Application of normalized decision measures to the new case classification

Dominik Ślęzak[1,2], Jakub Wróblewski[1,2]

[1] Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
[2] Institute of Mathematics, Warsaw University
Banacha 2, 02-097 Warsaw, Poland

Abstract. The optimization of rough set based classification models with respect to parameterized balance between a model's complexity and confidence is discussed. For this purpose, the notion of a parameterized approximate inconsistent decision reduct is used. Experimental extraction of considered models from real life data is described.

1 Introduction

While reasoning about a domain specified by our needs, we usually base on the information gathered by the analysis of a sample of objects. The rough set theory ([3]) assumes that a universe of known objects is the only source of knowledge, which can be applied to construct models of reasoning about new cases. Reasoning can be stated, e.g., as a classification problem, concerning prediction of a decision attribute under information provided over conditional attributes. For this purpose, one stores data within decision tables, where each training case drops into one of predefined decision classes.

Classification of new objects is performed by analogy, e.g., by the usage of "if...then..." decision rules calculated over the universe of a given table. Theoretical studies related to the *Minimum Description Length Principle (MDLP)* (cf. [5]), as well as practical experiences, lead to the same conclusion: *Optimal rule-based classification models should be extracted from data by tuning up a parameterized tradeoff between the overall confidence and complexity of the decision rule collections.*

Confidence of a rule-based model can be interpreted as the expected chance of correct classification of new cases. To express such a chance numerically, we need to set up the model of representing inexact *conditions*$\rightarrow$*decision* dependencies. Then, we are able to evaluate the degree of decision information provided by each particular subset of conditional attributes, and to express the dynamics of this degree under the attribute reduction. In the same way, one can interpret the complexity of a given collection as opposite to the expected chance of recognizing new cases by its decision rules. It leads to a rough set based version of MDLP, related to the fundamental concept of searching for approximate decision

W. Ziarko and Y. Yao (Eds.): RSCTC 2000, LNAI 2005, pp. 553−560, 2001.

reducts: *Given a rule-based decision model, any simplification which approximately preserves the expected chance of correct classification should be performed to increase the expected chance of the new case recognition.*

The above principle can be regarded as the starting point for the design of the process of the rule-based classification model optimization. In the paper, we discuss exemplary methodology of setting up the foregoing items, concerning the adjustment of thresholds, voting measures, etc.. Accordingly, Section 2 includes the basics of rough set based classification techniques (cf. [4], [6]). Sections 3 and 4 outline exemplary extensions of these methods by introducing the notion of an *approximate reduct* based on a *normalized decision function* (cf. [7]). In Sections 5 and 6 we present the main contribution – the classification algorithm based on the family of parameterized decision functions. Section 7 contains experimental verification of the performance of the proposed classification framework.

2 Decision rules and reducts

In the rough set theory sample of data takes the form of an information system $\mathbb{A} = (U, A)$, where each attribute $a \in A$ is a function $a : U \to V_a$ into the set of all possible values on a. Reasoning about data can be stated as, e.g., a classification problem, where a distinguished decision is to be predicted under information over conditional attributes. In this case, we consider a triple $\mathbb{A} = (U, A, d)$, called a decision table, where, for the decision attribute $d \notin A$, values $v_d \in V_d$ correspond to mutually disjoint decision classes of objects.

Definition 1. *Let $\mathbb{A} = (U, A, d)$, where $A = \langle a_1, \ldots, a_{|A|} \rangle$, be given. For any $B \subseteq A$, $B = \langle a_{i_1}, \ldots, a_{i_{|B|}} \rangle$, the* B**-information function** *over U is defined by*

$$\overrightarrow{Inf}_B(u) = \langle a_{i_1}(u), \ldots, a_{i_{|B|}}(u) \rangle \tag{1}$$

The B**-indiscernibility relation** *is the equivalence relation defined by*

$$IND_{\mathbb{A}}(B) = \{(u, u') \in U \times U : \overrightarrow{Inf}_B(u) = \overrightarrow{Inf}_B(u')\} \tag{2}$$

Each $u \in U$ induces a B**-indiscernibility class** *of the form*

$$[u]_B = \{u' \in U : (u, u') \in IND_{\mathbb{A}}(B)\} \tag{3}$$

which can be identified with vector $\overrightarrow{Inf}_B(u)$.

