Selecting relevant features using Rough Sets and the MDL principle

F. Díaz[1] and J. M. Corchado[2]

[1] Depto. de Informática. Universidad de Vigo
Campus Universitario As Lagoas s/n, 32004 Ourense (SPAIN)
e-mail: fdiaz@uvigo.es
[2] Depto. de Informática y Automática. Universidad de Salamanca
Plaza de la Merced, s/n, 37008 Salamanca (SPAIN)
e-mail: corchado@usal.es

Abstract. The paper proposes a feature subset selection algorithm as a pre-processing step to the induction of a classifier from data sets. The learning algorithm is based on a generalization of the rough set theory, the variable precision rough set (VPRS) model. The feature subset selection is based on the application of the minimum description length (MDL) principle. Finally, the precision of the rough classifiers induced after the proposed selection model on several data sets has been evaluated and compared with other algorithms.

1 Introduction

In supervised classification learning, one is provided with a training set containing labelled instances or examples. Each labelled instance contains a list of feature values (attribute values) and a discrete label value (class label). The induction task is then to build a classifier that will correctly predict the label of novel instances. Common machine learning algorithms, including top-down induction of decision trees, such as CART [3] and C4.5 [11], are known to suffer from irrelevant features. A "good" choice of features may not only help to improve performance accuracy, but also may aid to find smaller models for the data, resulting in better understanding and interpretation of the data.

The selection of relevant features, and the elimination of irrelevant ones, is one of the central problems in machine learning, and many induction algorithms incorporate some approach to address it. There have been many attempts to define what is a "relevant" feature, in the machine learning literature. John *et al.* [7] define two notions of relevance: *strong* and *weak relevance*. Strong relevance implies that the feature is indispensable in the sense that it cannot be removed without a loss of prediction accuracy. Weak relevance implies that the feature can sometimes contribute to prediction accuracy depending on which other attributes are considered. Features are relevant if they are either strongly or weakly relevant, and are irrelevant otherwise.

The rough set theory, proposed by Pawlak [8, 9], is an attempt to dispose a formal framework for the automated transformation of data into knowledge. It

is based on the idea that any inexact concept (for example, a class label) can be approximated from below and from above using an indiscernibility relationship (generated by information about objects). Pawlak points out that one of the most important and fundamental notions to the rough sets philosophy is the need to discover redundancy and dependencies between features [9]. Since then, this philosophy has been used successfully in several tasks as, for example, construction of rule based classification schemes, identification and evaluation of data dependencies, information-preserving data reduction, etc. [10].

In this paper, a generalization of Pawlak's model, that is known as variable precision rough set model (VPRS) [14], is used to construct a rule based classifier from data sets. Since the induction process is a large time-consuming task when many features are considered, the paper proposes a feature subset selection method as a pre-processing step to improve the induction process.

The paper is organized as follows. Section 2 briefly introduces the relevant rough set terminology. Section 3 shows how the rough set theory may be used in classification tasks and introduces the concept of β–rough classifier. Section 4 shows the relationship among rough set theory and the feature subset selection and describes an algorithm for selecting a "good" set of attributes. And finally, results and conclusions are presented in section 5.

2 Rough Set Theory

This section describes the basic concepts in rough set theory, viewed from the perspective of the supervised classification learning. An *information system* is a pair $S = \langle U, A \rangle$, where U is a non-empty, finite set called the *universe*, and A is a non-empty, finite set of attributes (or features). Every attribute $a \in A$ is a total function $a : U \longrightarrow V_a$, where V_a is the set of values of the attribute a, called the domain of a. An equivalence relationship, referred to as *indiscernibility relation*, can be associated with every subset of attributes $P \subseteq A$. This relation is defined as:

$$IND(P) = \{(x, y) \in U \times U : \text{ for every } a \in P,\ a(x) = a(y)\} \tag{1}$$

The elements of U that satisfy the relation $IND(P)$ are objects with the same values for the attributes P and they are indiscernible with respect to P. Therefore, the indiscernibility relationship induces a partition of the universe into disjoint classes (each one of them given by the equivalence classes in the quotient set $U/IND(P)$).

