Probabilistic Similarity-Based Reduct

Wojciech Froelich and Alicja Wakulicz-Deja

Institute of Computer Science, University of Silesia,
ul.Bedzinska 39, Sosnowiec, Poland
{Wojciech.Froelich,Alicja.Wakulicz-Deja}@us.edu.pl

Abstract. The attribute selection problem with respect to decision tables can be efficiently solved with the use of rough set theory. However, a known issue in standard rough set methodology is its inability to deal with probabilistic and similarity information about objects. This paper presents a novel type of reduct that takes into account this information. We argue that the approximate preservation of probability distributions and similarity of objects within reduced decision table helps to preserve the quality of its classification capability.

Keywords: rough set theory, probabilistic reduct, similarity.

1 Introduction

The information system [2] is an ordered pair $IS = (U, A)$, where U is a nonempty, final set of objects. The nonempty, final set A consists of labels which are the names of the attributes (i.e. features). An attribute $a \in A$ is a mapping $\mathrm{a : U \to V_a}$, where V_a is called a value set of the attribute a. For any subset of objects $X \subseteq U$, the B - indiscernibility relation: $xIND(B)y$ is defined as: $IND\left(B\right) = \left\{\left(u, u'\right) \in U^2 : \forall a \in B, a\left(u\right) = a\left(u'\right)\right\}$. The indiscernibility class determined by object u on the subset B of attributes can be denoted as $[u]_B$ and is called an elementary set E. Thus, U is partitioned into the family U/E of elementary sets. A decision table is defined as a special case of information system, $DT = (U, A, d)$, where $d \notin A$ is the decision attribute. The V_d is the set of values of d. Every value of decision d determines the decision class $D_k = \{u \in U : d\left(u\right) = k\}$. The family U/d is called the classification of objects in DT. It has been noted [4] that some formulas used in the decision-theoretic approach are in fact the estimators of the probability distributions within decision tables. Recently, a probabilistic interpretation of the variable precision rough set (VPRS) model was investigated [7]. A survey of the known types of reducts is given in [1].

A similarity between objects can be calculated by the function $sim : U \times U \to [0, 1]$. The properties of similarity are reflexivity $sim(x, x) = 1$ and symmetry $sim(x,\ y) =\ sim(y, x)$. For the attribute a with numeric values, similarity can be defined [5] as $sim_a(v_i, v_j) = 1 - \frac{|v_i - v_j|}{a_{max} - a_{min}}$, where a_{min}, a_{max} denote the minimum and maximum values of attribute a, respectively. For many attributes, it is possible to use [3]: $sim\left(x, y\right) = \frac{\sum_{a \in C} sim_a(x,y)}{card(C)}$, where $C \subseteq A$.

J.T. Yao et al. (Eds): RSKT 2011, LNCS 6954, pp. 610–615, 2011.

2 Assumptions

For the purpose of this paper, it is assumed that DT does not contain missing values and the values of attributes are discrete, numerical. To denote a particular subset of objects in DT, we give the value of the representative vector, e.g. $< -, 2 >$, where the sign "-" stands in place of the attribute that is not considered. The $A = \{A_1, A_2, \ldots, A_n\}$ is the set of random variables, where $n = card(A)$. The random variables are denoted using upper letters with the lower indexes identifying them within the set A. Lower letters are used to denote the values of variables. The expression $dom(A_i)$ denotes the domain of A_i. The $P(A_i = a_k)$ denotes the probability of assuming the value $a_k \in dom(A_i)$ by the variable A_i. Suppose we have a subset $B \subseteq A$, and b is a vector of values of corresponding random variables. The expression $b(u)$ denotes the vector of values that the object u takes on the set of random variables/attributes from B. We denote as $P(B = b)$ the probability of assuming by every variable from B the corresponding value from vector b. Suppose $D_1 \notin A$ is a distinguished random variable, we denote as $d \in dom(D_i)$ the constant value from its domain. The $P(D_1|A = a)$ is a vector of conditional probability distribution of D_1 given $A = a$. The matrix $P(D_1|A)$ stores a conditional probability distribution of D_1 for all values of A. Let us also notice that: $sim(x, y) = sim(a(x), a(y))$.

