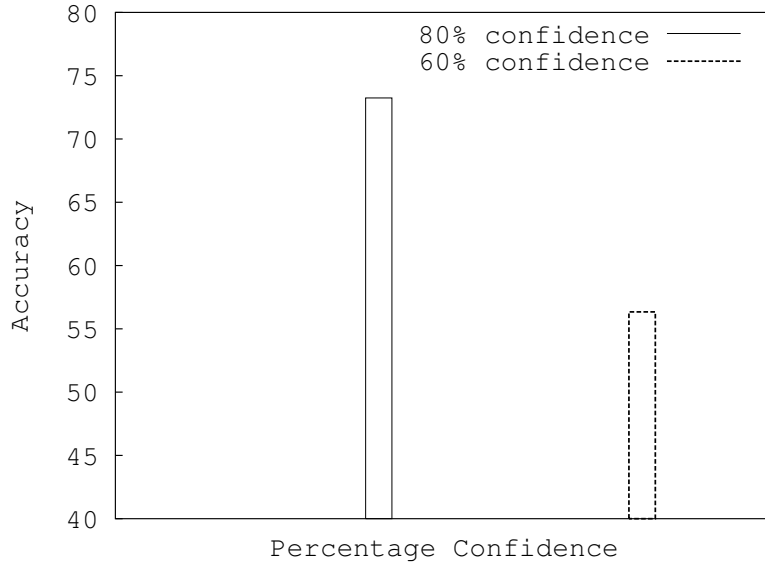
8124 instances, 2481 are used for testing and the remaining are used for training.

2. Dermatology dataset : The data consists of 366 instances with 34 attributes, out of which 8 having missing attribute values. There are 6 classes each containing 112, 61, 72, 49, 52, 20 instances respectively. Out of the total 366 instances, 80 are used for testing and the remaining are used for training.
3. Cervical Cancer data : The data consists of a set of 221 Cervical Cancer patient cases obtained from the database of the Chittaranjan National Cancer Institute (CNCI). There are 4 classes corresponding to the stages 1 - 4 of the cancer, each containing 19, 41, 139 and 19 patients respectively. This data has a lot of missing attribute values. Out of total 221 patient cases, 71 are used for testing and the remaining are used for training.

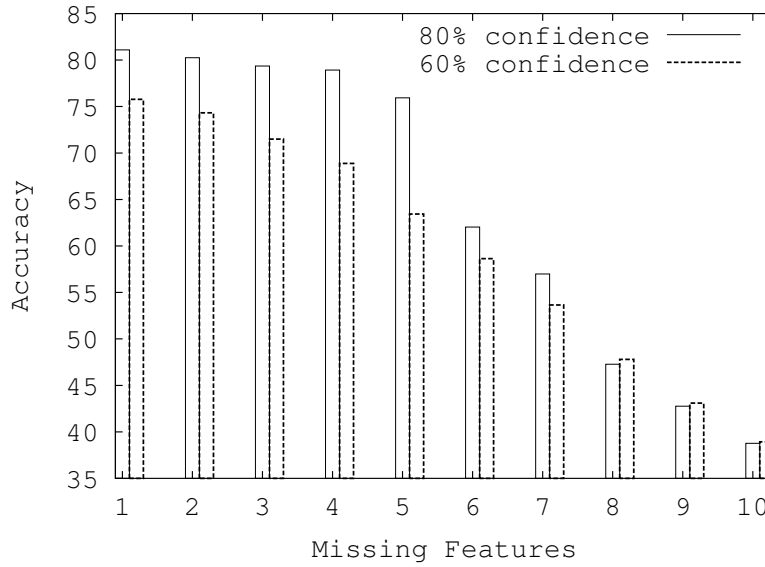
The experimental results ( Figures 3-4 and Tables 1-2) show that the accuracy of class prediction increases with increasing confidence value (the minimum probability with which the user wants to predict the class of the unseen instance with missing feature values), decreases with increase in number of missing feature values which is as expected. However, the accuracy can be very high if the features of the dataset are highly correlated with the class label irrespective of the confidence parameter and the number of missing feature values. This was the case with dermatology dataset. In this dataset the matching score of the predicted class was found to be far greater than the matching scores of the remaining classes, resulting in very accurate prediction.