In the rough set theory [8], Pawlak proposes that given any subset of features P, any concept $X \subseteq U$ can be defined approximately by the employment of two sets, called lower and upper approximations. The *lower approximation*, denoted by $\underline{P}X$, is the set of objects in U which can be certainty classified as elements in the concept X using the set of attributes P, and is defined as follows:

$$\underline{P}X = \bigcup \{Y \in U/IND(P) : Y \subseteq X\} \tag{2}$$

The *upper approximation*, denoted by $\overline{P}X$, is the set of elements in U that can be possibly classified as elements in X, formally:

$$\overline{P}X = \bigcup \{Y \in U/IND(P) : Y \cap X \neq \emptyset\} \tag{3}$$

The *boundary region* of the concept X in relation to P is defined as follows:

$$BND_P(X) = \overline{P}X - \underline{P}X \tag{4}$$

If $BND_P(X) = \emptyset$ then X is *definable* using P, otherwise X is a *rough set* with respect of P.

The *degree of dependency* of a set of features P on a set o features R is denoted by $\gamma_R(P)$, with $0 \leq \gamma_R(P) \leq 1$, and is defined as:

$$\gamma_R(P) = \frac{|POS_R(P)|}{|U|} \tag{5}$$

where $POS_R(P)$ is the set defined by:

$$POS_R(P) = \bigcup_{X \in U/IND(P)} \underline{R}X \tag{6}$$

namely, $POS_R(P)$ contains the objects of U which can be classified as belonging to one of the equivalence classes of $IND(P)$, using only features from the set R. If $\gamma_R(P) = 1$, then R functionally determines P.

P is an independent set of features if there does not exist a strict subset P' of P such that $IND(P) = IND(P')$. A set $R \subseteq P$ is a *reduct* of P if it is independent and $IND(R) = IND(P)$. Each reduct has the property that a feature can not be removed from it without changing the indiscernibility relation. Many reducts for a given set of features P may exists. An attribute $a \in P$ is indispensable if $IND(P) \neq IND(P \setminus \{a\})$. The core of P is the union of all the indispensable features in P.

The indispensable attributes, reducts, and core can be similarly defined relative to a decision attribute or output feature. The precise definitions of these concepts can be fount in Pawlak's book on rough sets [9].

A direct application of the rough set theory, that is shown in the next section, is the construction of rule based classifiers from data sets based on the approximation of the target concepts (defined by the class labels) in terms of the knowledge given by the feature values.

3 A β–rough Classifier

A *classifier* maps an unlabelled instance to a class label using some internally stored structures. Given a test set, an estimation of the accuracy of the classifier can be defined as the ratio of the number of correctly classified instances to the number of instances. An *inducer* generates a classifier from a training set. The (estimated) accuracy of an inducer, given a training set and a test set, is the accuracy of the classifier induced from the training set when run on the test set.

A *decision table* is an information system of the form $S = \langle U, A \cup \{d\} \rangle$, where $d \notin A$ is a distinguished attribute called the *decision attribute* or *class attribute*. The elements of the set A are referred to as condition attributes. A decision table is a classifier that has as its internal structure a table of labelled instances. Given a novel instance, the classification process is based on the search of all matching instances in the table. If no matching instances are found, unknown is returned; otherwise, the majority class of the matching instances is returned (there may be multiple matching instances with conflicting labels). In addition, it is important to dispose of these classification rules with the minimal effort, and therefore, the simplification of decision tables is of primary importance.

In the rough set framework, the simplification process of a decision table comprises two fundamental tasks. On the one hand, reduction of attributes consists of removing redundant or irrelevant attributes, without losing any essential classification information. The computation of the reducts for the condition attributes relative to the decision attribute is carried out to achieve this goal. On the other hand, the reduction of attribute values is related to the elimination of the greatest number of condition attribute values, maintaining also the classificatory power.

A rough inducer simply passes the training set to a reduced decision table, herein referred to as *rough classifier*, after the reduction process above mentioned. Each row of the reduced decision table represents a classification rule of the rough classifier.

Although, the original rough set theory provides an adequate framework for data analysis in general, and for classification tasks in particular, some limitations have been detected. One of the major limitations is the inability to extract knowledge from data with a controlled degree of uncertainty. In fact, all rules of the rough classifier must be deterministic, so that inconsistent instances must be discarded *a priori*.

