Order based genetic algorithms for the search of approximate entropy reducts *

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

1. Department of Computer Science University of Regina, Regina, Canada
2. Polish-Japanese Institute of Information Technology, Warsaw, Poland

Abstract. We use entropy to extend the rough set based notion of a reduct. We show that the order based genetic algorithms, applied to the search of classical decision reducts, can be used in exactly the same way in case of extracting optimal approximate entropy reducts from data.

1 Introduction

In the theory of *rough sets* [4] a universe of objects is the only source of knowledge usable to construct the reasoning models. In classification problems the goal is to *approximate* values of a *decision attribute* under information provided by *conditional attributes*. New objects are classified using *"if..then.." decision rules* learnt from the known cases.

Due to the *Minimum Description Length Principle (MDLP)* [7], adapted to rough sets e.g. in [2], we search for the simplest decision rule based models, using various heuristics. As one of such heuristics, the *order based genetic algorithm* for extraction of minimal *decision reducts* was proposed [11]. We modify it to search for *approximate reducts* – the attribute subsets inducing rules, which approximate decision classes accurately enough.

We label each subset B of available attributes A with its *entropy* $H(B)$ [3], opposite to the *strength* of the model induced by B, *conditional entropy* $H(d/B)$ – the model's *inaccuracy* in predicting decision [2, 8], and $H(B/d)$ – its *sensitivity* [6]. Due to the *Approximate Entropy Reduction Principle (AERP)* [10], we minimize $H(B)$ ($H(B/d)$) keeping $H(d/B)$ at a reasonable level.

2 Rough sets and probabilities

In [4] data is represented as an *information system* $\mathbf{A} = (U, A)$, where U is the *universe* of *objects* and each *attribute* $a \in A$ provides a function $a : U \to V_a$ into the set of *values* on a. For any $B \subseteq A$ and $u \in U$, we define *B-information vector* $B(u) = \langle b_1(u), \ldots, b_{|B|}(u) \rangle$, where $b_j(u)$ is the value of $b_j \in B$. The set of all such vectors equals to $V_B^U = \{B(u) : u \in U\}$. Each $B \subseteq A$ induces a U-partition with *B-indiscernibility classes* $\|(B, w)\|_{\mathbf{A}} = \{u \in U : B(u) = w\}$, for $w \in V_B^U$.

* Supported by Polish National Committee for Scientific Research (KBN) grant No. 8T11C02519, as well as the Research Centre of PJIIT.

Pairs (B, w), $B \subseteq A$, $w \in V_B^U$, are *patterns*, interpreted as conjunctions of *descriptors* (a, v), $a \in B$, $v \in V_a$. Classes $\|(B, w)\|_{\mathbf{A}}$ are their *supports*. The data driven probability $P(w) = |\|(B, w)\|_{\mathbf{A}}| \,/\, |U|$ reflects the *strength* of (B, w).

The task of analysis is often concerned with defining a distinguished *decision* by the rest of attributes. In this case, we represent data as a *decision system* $\mathbf{A} = (U, A \cup \{d\})$, where $d \notin A$, $V_d = \{1, \ldots, |V_d|\}$. For each $k \in V_d$, we define the k-th decision class $X_k = \{u \in U : d(u) = k\}$. The data driven probability $P(k/w) = |\|(B, w)\|_{\mathbf{A}} \cap X_k| \,/\, |\|(B, w)\|_{\mathbf{A}}|$ of $k \in V_d$ conditioned by $w \in V_B^U$ corresponds to the *precision* of *decision rule* $B = w \Rightarrow d = k$. In the rough set applications, one often operates with the *object related decision rules* $B = B(u) \Rightarrow d = d(u)$, for $u \in U$. Then, precision (strength) is expressed as $P(d(u)/B(u))$ $(P(B(u)))$.

The precision-strength balance refers to the MDLP principle [7]. The strength can be replaced with the *complexity* of the rule's description or, e.g., the rule's *sensitivity* $P(B(u)/d(u)) = |\|(B, B(u))\|_{\mathbf{A}} \cap X_{d(u)}| \,/\, |X_{d(u)}|$, opposing precision $P(d(u)/B(u))$ in the *Relative Operating Characteristic (ROC)* approach [6].

