nondeterministic decision rules,
classification, decision tables,
rough sets, rule-based classifiers

Barbara MARSZAŁ-PASZEK[1], Piotr PASZEK[1]

NONDETERMINISTIC DECISION RULES IN CLASSIFICATION PROCESS FOR MEDICAL DATA

In the paper, we discuss nondeterministic rules in decision tables, called the second type nondeterministic rules. They have a few decisions values on the right hand side but on the left hand side only one attribute that has two values. We show that these kinds of rules can be used for improving the quality of classification. It is important in rule-based diagnosis support systems, where classification error can lead to serious consequences. The well known greedy strategy to construct the new nondeterministic rules, have been proposed. Additionally, based on deterministic and nondeterministic (second type) rules, classification algorithm with polynomial computational complexity has been developed. This rule-based classifier was tested on the group of decision tables, containing medical data, from the UCI Machine Learning Repository. The reported results of experiments showing that by combining rule-based classifier based on deterministic rules with second type nondeterministic rules give us possibility to improve the classification quality.

1. INTRODUCTION

Over the years many methods based on rule induction and rule-based classification systems were developed [10,20]. A significant part of these systems concerned diagnosis support systems or medical expert systems [4,5,8,11,12,19,21,22]. Some of them are based on rough sets [3,9,17,18,21,22]. In this paper we show that exist possibility for improving the rule-based classification systems. It is particularly important in diagnosis support systems or medical expert systems where wrong diagnosis can lead to serious consequences [5,19,21,22].

We propose a method for rules inducing based on searching for strong rules for a union of a few relevant decision classes and for union of two values of one condition attribute - second type nondeterministic decision rules. There is an additional motivation for the use of such rules in medical expert systems. The nondeterministic decision rules have larger support than the deterministic ones [13]. Therefore they are more often accepted by medical experts in rule-based diagnosis support systems.

In the paper, an application of second type nondeterministic rules in construction of rule-based classifier is presented. These rules are of the following form:

$$a_1(x) = b_1 \wedge \ldots \wedge a_i(x) = (b_{i_1} \vee b_{i_2}) \wedge \ldots \wedge a_t(x) = b_t \Rightarrow d(x) = (c_1 \vee \ldots \vee c_s),$$

where $a_1,\ldots,a_t$, are conditional attributes of the decision table T with the values from the set $V_A(T)$. The decision attribute of T is d and $\emptyset \neq \{c_1,\ldots,c_s\} \subseteq V_d(T)$, where $V_d(T)$ is the value set of the decision d [15].

We include the results of experiments showing that by combining rule-based classifiers based on minimal decision rules [15] with the classifier based on second type nondeterministic decision rules, it is possible to improve the classification quality and reduce classification error.

The paper consists of six sections. In Section 2, we recall the notion of decision table. In Section 3 we describe notions of nondeterministic decision rules. Section 4 contains definitions of classification algorithm. Results of experiments are discussed in Section 5. Section 6 contains short conclusions.

[1] Institute of Computer Science, University of Silesia, Będzińska 39, 41-200 Sosnowiec, Poland, email:{bpaszek, paszek}@us.edu.pl.

2. DECISION TABLES

Let $T = (U, A, d)$ be a *decision table* [14], where $U = \{u_1, \ldots, u_n\}$ is a finite nonempty set of *objects*, $A = \{a_1, \ldots, a_m\}$ is a finite nonempty set of *conditional attributes* (functions defined on *U*), and *d* is the *decision attribute* (function defined on *U*).

We assume that for each $u_i \in U$ and each $a_j \in A$ the value $a_j(u_i)$ belongs to $V_A(T)$ and the value $d(u_i)$ belongs to $V_d(T)$, where $V_d(T) = \{0,1,2,\ldots\}$ is the set of nonnegative integers. By $V_d(T)$ is denoted the set of values of the decision attribute *d* on objects from *U*.