One extension of the Pawlak's theory that is aimed at handling uncertain information is the variable precision rough set model (VPRS)[14]. The VPRS model is a generalization of the rough set model that introduces a controlled degree of uncertainty within its formalism. This fact leads to more general notions of the upper and lower approximations. The fundamental notion introduced by the VPRS model is the *majority inclusion relationship*. To define this concept it is necessary to introduce first the notion of *misclassification error* $c(X, Y)$. This measure is defined as the ratio of objects in X that also belongs to Y, and it evaluates the misclassification error that is committed when the concept given by the set X is considered as a part of a target concept (given by the Y). Formally, it is defined as follows:

$$c(X, Y) = \begin{cases} 1 - \frac{|X \cap Y|}{|X|} & \text{if } |X| > 0 \\ 0 & \text{if } |X| = 0 \end{cases} \tag{7}$$

Based on this measure, it can be defined the standard set inclusion relationship between X and Y as $X \subseteq Y$ if and only if $c(X, Y) = 0$. The natural relaxation of this standard definition to allow $c(X, Y)$ greater than 0, causes the

extended definition of inclusion relationship. In addition, Ziarko [14] establishes as requirement that the number of elements of X in common with Y should be above 50%. This restriction on the admissible level of classification error is specified by the parameter β and, it must be within the range $0 \leq \beta < 0.5$. According to this requirement, the majority inclusion relationship is defined as $X \subseteq_\beta Y$ if and only if $c(X,Y) \leq \beta$. It follows directly from the above definition that the majority inclusion relationship becomes the standard inclusion relationship if $\beta = 0$.

The extended inclusion relationship defined has an effect on the original notions of lower and upper approximations, obtaining the following generalized notions. Given an information system $S = \langle U, A \rangle$, any subset of features $P \subseteq A$ and any concept $X \subseteq U$, the *β–lower approximation* of X is defined as:

$$\underline{P}_\beta X = \bigcup \{Y \in U/IND(P) : c(Y,X) \leq \beta\} \tag{8}$$

Similarly, the *β–upper approximation* of the concept $X \subseteq U$ is defined as:

$$\overline{P}_\beta X = \bigcup \{Y \in U/IND(P) : c(Y,X) < 1 - \beta\} \tag{9}$$

The remaining notions of the rough set theory can be defined immediately once the new definitions of the approximations of a concept are given. An induction algorithm considers all instances of the training set (including the inconsistent ones) and carries out the simplification process according to the VPRS model. The simplified decision table will be referred to as *β–rough classifier*.

The next section describes a method of feature selection proposed to improve the behaviour of the induction algorithms, and in particular the inducer of β–rough classifiers which have been introduced in this section.

4 Selection of Relevant Features

For the rough set theory the core of an information system is the set of indispensable features. Removal of any attribute from the core set changes the positive region with respect to the label [9]. This fact can be interpreted as a similarity between the core set and the notion of strong relevance introduced by John *et al.* [7]. On the other hand, the attributes of the core may be insufficient for defining all decision classes. Therefore, other attributes may be added to the core in order to maintain the same classification power that the one achieved with all the features. The minimal set of attributes that is sufficient for satisfying this property is called a reduct. Attributes in the union of all reducts but not belonging to the core set can be interpreted as features with weak relevance. In this sense, aiming at reducing irrelevant features, several methods have been developed [10].

This paper proposes a feature subset selection process that should be carried out before the induction of β–rough classifiers. The aim of this proposal is to reduce de original set of attributes and, therefore, decrease the computational effort for the calculation of optimal reducts. This pre-processing step is fitted in the *filter approach* to the feature subset selection. Filter methods select features

(based on properties of the data itself and independent of the induction algorithm) which are afterwards used by the induction mechanisms. Another class of feature subset selection methods is referred to as *wrapper approach* since these algorithms treat feature selection as a wrapper around the induction process. Namely, they conduct a search for a good feature subset using the induction algorithm itself as part of the evaluation function.