3 Information entropy

Entropy evaluates information, which we get from the fact that a given random event occurred [2, 3]. Given the event's probability $p > 0$, it is defined as $h(p) = -\log p$. For instance, given $w \in V_B^U$, one can state the entropy of pattern (B, w) as equal to $-\log P(w)$. In its generalized form, entropy evaluates random distributions $\overrightarrow{p} = \langle p_1, \ldots, p_r \rangle$ with the expected degree of information $H(\overrightarrow{p}) = -\sum_{k:\, p_k > 0} p_k \log p_k$. For instance, we can label $B \subseteq A$ with its entropy

$$H(B) = -\sum_{w \in V_B^U} P(w) \log P(w) \tag{1}$$

interpreted as the average degree of information about objects, obtainable from knowledge about their values on B. *Conditional entropy* $H(d/B)$ of d given B evaluates information obtainable from d, given already provided B [2, 3]. Similarly, we interpret $H(B/d)$. By definition, we have the following:

$$H(d/B) = H(B \cup \{d\}) - H(B) \qquad H(B/d) = H(B \cup \{d\}) - H(\{d\}) \tag{2}$$

We can use $H(d/B)$ to label each $B \subseteq A$ with the amount of uncertainty concerning d under information about B. We have the following inequalities:

$$0 \leq H(d/A) \leq H(d/B) \leq H(\{d\}) \tag{3}$$

where: $H(d/A) = 0$ holds, iff A *defines* d, i.e. iff $P(d(u)/B(u)) = 1$ for any $u \in U$; $H(d/A) = H(d/B)$ holds, iff B makes d *conditionally independent* from $A \setminus B$, i.e. iff $P(d(u)/B(u)) = P(d(u)/A(u))$ for any $u \in U$; and $H(d/B) = H(\{d\})$ holds, iff d is *independent* from B. We have also the following equalities [10]:

$$H(B) = -\log G(B) \quad H(d/B) = -\log G(d/B) \quad H(B/d) = -\log G(B/d) \tag{4}$$

where quantities $G(B) = \sqrt[|U|]{\Pi_{u \in U} P(B(u))}$, $G(d/B) = \sqrt[|U|]{\Pi_{u \in U} P(d(u)/B(u))}$ and $G(B/d) = \sqrt[|U|]{\Pi_{u \in U} P(B(u)/d(u))}$ are the average strength, precision and sensitivity of the object related decision rules induced by $B \subseteq A$.

4 Approximate entropy reducts

Due to the MDLP principle, we should tend to the model's simplification, unless it causes a loss of its accuracy. This idea corresponds to the rough set notion of a *decision reduct*: an irreducible $B \subseteq A$ defining d in $\mathbf{A} = (U, A \cup \{d\})$. If $\mathbf{A}$ is *inconsistent*, i.e. even the whole A does not define d, a question about the reduction criterion arises. In [8, 10] we considered *μ-decision reducts*: irreducible subsets $B \subseteq A$ such that $P(d(u)/B(u)) = P(d(u)/A(u))$ for any $u \in U$.

Subset $B \subseteq A$ is a μ-decision reduct, iff it is a *Markov boundary* of d for the product distribution P over $A \cup \{d\}$, i.e. it is an irreducible subset of attributes (random variables), which provides the same probabilistic information about d as A [5]. Equivalently, B is a μ-decision reduct, iff $H(d/B) = H(d/A)$ and $H(d/B) > H(d/B \setminus \{a\})$ for any $a \in B$. If $\mathbf{A} = (U, A \cup \{d\})$ is *consistent*, i.e. A defines d, then such reducts-boundaries coincide with classical decision reducts. B is a decision reduct, iff $H(d/B) = 0$ and $H(d/B \setminus \{a\}) > 0$ for any $a \in B$.

The reduction of attributes causes potential growth of conditional entropy, i.e., potential average decrease of precision of the object related decision rules. Let us consider constraint $G(d/B) \geq (1-\varepsilon)G(d/A)$, pointing at subsets $B \subseteq A$, which induce rules being on average *ε-almost* as precise as those induced by the whole A. By taking the logarithm of both sides, we get the following criterion:

$$H(d/B) + \log(1-\varepsilon) \leq H(d/A) \tag{5}$$

We say that B is an *ε-approximate μ-decision reduct*, iff it is an irreducible set satisfying (5). Originally, reducts were evaluated due to the number of attributes involved. In [10] the following generalization was proposed: Given $\varepsilon \in [0, 1)$, the *Minimal ε-Approximate Decision Reduct Problem (MεDRP)* is the task of finding a minimal (by cardinality) B satisfying (5). One can also evaluate reducts with numbers of distinct rules or the measures of strength and sensitivity. The *Minimal Rule (MRεDRP)*, *H-Strength Optimal (StεDRP)* and *H-Sensitivity Optimal (SeεDRP) ε-Approximate Decision Reduct Problems* are the tasks of finding such minimal (by cardinality) ε-approximate μ-decision reduct $C \subseteq A$ that

$$f(C) = \min_{B \subseteq A:\, B \text{ satisfies } (5)} f(B) \tag{6}$$

for $f : \mathcal{P}(A) \to \mathbf{R}$ defined by $f(B) = \left|V_B^U\right|$, $H(B)$, and $H(B/d)$, respectively.