3. SECOND TYPE NONDETERMINISTIC DECISION RULES

Second type nondeterministic decision rules are examples of nondeterministic decision rules. In general, nondeterministic decision rules in a given decision table *T* are of the form

$$a_{j_1}(x) = b_1 \wedge \ldots \wedge a_{j_t}(x) = b_t \Rightarrow d(x) = (c_1 \vee \ldots \vee c_s), \quad (1)$$

where $a_{j_1}, \ldots, a_{j_t} \in A$, $b_1, \ldots, b_t \in V_A(T)$, numbers $j_1, \ldots, j_t$ are pairwise different, and $\varnothing \neq \{c_1, \ldots, c_s\} \subseteq V_d(T)$. Some notations about rules of the form (1) are introduced in [7].

The second type nondeterministic decision rules are defined as follows:

$$a_{j_1}(x) = b_1 \wedge \ldots \wedge a_{j_i}(x) = (b_{i_1} \vee b_{i_2}) \wedge \ldots \wedge a_{j_t}(x) = b_t \Rightarrow d(x) = (c_1 \vee \ldots \vee c_s), \quad (2)$$

where $a_{j_1}, \ldots, a_{j_t} \in A$, $b_1, \ldots, b_t \in V_A(T)$, numbers $j_1, \ldots, j_t$ are pairwise different, and $\varnothing \neq \{c_1, \ldots, c_s\} \subseteq V_d(T)$.

The rule of the form (2) has nondeterminism on one condition attribute beside nondeterminism on decision part of rule. This attribute has two values but it is different from others attributes which have exactly one value. This type of nondeterministic rules appears as a result of shorting rules according to principle MDL (*Minimum Description Length*) [16].

Let us introduce some notation. If r is the nondeterministic rule (1) then by *lh*(*r*) we denote its left hand side, i.e., the formula $a_{j_1}(x) = b_1 \wedge \ldots \wedge a_{j_t}(x) = b_t$, and by *rh*(*r*) its right hand side, i.e., the formula $d(x) \in (c_1 \vee \ldots \vee c_s)$. If β is a boolean combination of descriptors, i.e., formulas of the form $a(x) = v$, where $a \in A \cup \{d\}$ and $v \in V_A(T)$ then by $\|\beta\|_T$ we denote all objects from *U* satisfying β [15].

The second type nondeterministic rule *r* can be distribute on two nondeterministic rules (r_1, r_2) of the form (1) such as:

$$r_1 : a_{j_1}(x) = b_1 \wedge \ldots \wedge a_{j_i}(x) = b_{i_1} \wedge \ldots \wedge a_{j_t}(x) = b_t \Rightarrow d(x) = (c_1 \vee \ldots \vee c_s),$$
$$r_2 : a_{j_1}(x) = b_1 \wedge \ldots \wedge a_{j_i}(x) = b_{i_2} \wedge \ldots \wedge a_{j_t}(x) = b_t \Rightarrow d(x) = (c_1 \vee \ldots \vee c_s).$$

To measure the quality of such rule we use coefficients called the *support* and the *confidence* [1]. They are defined as follows.

The support of this rule in the decision table *T* is defined by

$$supp_T(r) = supp_T(r_1) + supp_T(r_2)$$

where for $i = 1,2$

$$supp_T(r_i) = \frac{\left| \|lh(r_i)\|_T \cap \|rh(r_i)\|_T \right|}{|U|}.$$

We also use a normalized support of r in T defined by

$$norm\_supp_T(r) = \frac{supp_T(r)}{\sqrt{|V(r)|}},$$

where $V(r) \subseteq V_d(T)$ is a decision values set from right hand side of the rule ($rh(r)$).

The confidence of r in T is defined by

$$conf_T(r) = conf_T(r_1) + conf_T(r_2)$$

where for $i = 1,2$

$$conf_T(r) = \frac{|\|lh(r)\|_T \cap \|rh(r)\|_T|}{|\|lh(r)\|_T|}.$$

We can define a set of second type nondeterministic decision rules as the set of all second type nondeterministic rules r (over attributes in T) such that

$$1 \geq conf_T(r_1) + conf_T(r_2) = conf_T(r) \geq \alpha,$$

where parameter $\alpha \in [0.5,1]$.