As it has been mentioned above, filter methods are based on properties of the data itself to select features. In the rough set framework, the natural way to measure the prediction success (i.e., the goodness of a set of condition attributes to predict a decision attribute) is the degree of dependency (see (5)). However, Düntsch and Gediga [4] have shown the weakness of this measure in order to assess an estimation of the predictive accuracy of a set of condition attributes Q with regard to a class attribute d. To overcome this deficiencies, Düntsch and Gediga define the notion of *rough entropy* [5]. Based on this measure it is defined in this work a significant coefficient, which will be used by the proposed algorithm to select relevant features. The underlying principle is the *minimum description length principle* (*MDLP*) [13] since the definition of the rough set entropy comprises two different factors: the complexity of the hypothesis given by the set of condition attributes Q, and the accuracy of a given hypothesis Q to determine the value of the decision attribute d.

The associated complexity of a given set of condition attributes Q can be evaluated through the entropy of the partition $U/IND(Q)$, which will be denoted by $H(Q)$. On the other hand, the conditional rough entropy $H(d|Q)$ can be used to evaluate the accuracy that is achieved when the condition attributes Q are used to predict the value of the condition attribute d. The formal definition of the rough entropy, denoted by $RH(Q,d)$, is given by the following expression:

$$\begin{gathered} RH(Q,d) = H(Q) + H(d|Q) = \\ = H(Q) + \left\{ \{1-\gamma_Q(d)\} \log_2 |U| + \sum_{X_i \subseteq U \setminus POS_Q(d)} \frac{|X_i|}{|U|} \log_2 \frac{|X_i|}{|U|} \right\} = \\ = \{1-\gamma_Q(d)\} \log_2 |U| - \sum_{X_i \subseteq POS_Q(d)} \frac{|X_i|}{|U|} \log_2 \frac{|X_i|}{|U|} \end{gathered} \tag{10}$$

where X_i represents each one of the classes of the partition $U/IND(Q)$, the set $POS_Q(d)$ is the positive region of Q with regard to the decision attribute d, and $\gamma_Q(d)$ is the degree of dependence of the attribute d on the set of attributes Q.

In order to evaluate the significance of a condition attribute, $a \in Q$, with regard to the decision attribute d, it is evaluated the variation that the rough entropy suffers when the considered attribute is discarded from Q. Namely, it is computed the term $\Delta_a RH(Q,d)$, given by the difference between $RH(Q,d)$ and $RH(Q \setminus \{a\}, d)$. Formally,

$$\begin{gathered} \Delta_a RH(Q,d) = RH(Q,d) - RH(Q \setminus \{a\}, d) = \\ \{H(Q) - H(Q \setminus \{a\})\} - \{H(d|Q \setminus \{a\}) - H(d|Q)\} = \\ = \Delta_a H(Q) - \Delta_a H(d|Q) \end{gathered} \tag{11}$$

where $\Delta_a H(Q)$ and $\Delta_a H(d|Q)$ are defined so that both terms are positive.

The significance of an attribute $a \in Q$ is defined in a way that its value is greater when the removal of the attribute leads to a greater diminution of the complexity of the hypothesis $Q \setminus \{a\}$, and simultaneously, to a smaller loss of accuracy of the hypothesis. Before the formal definition of the significant rough coefficient, the terms $H(Q)$ and $H(d|Q)$ can be normalized between the values 0 and 1, in the following way:

$$S(Q) = 1 - \frac{H(B)}{\log_2(|U|)} \tag{12}$$

since $0 \leq H(Q) \leq \log_2(|U|)$, and

$$S(d|Q) = 1 - \frac{H(d|B)}{\{1 - \gamma_Q(d)\} \log_2 (|U| - |POS_Q(d)|)} \tag{13}$$

given that $H(d|Q) \leq \{1 - \gamma_Q(d)\} \log_2 (|U| - |POS_Q(d)|)$, as Düntsch and Gediga have shown [5].

Once the entropies have been normalized, the significant rough coefficient of the attribute a within the set of condition attributes Q with respect to the decision attribute d, denoted by $\sigma_a(Q, d)$, is defined as follows:

$$\sigma_a(Q, d) = \frac{\{-\Delta_a S(Q)\} + \{1 - \Delta_a S(d|Q)\}}{2} \tag{14}$$

where $\Delta_a S(Q)$ and $\Delta_a S(d|Q)$ are defined in a similar way that the terms $\Delta_a H(Q)$ and $\Delta_a H(d|Q)$, respectively, in expression (11). The definition of the significant coefficient $\sigma_a(Q, d)$ is done according to what has been stated above, concerning both the decrease of complexity and accuracy of the resulting model.