Indiscernibility enables us to express global dependencies as follows:

Definition 2. *Given $\mathbb{A} = (U, A, d)$, we say that $B \subseteq A$* **defines** *d* **in** *$\mathbb{A}$ iff*

$$IND_{\mathbb{A}}(B) \subseteq IND_{\mathbb{A}}(\{d\}) \tag{4}$$

or, equivalently, iff for any $u \in U$ the following u-oriented rule is valid in $\mathbb{A}$:

$$\bigwedge_{a \in B}(a = a(u)) \Rightarrow (d = d(u)) \tag{5}$$

We say that $B \subseteq A$ is a **decision reduct** *iff it defines d and none of its proper subsets does it.*

Given a collection of subsets $\mathcal{B} \subseteq \mathcal{P}(A)$ which define d, we can classify any new case $u_{new} \notin U$ by using the bunch of decision rules of the form (5). The only requirement is that u_{new} must be comparable with U with respect to at least one (several) of $B \in \mathcal{B}$. To improve such understood recognition of new cases, we base on (approximate) decision reducts of possibly low complexity, expressible in various terms (cf. [2], [8]).

3 Normalized decision functions

In a consistent decision table $\mathbb{A} = (U, A, d)$ – where each indiscernibility class of $IND_{\mathbb{A}}(A)$ has one decision value – decision rules lead to deterministic classification within the universe U. In case of inconsistent decision tables (non-deterministic dependencies among attributes), we should specify the way of dealing with uncertainty.

Definition 3. *Let $\mathbb{A} = (U, A, d)$, linear ordering $V_d = \langle v_1, \ldots, v_r \rangle$, $r = |V_d|$, and $B \subseteq A$ be given. By a B-***rough membership distribution** *we call the function $\overrightarrow{\mu}_{d/B} : U \to \triangle_{r-1}$ defined by*[1]

$$\overrightarrow{\mu}_{d/B}(u) = \langle \mu_{d=1/B}(u), \ldots, \mu_{d=r/B}(u) \rangle \tag{6}$$

where, for $k = 1, \ldots, r$, $\mu_{d=k/B}(u) = |\{u' \in [u]_B : d(u') = v_k\}| \,/\, |[u]_B|$ is the rough membership function (cf. [4]) labeling $u \in U$ with the degree of hitting the k-th decision class with its B-indiscernibility class $[u]_B$.

Distributions of the form (6) seem to express the most accurate knowledge about dependencies of the decision on conditions (cf. [4], [7], [9]). Thus, it should be possible to model various B-based reasoning strategies as functions acting over $\overrightarrow{\mu}_{d/B}$ by "forgetting" a part of frequency information, which is redundant with respect to a given approach.

Definition 4. *([7]) Let $\mathbb{A} = (U, A, d)$, $B \subseteq A$ and $\phi : \triangle_{r-1} \to \triangle_{r-1}$, $r = |V_d|$, be given. We say that ϕ is a* **normalized decision function** *(ND-function) iff it satisfies the following, logical and monotonic consistency assumptions:*

$$\forall_k \, (s[k] = 0 \Rightarrow \phi(s)[k] = 0) \quad \wedge \quad \forall_{k,l} \, (s[k] \leq s[l] \Rightarrow \phi(s)[k] \leq \phi(s)[l]) \tag{7}$$

Function $\overrightarrow{\phi}_{d/B}(u) = \phi(\overrightarrow{\mu}_{d/B}(u))$ is called a normalized $\phi_{d/B}$-decision function.

According to (7), a positive weight cannot be attached to a non-supported event and the relative chances provided by the reasoning strategy cannot contradict those derived directly from an information source.