**Fig. 3.** Dermatology data: Class prediction accuracy of the proposed algorithm.



**Fig. 4.** Cervical Cancer data: Class prediction accuracy of the proposed algorithm.



**Fig. 5.** Mushroom data: Class prediction accuracy of the proposed algorithm.

## 6 Conclusions and Discussion

In this article we propose a method for predicting the class of an unseen instance with missing attribute values, without estimating the missing values, us-

**Table 1.** Dermatology data: Classification accuracy for different numbers of missing features and confidence values

No. of missing feature values	Confidence (%)	Classification accuracy (%)
1	60	91.25
	80	91.25
2	60	91.25
	80	91.25
3	60	91.25
	80	91.25
4	60	91.25
	80	91.25
5	60	88.75
	80	88.75
6	60	87.5
	80	90.0
7	60	86.25
	80	87.5
8	60	85.0
	80	86.25
9	60	83.75
	80	85.0
10	60	82.5
	80	82.5

**Table 2.** Cervical Cancer data: Classification accuracy for different confidence values

Confidence (%)	Classification accuracy (%)
60	56.34
80	73.24

ing known feature values based on rough association rules. The rough association rules capture the statistical dependency relations. Our future work would be to study the effect of correlation between the features of the dataset on the accuracy of the class prediction algorithm. The accuracy of the class prediction may be improved further using the concept of active classifier and other soft-computing methods.

**Table 3.** Mushroom data: Classification accuracy for different numbers of missing features and confidence values

No. of missing feature values	Confidence (%)	Classification accuracy (%)
1	60	75.77
	80	81.09
2	60	74.32
	80	80.25
3	60	71.5
	80	79.36
4	60	68.88
	80	78.92
5	60	63.44
	80	75.94
6	60	58.64
	80	62.03
7	60	53.65
	80	56.99
8	60	47.8
	80	47.28
9	60	43.1
	80	42.76
10	60	38.93
	80	38.77

## References

1. C. L. Blake and C. J. Merz. *UCI Repository of machine learning databases*. University of California, Irvine, Dept. of Information and Computer Sciences, <http://www.ics.uci.edu/~mllearn/MLRepository.html>, 1998.
2. J.W.Grzymala-Busse. On the unknown attribute values in learning from examples . In *Proc. of the 6th International Symposium on Methodologies for Intelligent Systems (ISMIS-91)*, pages 368–377. Springer-Verlag, 1991.
3. J.W.Grzymala-Busse. Rough set strategies to data with missing attribute values . In *Proceedings of the Workshop on Foundations and New Directions in Data Mining (ICDM-2003)*, pages 56–63. Springer-Verlag, 2003.
4. J.W.Grzymala-Busse. Characteristic relations for incomplete data : A generalization of the indiscernibility relation . In *Proceedings of the fourth international conference on Rough sets and Current Trends in Computing (RSTC'2004)*, pages 244–253. Springer-Verlag, 2004.
5. J.W.Grzymala-Busse. Rough set approach to incomplete data . In *Proceedings of the seventh international conference on Artificial Intelligence and Soft computing (ICAISC'2004)*, pages 50–55. Springer-Verlag, 2004.
6. J.W.Grzymala-Busse and A.Y.Wang. Modified algorithms lem1 and lem2 for rule induction from data with missing attribute values. In *Proc. of the fifth International workshop on Rough Sets and Soft computing (RSSC'97)*, pages 69–72,

- 1997.
7. J.W.Grzymala-Busse and M.Hu. A comparison of several approaches to missing attribute values in data mining. In *Proceedings of the Second International conference on Rough Sets and Current Trends in Computing (RSCTC'2000)*, pages 340–347, 2000.
  8. P. Mitra, S. Mitra, and S.K. Pal. Staging of cervical cancer using soft computing. *IEEE Trans. Biomedical Engg.*, 47(7):934–940, 2000.
  9. R.Agrawal and R.Srikant. Fast algorithms for mining association rules. In *Proc. of 20th Int'l conference on Very Large Databases*, 1994.
  10. R.Agrawal, T.Imielinski, and A.Swami. Mining associations between sets of items in massive databases. In *Proc. of the ACM-SIGMOD 1993 , Int'l conference on Management of Data*, pages 207–216, 1993.
  11. X.Hu, T.Y.Lin, and J.Han. A new rough sets model based on database systems. *Fundamenta Informaticae*, XX:1–18, 2004.