Moreover, it is possible to define a measure of the significance of an attribute that includes the β parameter of the VPRS model. This new coefficient, referred to as *significant β–rough coefficient* and denoted by $\sigma_{a,\beta}(Q, d)$, is defined by:

$$\sigma_{a,\beta}(Q, d) = \frac{\{-\Delta_a S(Q)\} + \{1 - \Delta_a S_\beta(d|Q)\}}{2} \tag{15}$$

where $\Delta_a S_\beta(d|Q)$ is the variation of the conditional β–rough entropy (defined in a similar way that the conditional entropy, see (10)), when the condition attribute a is removed from the attribute set Q.

Once the metric that is used to evaluate the significance of an attribute (the significant β–rough coefficient)is defined, the proposed algorithm for selecting relevant features is described according to Blum and Langley [2]. These authors state that a convenient paradigm for viewing feature selection methods is that of heuristic search, with each state in the search space specifying a subset of the possible features. Following Blum and Langley viewpoint the four basic issues that characterize this method are:

1. The starting point in the space, which in turn influences the direction of search and the operators used to generate successor states. The proposed algorithm starts with all attributes and successively removes them. This approach is known as backward elimination.

Table 1. Reduction in the number of attributes after feature selection.

	Cases	*Attributes*		*Classes*	*Feature Selection*	
		Cont	*Discr*		β	*Attributes*
anneal	898	9	29	6	0.025	14
breast-c	699	9	–	2	0.025	6
colic	368	10	12	2	0.04	8
credit-a	690	6	9	2	0.0125	11
heart-c	303	8	5	2	0.05	8
hypo	3772	7	22	5	0.0025	17
vehicle	846	18	–	4	0.35	13

2. The organization of the search. Any realistic approach relies on a greedy method to traverse the space considering that an exhaustive search is impractical. At each point in the search, the proposed algorithm considers all local changes, namely, it evaluates the significance of each attribute of the current set of attributes.
3. The strategy used to evaluate alternative subsets of attributes. In this paper, the variation of the normalized β–rough entropy has been chosen for this purpose. Specifically, at each decision point the next state that is selected is that one which results of remove the attribute with the least significant β–rough coefficient.
4. A criterion for halting the search. In the algorithm, the search terminates when the difference between the degree of dependency at initial state and the current state (both with respect to the decision) do not go beyond a predefined threshold.

The next section describes the experiments that have been carried out in order to test the performance of a inducer of β–rough classifiers, which in addition perform the pre-processing step of the feature subset selection, described in this section.

5 Results and Conclusions

This section describes the experiments carried out using the previously introduced method to evaluate the accuracy of the β–rough classifiers. The classifiers induced using the proposed method are referred to as β–*RS+FS* classifiers. The data sets used in the experiments are available at the repository of the University of California in Irvine [1]. Besides, and given that both the inducer and the feature-selection algorithms deal with discrete variables and without missing values, it has been necessary to discretize and complete the original data sets. On the one hand, the entropy based method of Fayyad and Irani [6] have been used to discretize continuous variables. On the other, and once the original data have been discretized, the missing values have been replaced by the value of the mode

Table 2. Comparison of induction algorithm of β–rough classifiers with feature selection and C4.5 algorithm.

	β–RS+FS		*C4.5 v8*		*C4.5 v8+D*	
	β	*CV10 Error*	*CV10 error*	*t-test*	*CV10 error*	*t-test*
anneal	0.025	4.59 ±0.12	7.67 ±0.12	>	9.48 ±0.14	>
breast-c	0.025	4.00 ±0.27	5.26 ±0.19	>	9.48 ±0.14	>
colic	0.04	11.43 ±1.27	15.00 ±0.20	>	15.10 ±0.10	>
credit-a	0.0125	17.23 ±1.51	14.70 ±0.20	=	14.00 ±0.10	<
heart-c	0.05	17.51 ±1.59	23.00 ±0.50	>	21.70 ±0.60	>
hypo	0.0025	0.54 ±0.07	0.48 ±0.01	=	0.72 ±0.03	>
vehicle	0.35	38.43 ±1.36	27.10 ±0.40	<	31.50 ±0.50	<

within the same decision class. For each data set, the inducer takes the output of the algorithm of feature subset selection, and learns a β–rough classifier from the training set. The accuracy of the induced classified is then estimated using ten-fold stratified cross validation (CV10).