[1] For any $r \in \mathbb{N}$, $\triangle_{r-1}$ denotes the $(r-1)$-dimensional simplex of real valued vectors $s = \langle s[1], \ldots, s[r] \rangle$ with non-negative coordinates, such that $\sum_{k=1}^{r} s[k] = 1$.

Example 1. Consider ND-functions $\partial, m : \triangle_{r-1} \to \triangle_{r-1}$ defined by

$$\begin{array}{rl} \partial(s)[k] = & \left\{ \begin{array}{ccc} |\{l : s[l] > 0\}|^{-1} & for & s[k] > 0 \\ \left(|\{l : s[l] = \max(s)\}|^{-1} \right. & for & \left. s[k] = \max(s) \right) \\ 0 & otherwise & \end{array} \right. \\ \left(m(s)[k] = \right) & \end{array} \tag{8}$$

where $\max(s) = \max_k s[k]$. One can see that by combining $\overrightarrow{\mu}_{d/B}(u)$, $u \in U$, with ∂ and m we obtain the uniform distributions spanned over the subsets: (1) $\partial_{d/B}(u) \subseteq V_d$ induced by the generalized decision (cf. [6]); (2) $m_{d/B}(u) \subseteq V_d$ of decision values taking the maximum over the coordinates of $\overrightarrow{\mu}_{d/B}(u)$.

4 Normalized decision measures

In real life applications, the search for attributes which approximately preserve ϕ-decision distributions seems to be promising. We are likely to understand an approximate ϕ-decision reduct as a minimal subset of conditions, which almost preserves information about decision in terms of a given ND-function.

Definition 5. *Let $\mathbb{A} = (U, A, d)$ and $\phi \in \triangle_{r-1} \to \triangle_{r-1}$, $r = |V_d|$, be given. The* **normalized ϕ-decision measure** *$E_{\phi/\mathbb{A}} : \mathcal{P}(A) \to [0, 1]$ is defined by*[2]

$$E_{\phi/\mathbb{A}}(B) = \frac{1}{|U|} \sum_{u \in U} \langle \overrightarrow{\mu}_{d/B}(u) | \overrightarrow{\phi}_{d/B}(u) \rangle \tag{9}$$

The value of (9) equals to the average probability that objects $u \in U$ will be correctly classified by a random $\overrightarrow{\phi}_{d/B}(u)$-weighted choice among decision classes ([7]). Thus, ϕ-decision measures enable us to evaluate subsets numerically with respect to their capabilities of ϕ-defining decision.

Definition 6. *Let $\mathbb{A} = (U, A, d)$, $\varepsilon \in [0, 1)$ and $\phi \in \triangle_{r-1} \to \triangle_{r-1}$, $r = |V_d|$, be given. We say that subset $B \subseteq A$* **ε-approximately ϕ-defines** *d iff*

$$E_{\phi/\mathbb{A}}(B) \geq (1 - \varepsilon) E_{\phi/\mathbb{A}}(A) \tag{10}$$

We say that $B \subseteq A$ is an **ε-approximate ϕ-decision reduct** *iff it ϕ-defines d ε-approximately and none of its proper subsets does it.*

Two parameters can be tuned up while searching for optimal conditions for classification: (1) ND-function ϕ responsible for the way of understanding inexact *conditions→decision* dependencies, and (2) the degree $\varepsilon \in [0, 1)$ up to which we are likely to neglect the decrease of ϕ-decision information provided by smaller subsets $B \subseteq A$ with respect to the whole of A. In case of the first parameter, it is easier to handle a numeric factor responsible for adjusting a specific function:

[2] By "$\langle \cdot | \cdot \rangle$" we mean the inner product of two distribution vectors.