The algorithm presented in next section is searching for second type nondeterministic rules. Such rules are combined with minimal rules [15] for increasing the classification quality.

4. CLASSIFICATION ALGORITHM BASED ON SECOND TYPE NONDETERMINISTIC RULES

In this section, we present an application of second type nondeterministic rules for classification process. An algorithm was developed for generation of such rules from decision table. It used greedy algorithm *ND*, presented in [13], for calculating nondeterministic decision rules of the form (1). The main steps of this *ND*2 algorithm are as follows.

Algorithm for second type nondeterministic decision rule construction *ND*2

Input: Decision table T, real number $\alpha \in [0.5,1]$;

Output: $RULE_{ND2}(T,\alpha)$ a set of second type nondeterministic decision rules for T.

Step 1. $RULE_{ND2}(T,\alpha)$ is empty set and parameter $\alpha \in [0.5,1]$.

Step 2. For all condition attributes of T do the following:

- Find often appearing two values for this attribute,
- Generate subtable with restriction to these attribute values,
- Delete an attribute which was chosen,
- Generate the set of rules of type (1), // algorithm *ND*
- Ad to these rules attribute which was deleted; // now rules have the form (2)
- Add these generated rules to the set $RULE_{ND2}(T,\alpha)$.

Step 3. Return $RULE_{ND2}(T,\alpha)$.

In our experiments, we used classification algorithms constructed by the combination of two classification algorithms. The first one C_1 is the classification algorithm based on minimal rules generated by using method from *RSESlib* (Rough Set Exploration System library) [3]. This algorithm uses the standard voting procedure.

The second classification algorithm C_2 is based on second type nondeterministic rules generated by the *ND*2 algorithm. The voting procedure on these rules looks as follows. First, all second type nondeterministic rules matching this object are extracted. Next, from these matched rules, a rule with the largest normalized support is selected. In the case when several rules have the same normalized support, the decision value set $V(r)$ of the second type nondeterministic rule r with the smallest decision value set ($|V(r)|$) is selected. If still several second type nondeterministic rules with the above property exist then first of them is selected.

The prediction process for any new object looks as follows. For this object we obtain a decision value c (given by the C_1 classification algorithm) and a decision value set $V(r)$ (given by the C_2 classification algorithm). It should be noted that each of the considered classification algorithms can leave the new object unclassified (if there are no rules matching this object). The final decision for a given new object is obtained from the decision c and decision value set $V(r)$. The method of conflict resolution is described in detail in [13].

5. EXPERIMENTS

We have performed experiments on decision tables from UCI Machine Learning Repository [2] using combination of two classification algorithms C_1, C_2.

The algorithm C_1 is classification algorithm from RSESlib based on all minimal decision rules and standard voting. The classification algorithm C_2 is based on second type nondeterministic rules generated by the *ND*2 algorithm.

Some attributes in decision tables used for experiments were discretized by algorithm from RSESlib. In evaluation of the accuracy of classification algorithms on decision table (i.e., the percentage of correctly classified objects) the 5-fold cross-validation method was used.

For any considered data table, we used proposed classification method (denoted by C) based on combination of classifiers C_1 and C_2 for different values of parameter α. On testing sets the accuracy and the coverage factor were calculated. Also the *maximal relative deviation* (mrd) was calculated.

The majority of decision tables used for experiments concern medical data.

Table 1. Accuracy of classifiers based on second type nondeterministic decision rules ¯ cross-validation method.