3 A New Type of Reduct

For every pair of decision tables $< DT, DT' >$, the probabilistic decision table PDT is defined. It possess the following attributes: L - a number of the row within PDT, $A = \{A_1, A_2, \ldots, A_n\}$ - the set of discrete random variables corresponding to the attributes of DT; the domain of the random variable A_i from PDT is the set of values of the corresponding attribute from DT, i.e.: $dom(A_i) = V_a$, D_1 - a distinguished discrete random variable corresponding to the decision attribute, $dom(D_1) = V_d$, $P(D_1|A)$ - the conditional probability of decision given the values of attributes from DT, $P(A)$ - the prior probability of assuming by A the combination of values given in the row of PDT, $P(D_1|B)$ - the conditional probability of decision given the values of attributes from DT', $P(B)$ - the prior probability of assuming by A the combination of values given in the row of PDT, $P(A|D_1)$ - the probabilistic distribution of indiscernibility classes existing in original DT within every decision class, SIM - the set of attributes $SIM = \{SIM(d_1), SIM(d_2), \ldots, SIM(d_m)\}$, each of which corresponds to the value of the decision attribute; i.e., for every $d_i \in dom(D_1)$, there is a corresponding real valued attribute $SIM(d_i) \in \mathbb{R}$ in PDT. Note that every indiscernibility class from DT is uniquely mapped to one record in PDT. Suppose that l_c is a row number in the PDT. We denote as $A[l_c]$ a vector stored in row l_c using the variables from the set A. Similarly, we use notation $B[l_c], C[l_c]$, where $C = A - B$.

3.1 Probabilistic Similarity-Based Classification

Suppose that for every decision d within $[u]_B$ a similarity-weighted probabilistic function $S(d, b) = \sum_{a \in dom(A)} P(D_1 = d|B = b) \cdot P(A = a|D_1 = d) \cdot simd(pd = d, pa = a, pb = b)$ is computed. The $P(D_1 = d|B = b)$ is the conditional probability of choosing the decision d given the values b of the attributes from DT'. The $P(A = a|D_1 = d)$ is the probability of asigning objects from indiscernibility class $A = a$ to the decision class d. Additionaly, all indiscernibility classes $A = a$ assigned in DT' to $D_1 = d$ are weighted by their similarity $simd(pd, pa, pb)$ to the most similar class from the set of these classes that were in DT originaly assigned to the given decision d. In order to calculate such specific similarity we propose the function $simd(pd, pa, pb)$. The pd is the parameter of one of the possible decisions and pa, pb are the parameters of the possible values of A and B, respectively. The values of $simd(pd, pa, pb)$ are stored as the values of attributes $SIM(d)$ of the PDT. For any given decision d and a vector $b(u)$ a subset $K \subset L$ of rows in PDT with $B = b(u)$ is selected. Then, it is checked whether the corresponding to K class $[u]_B$ is consistent. The function $simd$ works in two steps:

1. In the case of consistency of $[u]_B$, for every row $l_c \in K$:
   - If $D_1[l_c] = pd$, then $simd = 1$.
   - If $D_1\,[l_c] \neq pd$, then $simd = 0$.
2. In the case of inconsistency of $[u]_B$,
   (a) First, for all rows with $D_1\,[l_c] = pd$, the $simd = 1$.
   (b) Afterwards, for all the remaining in K rows (not considered in step 2a) for which $D_1[l_c] \neq pd$, we use $C = A - B$ and the similarity measure to compute the minimum value $s = \min sim(C[l_c\,], C[l_i])$, where l_c is a number of the current row in PDT and l_i any other row in K for which $D_1\,[l_i] \neq pd, A = pa, B = pb$ and $simd = 1$ (computed in the Step 2a). Then, we assign $simd = s$.

The classification of object u in table DT' can be performed by selecting the decision d' with the maximal value of $S(d, b(u))$ within the indiscernibility class represented by $b(u)$: $d'(u) = \arg max_d S(d, b(u))$.

3.2 Testing the Subsets of Attributes

After computing the dominant similarity-based decisions for every indiscernibility class, it is possible to calculate what part of objects is approximately correctly classified with the use of DT'. For this purpose, we use the function $ptest(DT, B)$ given as: $ptest\,(DT, B) = \sum_{b \in dom(B)} P(B = b) S(d', b)$, where d' is the probabilistic similarity-based dominant decision assigned for the given $B = b$. Note that the number of objects that are approximately correctly classified within DT' can be calculated as $ptest(DT, B) \cdot card(DT)$. The definition of negative, boundary and positive regions within DT is possible (similarily as in VPRSM) by the the introduction of appriopriate thresholds for the value of $ptest(DT, B)$.