Definition 7. *Let $\mathbb{A} = (U, A, d)$ be given. For any $x \in (0, +\infty)$, we define the* **normalized x-decision function** *by putting, for any $s \in \triangle_{r-1}$, $r = |V_d|$,*

$$x(s)[k] = (s[k])^x \,/\, \textstyle\sum_{l=1}^{r}(s[l])^x \tag{11}$$

Proposition 1. *All x-decision functions satisfy (7). For any $s \in \triangle_{r-1}$,*

$$\lim_{x\to 0^+} x(s) = \partial(s) \quad \wedge \quad \lim_{x\to+\infty} x(s) = m(s) \tag{12}$$

For any $\mathbb{A} = (U, A, d)$, $B \subseteq A$, $0 < x_1 < x_2 < +\infty$, we have

$$E_{\partial/\mathbb{A}}(B) \leq E_{x_1/\mathbb{A}}(B) \leq E_{x_2/\mathbb{A}}(B) \leq E_{m/\mathbb{A}}(B) \tag{13}$$

where equalities hold iff $E_{\partial/\mathbb{A}}(B) = E_{m/\mathbb{A}}(B)$.

One can see that the obtained x-parameterized family covers densely enough all possible ways of performance of ND-functions over training data.

5 Optimization of approximate reducts

In Section 1 we suggested to relate the overall confidence of a rule-based decision model to the expected chance of correct classification of new cases. Above, we argued that the quantities of normalized decision measures can be interpreted in that way. Analogously, let us now propose an exemplary measure of the expected chance of recognizing new cases by decision rules generated by a given $B \subseteq A$.

Definition 8. *Let $\mathbb{A} = (U, A, d)$ be given. The* **normalized coverage measure** *$cov_{\mathbb{A}} : \mathcal{P}(A) \to [0, 1]$ is defined by*

$$cov_{\mathbb{A}}(B) = \tfrac{1}{|U|} \textstyle\sum_{u\in U} \mu_B(u) \tag{14}$$

where $\mu_B(u) = |[u]_B|/|U|$ is the frequency of occurrence of vector $\overrightarrow{Inf}_B(u)$ in $\mathbb{A}$.

One should realize that this is just one of possibilities of estimating the recognition probability (cf. [8]). Still, we would like to proceed with this measure, because it turns out to be flexible enough with respect to applications.

The exemplary procedure presented below searches for an ε-approximate x-decision reduct $B \subseteq A$ with the highest possible coverage $cov_{\mathbb{A}}(B)$, by following a randomly generated permutation $\sigma \in \Sigma_{|A|}$ over conditional attributes[3,4]. First, starting with $B = \{a_{\sigma(1)}\}$, we add the foregoing attributes until B begins to x-define d in ε-approximate way. The second part reflects the optimization principle formulated at the beginning of this paper: *We try to reduce a model until it ε-approximately preserves x-decision information, to increase $cov_{\mathbb{A}}$ as the measure of predicted average chance of the new case recognition.*

[3] We denote by Σ_n the set of all n-element permutations, i.e. "1–1" functions $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$. We use $\sigma \in \Sigma_{|A|}$ to re-order $A = \langle a_1, \ldots, a_{|A|}\rangle$.

[4] We base the search for $cov_{\mathbb{A}}$-maximal approximate decision reducts on random heuristics, because this optimization problem is NP-hard (cf. [7]).

Algorithm: Approximate reducts generation
Input: Decision table $\mathbb{A} = (U, A, d)$, permutation σ of A, $\varepsilon \in [0, 1)$, $x \in (0, +\infty)$.
Output: ε-approximate x-decision reduct.

```
1.  B := {a_σ(1)}; cov := cov_𝔸(B); app := E_x/𝔸(B); i := 2
2.  while app < (1 − ε)E_x/𝔸(A) begin
3.      B := B ∪ {a_σ(i)}
4.      (cov, app) = Test(B, x)
5.      i := i + 1
6.  end while
7.  maxcov := 0
8.  do
9.      stop := 1
10.     for each a_j ∈ B
11.       cov := cov_𝔸(B \ {a_j})
12.       app := E_x/𝔸(B \ {a_j})
13.       if app < (1 − ε)E_x/𝔸(A) and cov > maxcov then begin
14.         maxcov := cov
15.         maxj := j
16.         stop := 0
17.       end if
18.     end for
19.     if stop = 0 then B := B \ {a_maxj}
20. while stop = 0
21. return B
```