Decision table	Classification factor	Classification algorithm						
		C_1 [(1)]	C [(1)], α [(2)]					
			1.0	0.9	0.8	0.7	0.6	0.5
Dermatology	acc × cover	84.62	85.04	84.97	85.03	**85.11**	84.62	84.59
	mrd	0.012	0.006	0.014	0.009	0.015	0.012	0.009
Ecoli	acc × cover	54.99	55.51	**56.01**	54.52	54.40	50.63	50.63
	mrd	0.038	0.037	0.033	0.026	0.027	0.020	0.020
Lymphography	acc × cover	37.70	**38.06**	38.06	38.06	38.06	38.06	38.06
	mrd	0.042	0.022	0.022	0.022	0.022	0.022	0.022
Post operative	acc × cover	65.00	65.44	65.44	65.44	65.67	67.89	**69.11**
	mrd	0.061	0.066	0.066	0.034	0.034	0.034	0.024
Primary tumor	acc × cover	59.71	**60.09**	60.09	60.09	60.09	60.09	60.09
	mrd	0.016	0.020	0.020	0.020	0.020	0.020	0.020
Iris	acc × cover	90.47	90.47	90.47	90.47	90.47	90.47	90.47
	mrd	0.018	0.018	0.018	0.018	0.018	0.018	0.018

(1) In the column marked by C_1 the classification is defined by the classification algorithm based on deterministic rules. In the column marked by C the classification is defined by the classification algorithm based on nondeterministic and deterministic rules.

(2) Confidence of nondeterministic rules generated by the algorithm is not smaller than the parameter α.

Decision table *Dermatology* contains data about the diagnosis of erythemato-squamous diseases, a real problem in dermatology [8]. Decision table *Ecoli* concerns the protein localization sites in

Escherichia coli bacteria [12]. The classification task of decision table *Postoperative* is to determine when patients in a postoperative recovery area should be sent to the next one [4]. *Lymphography* and *Primary Tumor* data are two of three domains provided by the University Medical Centre, Institute of Oncology from Ljubljana that has repeatedly appeared in the machine learning literature [6,11].

Table 1 includes the results of our experiments.

For almost all, except one, decision tables the classification quality measured by *accuracy* × *coverage* was better for the proposed classification algorithm C than in the case of the classification algorithm from RSESlib based only on minimal rules with standard voting C_1. For one decision table Iris, the classification quality for the classification algorithms C and C_1 was the same. For all decision tables, the maximal relative deviation was no greater than 5% in the case when we used the classification algorithm C. Hence, using the classification algorithm C lead to stable classification.

For obtaining those results, it was necessary to optimize the threshold α for each data table. This means that the parameter α should be tuned to the data.

6. CONCLUSIONS

Results of experiments with second type nondeterministic rules - generated by the ND2 algorithm - are showing that these rules can lead to improve the classification quality. We have demonstrated this by using a classification algorithm based on minimal decision rules and nondeterministic rules. Experiments have shown that proposed classifier make sense, because improve classification accuracy for the most decision tables. This is very important in diagnosis support systems where classification error can be connected with serious consequences for the patient.

Second type nondeterministic rules can be used in any rule-based classification systems. Especially in diagnosis support systems or medical expert systems, because this kind of rules have usually large support (patients proves the rule), which makes them more accepted by physicians.

At this moment the proposed classification algorithm uses minimal rules (from RSESlib) and second type nondeterministic rules (ND2 algorithm). Since the algorithm for constructing minimal rules has exponential computational complexity, and ND2 algorithm has polynomial computational complexity, we plan to use others rule-based classifiers instead of the minimal rules (e.g. based on subsets of minimal decision rules or decision trees).

BIBLIOGRAPHY

[1] AGRAWAL R., IMIELINSKI T, SWAMI A., Mining Association Rules Between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington D.C., New York, ACM Press, 1993, pp. 207–216.

[2] ASUNCION A., NEWMAN D.J., UCI Machine Learning Repository, University of California, Irvine School of Information and Computer Sciences, 2007.

[3] BAZAN J.G., SZCZUKA M.S., WOJNA A., WOJNARSKI M., On the Evolution of Rough Set Exploration System, RSCTC 2004, LNAI, Vol. 3066, Springer, Heidelberg, 2004, pp. 592–601.