5 Order based genetic algorithms

The MεDRP, MRεDRP, StεDRP, and SeεDRP problems are NP-hard for any $\varepsilon \in [0, 1)$ [10]. Therefore, one cannot expect fast and reliable tools for solving them in a deterministic way. We propose to extend the *order based genetic algorithm (o-GA)* for searching for minimal decision reducts [11], in purpose of finding (sub)optimal ε-approximate μ-decision reducts. As a *hybrid algorithm* [1], our o-GA consists of two parts:

1. *Genetic part*, where each chromosome encodes a permutation of attributes
2. *Heuristic part*, where permutations τ are put into the following algorithm:

ε-REDORD algorithm:

1. Let $\mathbf{A} = (U, A \cup \{d\})$ and $\tau : \{1, .., |A|\} \to \{1, .., |A|\}$ be given; Let $B_\tau = A$;
2. For $i = 1$ to $|A|$ repeat steps 3 and 4;
3. Let $B_\tau \leftarrow B_\tau \setminus \{a_{\tau(i)}\}$;
4. If B_τ does not satisfy condition (5), undo step 3.

Comparing to [11], we replace the condition of defining d with criterion (5).

Proposition 1. *ε-REDORD always gives an ε-approximate μ-decision reduct. For any ε-approximate μ-decision reduct B there exists such τ that $B = B_\tau$.*

Each genetic algorithm simulates the evolution of *individuals* [1, 11]. Its behavior depends on specification of the *fitness function*, which evaluates individuals. In the proposed o-GA, we define fitness of a given permutation-individual τ due to the quality of B_τ resulting from ε-REDORD. It can be done, e.g., as follows:

$$fitness(\tau) = 2^{-f(B_\tau)} \tag{7}$$

for $f : \mathcal{P}(A) \to \mathbf{R}$ defined at the end of Section 4. The following result, together with Proposition 1, assures that o-GA with fitness (7) can be applied to search for solutions of the ε-approximate μ-decision reduct optimization problems:

Proposition 2. *B is a solution of MεDRP, iff $B = B_\tau$ for τ maximizing (7), for $f(B_\tau) = |B_\tau|$. If B is a solution of MRεDRP, StεDRP, or SeεDRP, then $B = B_\tau$ for τ maximizing (7), for $f(B_\tau) = \left|V^U_{B_\tau}\right|$, $H(B_\tau)$, $H(B_\tau/d)$, respectively.*

In some cases, the obtained reduct B_τ, which maximizes $fitness(\tau)$, may be not the one with minimal cardinality. This is, however, very improbable, because smaller attribute subsets are obtained for larger families of permutations. It is an important advantage of the proposed hybrid algorithm.

References

1. Davis, L. (ed.): Handbook of Genetic Algorithms. Van Nostrand Reinhold (1991).
2. Duentsch, I., Gediga, G.: Uncertainty measures of rough set prediction. Artificial Intelligence **106** (1998) pp. 77–107.
3. Kapur, J., Kesavan, H.: Entropy Optimization Principles with Applications. Academic Press (1992).
4. Pawlak, Z.: Rough sets – Theoretical aspects of reasoning about data (1991).
5. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann (1988).
6. Provost, F., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In: Proc. of IMLC'98 (1998).
7. Rissanen J.: Minimum-description-length principle. In: S. Kotz, N.L. Johnson (eds), Encyclopedia of Statistical Sciences. Wiley (1985) pp. 523–527.
8. Ślęzak, D.: Approximate reducts in decision tables. In: Proc. of IPMU'96 (1996).
9. Ślęzak, D.: Approximate decision reducts (in Polish). Ph.D. thesis, Institute of Mathematics, Warsaw University (2001).
10. Ślęzak, D.: Approximate Entropy Reducts. Accepted to Fundamenta Informaticae.
11. Wróblewski, J.: Theoretical Foundations of Order-Based Genetic Algorithms. Fundamenta Informaticae **28/3-4** (1996) pp. 423–430.
12. Wróblewski, J.: Adaptive methods of object classification (in Polish). Ph.D. thesis, Institute of Mathematics, Warsaw University (2001).