6 Optimization of the approximate reduct collections

A rule-based decision model should correspond to more than one subset of conditions. Thus, we construct systems composed of collections of *classifying agents* based on different ε-approximate x-decision reducts, obtained as $cov_{\mathbb{A}}$-optimal while following a number of randomly generated permutations. Since it is not known how to adjust the best configuration of $\varepsilon \in [0, 1)$ and $x \in (0, +\infty)$, we consider collections of (ε, x)-parameterized agents initialized randomly, to simulate a kind of the adaptation process searching through the space of $[0, 1) \times (0, +\infty)$.

Given $\mathcal{B} \subseteq \mathcal{P}(A) \times (0, +\infty)$ as the collection of obtained parameterized reducts, one needs also to specify the way of voting between particular agents. In general, negotiations concerning prediction of the decision value for a given u lead to the choice of $v_k \in V_d$ with a maximal value of a voting measure, calculated from u-oriented x-decision rules induced by particular elements of $\mathcal{B}$. Below, one can find examples of such voting measures:

$$\begin{aligned} VOTE1(\mathcal{B}) &= \textstyle\sum_{(B,x)\in\mathcal{B}:\, \mu_B(u)>0} \mu_B(u) x_{d=k/B}(u) \\ VOTE2(\mathcal{B}) &= \textstyle\sum_{(B,x)\in\mathcal{B}:\, \mu_B(u)>0} \mu_{d=k/B}(u) x_{d=k/B}(u) \\ VOTE3(\mathcal{B}) &= \textstyle\sum_{(B,x)\in\mathcal{B}:\, \mu_B(u)>0} x_{d=k/B}(u) \end{aligned} \tag{15}$$

Another problem concerns the fact that although all agents can be used to classify new objects, a subset of them often performs much better (cf. Fig. 1). We apply a specific genetic algorithm to search for optimal sub-collections of agents. In particular, it results with an indirect optimization process concerned with the ranges of (ε, x)-parameters.

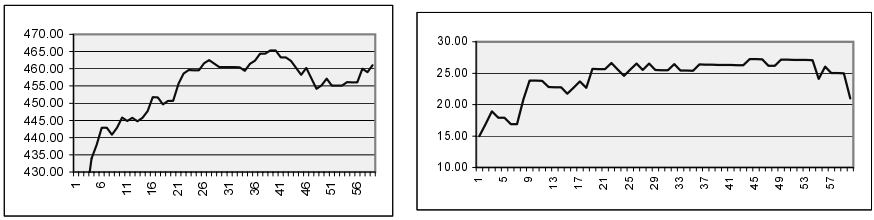

Fig. 1. Classification quality (vertical axis) and the number of agents in a team (horizontal axis) – examples obtained for "DNA splices" and "primary tumor" data sets.

7 Experimental results

The methodology described in the paper was implemented and tested on several data sets obtained from [1]. Results presented in Table 1 concern two of them: (1) "DNA splices" – 2000 objects, 20 symbolic attributes, 3 decision classes; (2) "Primary tumor" – 339 objects, 17 symbolic attributes, 22 decision classes.