3.3 The Formulation of Reduct

The probabilistic similarity-based reduct (PSBR) of the decision table DT is a subset of attributes $R \subset A$ that assures an approximate correct classification of at least $\eta \in [0, 1]$ fraction of objects calculated as $ptest\,(DT, R) \geq \eta$, where $\eta = \lambda \cdot ptest(DT, A)$. The above condition should be true for R and false for any proper subset of R. The $ptest(DT, A)$ is a fraction of objects that are approximate correct classified by DT with the use of all attributes A. For consistent DT the $ptest(DT, A) = 1$ and $\eta = \lambda$. The value of λ is the requirement given by expert and reflects in fact the degree in which the quality of classification is preserved.

4 An Illustrative Example

Due to the available page limit of this paper, it was possible solely to illustrate the calculation of function $ptest(DT, B)$, used for the evaluation of the subsets of attributes. Without doubt, the presented simple example cannot verify the practical effectiveness of the proposed reduct, the analysis of larger and complex decision tables is needed. Some of the theoretical strong points of our approach are given in section 5. A simple exemplary DT is given as Table 1(a). It is assumed that $A = \{a_1, a_2\}$. The DT is consistent. The reduced DT' is presented as Table 1(b) with the set of attributes $B = \{a_2\}$. The set $C = \{a_1\}$. The DT' is inconsistent. Bold lines mark the borders between indiscernibility classes. Suppose that we use the maximum distribution criterion for classification. In this case, the reduction of attributes leads to the incorrect classification of three objects from the original table: u_6, u_7 and u_8. In DT', these objects are assigned to decision $d = 1$ instead of $d = 0$ as in DT. Misclassified in the DT' objects are rare in the DT; e.g., $P(u_7) = 2/10$ in DT, by reduction of the attribute a_1, $P(u_7) = 8/10$ in DT'. This situation is better then incorporating frequent objects into the rare indiscernibility class. Note also that the object u_8 in DT is less similar than objects u_6 or u_7 to any (with $a_2 = 1$) of indiscernibility classes

Table 1. Decision table

(a) DT - original

U	a_1	a_2	d
u_1	0	1	1
u_2	0	1	1
u_3	0	1	1
u_4	0	1	1
u_5	1	1	1
u_6	2	1	0
u_7	2	1	0
u_8	5	1	0
u_9	1	0	0
u_{10}	1	0	0

(b) DT' - reduced

U	a_2	d
u_1	1	1
u_2	1	1
u_3	1	1
u_4	1	1
u_5	1	1
u_6	1	0
u_7	1	0
u_8	1	0
u_9	0	0
u_{10}	0	0

that exist in DT and that are merged in DT'. In the context of preserving the original similarity relations between objects (such as they were in DT), the incorrect classification of u_6 and u_7 in DT' can be evaluated as better than the misclassification of u_8. For the previously presented DT and DT', we constructed the PDT given in Table 2.

As an example, we show how to calculate $simd\,(pd, pa, pb)$ for the row l_3 of PDT. The following parameters are used: $\min(A_1) = 0, \max(A_1) = 5$, $\max(A_1) - \min(A_1) = 5$. For the row l_3 of PDT, and $d = 1$ we have $pd = 1, pa =< 2, 1 >$, $pb =< 1 >$. The most similar to l_3 row with $D_1 = 1$ is the row l_2. The similarity between l_3 and l_2 is $simd(1, < 2, 1 >, < 1 >) = 1 - ((2 - 1)/5) = 1 - 1/5 = 0.8$. This value is stored as $S(d = 1)[l_3]$. For $d = 0$ and l_3 we have: $SIM(d = 0)[l_3] = simd(0, < 2, 1 >, < 1 >)$; in this case we have $D_1 = d$, therefore $simd = 1$. After computing in PDT all values for columns: $SIM(d = 1)$ and $SIM(d = 0)$, we compute the dominant-similarity based decisions for every indiscernibility class in DT'. The class $< -, 1 >$ in DT' is inconsistent, so we calculate: $S(d = 1, B =< 1 >) = \frac{5}{8} \cdot \frac{4}{5} \cdot 1 + \frac{5}{8} \cdot \frac{1}{5} \cdot 1 + \frac{3}{8} \cdot \frac{2}{3} \cdot 0.8 + \frac{3}{8} \cdot \frac{1}{3} \cdot 0.2 = 0.85$ $S(d = 0, B =< 1 >) = \frac{5}{8} \cdot \frac{4}{5} \cdot 0.6 + \frac{5}{8\cdot} \frac{1}{5} \cdot 0.8 + \frac{3}{8} \cdot \frac{2}{3} \cdot 1 + \frac{3}{8} \cdot \frac{1}{3} \cdot 1 = 0.775$ The dominant decision for the class $< -, 1 >$ is $d' = 1$. For the DT' we have: $ptest(DT, B) = 0.8 \cdot (\frac{5}{8}\frac{4}{5} \cdot 1 + \frac{5}{8}\frac{1}{5} \cdot 1 + \frac{3}{8}\frac{2}{3} \cdot 0.8 + \frac{3}{8}\frac{1}{3} \cdot 0.2\ ) + 0.2 \cdot 1 = 0.68 + 0.2 = 0.88$