Table 1 shows the reduction in the number of attributes that the algorithm of feature selection achieves. The available cases for each data set, the number of decision classes and condition attributes, together with the value of the β parameter used in the feature-selection process and the number of selected attributes, are shown in this table. It should be emphasized that the algorithm of feature selection gets an average reduction of the 42.1% (with a standard error of 5.4%) in the number of attributes. This considerable reduction yet implies a greater reduction in the dimension of feature space (taking into account that the dimension of feature space grows exponentially with the number of attributes), and therefore, the proposed algorithm is beneficial as for the computational effort that is necessary to induce β–rough classifiers.

Now then, an assessment of the selected features is also required in order to validate the algorithm. In this sense, the CV10 accuracy of the β–rough classifiers inducer after the feature selection process has been compared to the CV10 accuracy of the C4.5 algorithm [11]. Table 2 shows the CV10 errors (together with the standard error) of the *β-RS+FS* classifier for each data set. The column headed '*C4.5 v8*' shows the CV10 error of the version 8 of C4.5 algorithm, while the column headed '*C4.5 v8+D*' shows the results of the same version but, once the continuous variables of the data sets have been discretized according to the algorithm of Fayyad and Irani. The values of the accuracy for the C4.5 algorithm have been obtained from [12]. The table also shows the result of the t-test (with a level of significance of the 5%) which is used to compare the CV10 error of the β–RS+FS classifier and both versions of C4.5 algorithm. As it shown in Table 2 the error of the β–RS+FS is lesser than the C4.5 error in most cases. It is then concluded that the proposed feature-selection algorithm leads to a "good" selection of condition attributes.

Summarizing, this paper proposes and tests a feature subset selection as a

pre-processing step to the induction of a classifier from data sets. The induced classifiers are decision tables which are simplified according to a generalization of the rough set theory, the VPRS model. The use of the VPRS model allows that rules are not necessarily deterministic, in the sense, that a classification error (over the training set) not greater than β ($0 \leq \beta < 0.5$) is admissible. The feature subset selection is based on the application of the MDL principle as selecting criterion. The algorithm tries to establish an appropriate balance between the complexity of the resulting subset of features and its accuracy when it is used to predict the class label of a novel instance.

References

1. Blake, C.L. and Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.
2. Blum, A.L., and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, **97**:245–271.
3. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and regression trees*. Wadsworth International Group.
4. Düntsch, I., and Gediga, G. (1997). Statistical evaluation of rough set dependency analysis. *International Journal of Human-Computer Studies*, **46**:589–604.
5. Düntsch, I., and Gediga, G. (1998). Uncertainty measures of rough set prediction. *Artificial Intelligence*, **106**:77–107.
6. Fayyad, U. M., and Irani, K. B. (1992). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence* (pp. 1024–1027). San Francisco: Morgan Kaufmann.
7. John, G. H., Kohavi, R., and Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 121–129). New Brunswick, NJ: Morgan Kaufmann.
8. Pawlak, Z. (1982). Rough Sets. *International Journal of Computer and Information Sciences*, **11**:341–356.
9. Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht.
10. Polkowski, L., and Skowron, A. (Eds.) (1998). *Rough Sets in Knowledge Discovery*. Springer Verlag, Parts 1, 2.
11. Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.
12. Quinlan, J. R. (1996). Improved Use of Continuous Attributes in C4.5, *Journal of Artificial Intelligence Research* **4**:77–90.
13. Rissanen, J. (1985). Minimum description length principle. In Kotz, S., and Johnson, N. L. (Eds.). *Encyclopedia of Statistical Sciences* (pp. 523–527). John Wiley and Sons, New York.
14. Ziarko, W. (1993). Variable Precision Rough Set Model. *Journal of Computer and System Sciences*, **46**:39–59.

This article was processed using the LaTeX macro package with LLNCS style