[4] BUDIHARDJO A., GRZYMAŁA-BUSSE J., WOOLERY L., Program LERS_LB 2.5 as a tool for knowledge acquisition in nursing, Proceedings of the 4th Int. Conference on Industrial & Engineering Applications of AI & Expert Systems, 1991, pp. 735–740.

[5] CARLIN U., KOMOROWSKI J., OHRN A., Rough set analysis of patients with suspected acute appendicitis, in: BOUCHON-MEUNIER G., YAGER R.R. (eds.), Proc. Seventh Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'98), Paris, France, 1998, pp. 1528–1533.

[6] CESTNIK G., KONENENKO I., BRATKO I., Assistant-86: A Knowledge-Elicitation Tool for Sophisticated Users, Progress in Machine Learning, Sigma Press, 1987, pp. 31–45.

[7] DELIMATA P., MARSZAŁ-PASZEK B., MOSHKOV M., PASZEK P., SKOWRON A., SURAJ Z., Comparison of Some Classification Algorithms Based on Deterministic and Nondeterministic Decision Rules, Transactions on Rough Sets XII, LNCS 6190, Springer, Heidelberg, 2010, pp. 90–105.

[8] DEMIROZ G., GOVENIR H.A., ILTER N., Learning Differential Diagnosis of Eryhemato-Squamous Diseases using Voting Feature Intervals, Artificial Intelligence in Medicine, Vol. 13, 1998, pp. 147–165.

[9] GRZYMAŁA-BUSSE J.W., LERS – A Data Mining System. The Data Mining and Knowledge Discovery Handbook, Springer, New York, 2005, pp. 1347–1351.
[10] MICHALSKI R., http://www.mli.gmu.edu/michalski/
[11] MICHALSKI R., MOZETIC I., HONG J., LAVRAC N., The Multi-Purpose Incremental Learning System AQ15 and its Testing Applications to Three Medical Domains, Proceedings of the Fifth National Conference on Artificial Intelligence, Philadelphia, Morgan Kaufmann, 1986, pp. 1041–1045.
[12] NAKAI K., KANEHISA M., Expert System for Predicting Protein Localization Sites in Gram-Negative Bacteria, Proteins: Structure, Function, and Genetics, Vol. 11, 1991, pp. 95–110.
[13] PASZEK P., MARSZAŁ-PASZEK B., Deterministic and Nondeterministic Decision Rules in Classification Process, Journal of Medical Informatics and Technologies, Vol. 15, 2010, pp. 87–92.
[14] PAWLAK Z., Rough Sets: Theoretical aspects of reasoning about data, Boston: Kluwer Academic Publishers, 1991.
[15] PAWLAK Z., SKOWRON A., Rudiments of Rough Sets, Information Sciences 177, pp. 3–27; Rough Sets: Some Extensions, Information Sciences 177, pp. 28–40; Rough Sets and Boolean Reasoning, Information Sciences, Vol. 177, 2007, pp. 41–73.
[16] RISSANEN J., Modelling by Shortest Data Description, Automatica 14, 1978, pp. 465–471.
[17] Rosetta: http://www.lcb.uu.se/tools/rosetta/.
[18] Rough Set Exploration System: http://logic.mimuw.edu.pl/rses.
[19] SWINIARSKI R., Rough sets Bayesian methods applied to cancer detection, RSCTC'98, LNAI, Vol. 1424, Springer-Verlag, 1998, pp. 609–616.
[20] TRIANTAPHYLLOU E., FELICI G., (eds.), Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques, Springer Science and Business Media, LLC, New York, 2006.
[21] TSUMOTO S., Modelling Medical Diagnostic Rules Based on Rough Sets, RSCTC 1998, LNCS, Vol. 1424, Springer-Verlag, Berlin, 1998, pp. 475–482.
[22] WOOLERY K., GRZYMAŁA–BUSSE J., Machine learning for an Expert System to Predict Preterm Birth Risk, J. Am. Med. Informatics Assoc. 1, 1994, pp. 439–446.