Voting	*Approx*	*Mode*	*Result*
1	–	–	82.06
2	–	–	86.96
3	–	–	**93.75**
3	0.5	–	68.74
3	0.4	–	87.05
3	0.3	–	93.39
3	0.2	–	**94.97**
3	0.1	–	93.14
3	0.2	exp(−2)	94.27
3	0.2	exp(−1)	94.68
3	0.2	exp(0)	95.01
3	0.2	exp(1)	94.93
3	0.2	exp(2)	**95.10**

Voting	*Approx*	*Mode*	*Result*
1	–	–	43.01
2	–	–	**43.29**
3	–	–	42.44
2	0.5	–	43.28
2	0.4	–	**44.08**
2	0.3	–	43.48
2	0.2	–	40.80
2	0.1	–	39.53
2	0.4	exp(−2)	44.46
2	0.4	exp(−1)	44.67
2	0.4	exp(0)	**44.70**
2	0.4	exp(1)	42.85
2	0.4	exp(2)	42.79

Table 1. Experimental results for "DNA splices" and "primary tumor" data sets, obtained by voting among optimized collections of ε-approximate x-decision reducts: **(1)** The choice of a measure from (15) corresponds to the *Voting* column; **(2)** Quantities of ε and x are chosen randomly from small intervals around values in the *Approx* and *Mode* columns, where symbol "–" means the uniform random choice from a wider interval; **(3)** Average percent of tested objects classified correctly for particular settings is presented in the *Result* column. Cross-validation (CV-5) was used in case of "primary tumor" data.

Experiments presented in Table 1 were performed with various settings:

1. Optimal voting measure was selected by setting other parameters randomly;
2. For the best voting method, several values of $\varepsilon \in [0, 0.5]$ were tested;
3. For the best voting method and ε-thresholds selected from the small interval around the best value found previously, different values of parameter $x \in [\exp(-2), \exp(2)]$ were tested.

It is worth noting that the best results obtained in our experiments are close to the best results ever found.

8 Conclusions

Parameterized tradeoff between model complexity and its accuracy was discussed. To handle it in a flexible way, the notion of a parametrized ε-approximate x-decision reduct was used. Main issues concerning implementation of the classification algorithm based on described methodology were outlined.

Experiments performed on two "benchmark" data sets show that our technique is relatively fast and very efficient. It is worth noting that best results for these sets were obtained using significantly different voting and (x, ε)-settings. It suggests us to consider the adaptive mechanisms of tuning up these parameters in the nearest future.

Acknowledgements

This work was supported by the grants of Polish National Committee for Scientific Research (KBN) No. $8T11C02319$ and $8T11C02419$. "Primary tumor" was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklic for its providing.

References

1. Michie D., Spiegelhalter D.J., Taylor C.C. (eds.): Machine Learning, Neural and Statistical Classification. Ellis Horwood Limited (1994). Data avaliable at: http://www.ics.uci.edu/~mlearn/MLRepository.html.
2. Nguyen, S.H., Skowron, A., Synak, P.: Discovery of data patterns with applications to decomposition and classification problems. In: L. Polkowski, A. Skowron (eds.), Rough Sets in Knowledge Discovery, Physica Verlag, Heidelberg (1998) pp. 55–97.
3. Pawlak, Z.: Rough sets – Theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht (1991).
4. Pawlak, Z., Skowron, A.: Rough membership functions. In: R.R. Yaeger, M. Fedrizzi, and J. Kacprzyk (eds.), Advances in the Dempster Shafer Theory of Evidence, John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore (1994) pp. 251–271.
5. Rissanen, J.: Modeling by the shortest data description. Authomatica, **14** (1978) pp. 465–471.
6. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. In: R. Słowiński (ed.), Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, Kluwer Academic Publishers, Dordrecht (1992) pp. 311–362.
7. Ślęzak, D.: Normalized decision functions and measures for inconsistent decision tables analysis. To appear in Fundamenta Informaticae (2000).
8. Wróblewski J.: Genetic algorithms in decomposition and classification problem. In: L. Polkowski, A. Skowron (eds.), Rough Sets in Knowledge Discovery, Physica Verlag, Heidelberg (1998) pp. 471–487.
9. Ziarko, W.: Decision Making with Probabilistic Decision Tables. In: N. Zhong, A. Skowron and S. Ohsuga (eds.), Proc. of the Seventh International Workshop RSFDGrC'99, Yamaguchi, Japan, LNAI **1711** (1999) pp. 463–471.