Table 2. *PDT* - probabilistic decision table

L	A_1	A_2	D_1	$P(D_1\|A)$	$P(A)$	$P(D_1\|B)$	$P(B)$	$P(A\|D_1)$	$SIM(d = 1)$	$SIM(d = 0)$
l_1	0	1	1	1	0.4	5/8	0.8	4/5	1	0.6
l_2	1	1	1	1	0.1	5/8	0.8	1/5	1	0.8
l_3	2	1	0	1	0.1	3/8	0.8	2/3	0.8	1
l_4	5	1	0	1	0.2	3/8	0.8	1/3	0.2	1
l_5	1	0	0	1	0.2	1	0.2	1	0	1

5 Relation to the Existing Approaches

To the best of our knowledge, all existing approaches to combining similarity with rough sets: a) use a similarity measure to construct similarity classes that partition U as U/S, and b) approximate the decision partition U/D using this similarity partition, i.e. cover U/D with U/S as much as possible. In contrast, in our approach the similarity classes on the global level of U are not constructed. Moreover, the similarity is applied with respect to the subsets of attributes and only for objects that changed the originally assigned decision after reduction of attributes. By this way, the similarity between objects depends on the subsets of reduced attributes. The maximum distribution reduct [6] preserves the dominant decision for every indiscernibility class and thus works locally. In our approach, it is possible that most objects from some rare indiscernibility classes will be classified incorrectly in DT' and the maximum distribution criterion does not hold. The advantage is that the other objects in frequent indiscernibility classes are better classified, thus our approach works globally. The parameter-based reducts preserve to certain extent the conditional probability distribution $P(D_1|B)$ and thus work globally. However, they do not take into account the probability distribution $P(A|B)$ that makes some indiscernibility classes from DT more or less important considering the number of correctly classified objects in DT'. This fact is considered in our approach.

6 Final Remarks

The proposed approach for reducing attributes of decision tables exploits the following types of information: 1) prior and conditional probability distributions regarding indiscernibility classes, 2) conditional probability distribution of decision classes given the indiscernibility classes, and 3) similarity between objects with respect to the reduction of attributes. The advantage of exploiting the above information should lead to the generation of better reducts, evaluated as it was shown within the paper, by the calculation of the fraction of objects correctly classified by the reduced decision table. Practical verification of the above claim requires to perform extensive experiments with real-world data, it is planned for future research. The limitation of the above presented approach is the prior assumption of the similarity function between objects.

References

1. Kryszkiewicz, M.: Comparative study of alternative types of knowledge reduction in inconsistent systems. Int. J. Intell. Syst. 16(1), 105–120 (2001)
2. Pawlak, Z.: Information systems – theoretical foundations. Information Systems 6, 205–218 (1981)
3. Shen, Q., Jensen, R.: Rough sets, their extensions and applications. International Journal of Automation and Computing 4, 217–228 (2007)
4. Slezak, D., Ziarko, W.: Attribute reduction in the bayesian version of variable precision rough set model. Electr. Notes Theor. Comput. Sci. 82(4) (2003)
5. Stefanowski, J., Tsoukiàs, A.: Induction of decision rules and classification in the valued tolerance approach. In: Rough Sets and Current Trends in Computing, pp. 271–278 (2002)
6. Zhang, W., Mi, J., Wu, W.: Approaches to knowledge reductions in inconsistent information systems. International Journal of Intelligent Systems (2003)
7. Ziarko, W.: Stochastic approach to rough set theory. In: Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Słowiński, R. (eds.) RSCTC 2006. LNCS (LNAI), vol. 4259, pp. 38–48. Springer, Heidelberg (2006)