Heuristic for Attribute Selection Using Belief Discernibility Matrix

Salsabil Trabelsi[1], Zied Elouedi[1], and Pawan Lingras[2]

[1] Larodec, Institut Superieur de Gestion de Tunis, Tunisia
[2] Saint Mary's University Halifax, Canada

Abstract. This paper proposes a new heuristic attribute selection method based on rough sets to remove the superfluous attributes from partially uncertain data. We handle uncertainty only in decision attributes (classes) under the belief function framework. The simplification of the uncertain decision table which is based on belief discernibility matrix generates more significant attributes with fewer computations without making significant sacrifices in classification accuracy.

Keywords: Uncertainty, belief function theory, rough sets, attribute selection, classification.

1 Introduction

A problem of relevant feature selection is one of the important problems in preprocessing stage of the modeling process in machine learning. There are several attempts to solve this problem based on rough set theory [5,14]. Using features from the reduct of the decision system adds to the efficiency of the classification process [8,9]. However, finding optimal reduct is an NP-hard problem. Researchers have proposed credible heuristics that compute acceptable reducts in reasonable time [19,20]. Another issue in real world database is the uncertainty, imprecision or incompleteness. There is some research that adapts exhaustive or heuristic feature selection methods based on rough sets to uncertain environment [16,17].

In this paper, we develop a new heuristic attribute selection method from partially uncertain data based on rough sets. The uncertainty exists only in the decision attribute and is represented by the belief function theory. It is considered as a useful theory for representing and managing total or partial uncertain knowledge because of its relative flexibility. The belief function theory is widely applied in artificial intelligence and to real life problems for decision making and classification. In this paper, we use the Transferable Belief Model (TBM) - one interpretation of belief function theory [13]. To remove the superfluous and the redundant attributes from the uncertain decision table, we adapt the concept of discernibility matrix in the new context to be used in the heuristic algorithm proposed originally in [12] to compute sufficient reduct without costly calculations.

T. Li et al. (Eds.): RSKT 2012, LNAI 7414, pp. 129–138, 2012.

This paper is organized as follows: Section 2 provides an overview of the rough set theory. Section 3 introduces the belief function theory as understood in the TBM. Section 4 describes the proposed heuristic for attribute selection based on rough sets under uncertainty. Section 5 details a belief rough set classifier a new classification system able to generate uncertain decision rules from partially uncertain data where the feature selection is one the important steps in the construction procedure. Section 6 reports the experimental results obtained from modified uncertain databases to evaluate the performance of our solution based on two evaluation criteria: the time requirement and the classification accuracy.

2 Rough Set Theory

In this section, we recall some basic notions related to information systems and rough sets [6,7]. An information system is a pair $A = (\mathrm{U}, \mathrm{C})$, where U is a non-empty, finite set called the *universe* and C is a non-empty, finite set of attributes. We also consider a special case of information systems called decision tables. A decision table is an information system of the form $A = (\mathrm{U}, \mathrm{C} \cup \{\mathrm{d}\})$, where $d \notin C$ is a distinguished attribute called *decision*. In this paper, the notation $c_i(o_j)$ is used to represent the value of a condition attribute $c_i \in C$ for $o_j \in U$.

For every set of attributes $B \subseteq C$, an equivalence relation denoted by IND_B and called the B-indiscernibility relation, is defined by

$$IND_B = U/B = \{[o_j]_B | o_j \in U\} \tag{1}$$

Where

$$[o_j]_B = \{o_i | \forall c \in B\, c(o_i) = c(o_j)\} \tag{2}$$

Let $B \subseteq C$ and $X \subseteq U$. We can approximate X by constructing the $B - lower$ and $B - upper\ approximations$ of X, denoted $\underline{\mathrm{B}}(\mathrm{X})$ and $\bar{\mathrm{B}}(\mathrm{X})$, respectively, where

$$\underline{\mathrm{B}}(X) = \{o_j | [o_j]_B \subseteq X\}\ and\ \bar{\mathrm{B}}(X) = \{o_j | [o_j]_B \cap X \neq \emptyset\} \tag{3}$$

A reduct is a minimal subset of attributes from C that preserves the positive region and the ability to perform classifications as the entire attributes set C. A subset $B \subseteq C$ is a reduct of C with respect to d, iff B is minimal and:

$$Pos_B(\{d\}) = Pos_C(\{d\}) \tag{4}$$

Where $Pos_C(\{d\})$, called a positive region of the partition $U/\{d\}$ with respect to C.

$$Pos_C(\{d\}) = \bigcup_{X \in U/\{d\}} \underline{\mathrm{C}}(X) \tag{5}$$

The core is the most important subset of attributes, it is included in every reduct.

$$Core(A, \{d\}) = \bigcap RED(A, \{d\}) \tag{6}$$

Where $RED(A, \{d\})$ is the set of all reducts of A relative to d.

Let's not that finding a minimal reduct (reduct with a minimal number of attributes) among all reducts is NP-hard. This means that computing reducts is not a trivial task. Fortunately, there exist good heuristics that compute sufficiently many reducts in often acceptable time [19,20].

3 Belief Function Theory

The belief function theory is proposed by Shafer [10] as a useful tool to represent uncertain knowledge. Here, we introduce only some basic notations related to the TBM [13], one interpretation of the belief function theory. Let Θ, frame of discernment, be a finite set of exhaustive elements to a given problem. All the subsets of Θ belong to the power set of Θ, denoted by 2^{Θ}. The bba (basic belief assignment) is a function representing the impact of a piece of evidence on the subsets of the frame of discernment Θ and is defined as follows:

$$m : 2^{\Theta} \rightarrow [0, 1]$$

$$\sum_{E \subseteq \Theta} m(E) = 1 \tag{7}$$

Where *m(E)*, named a basic belief mass (bbm), shows the part of belief exactly committed to the element E. The bba's induced from distinct pieces of evidence are combined by the cojunctive rule of combination [11].

$$(m_1 \textcircled{\cap} m_2)(E) = \sum_{F,G \subseteq \Theta : F \cap G = E} m_1(F) \times m_2(G) \tag{8}$$

To make decisions in the TBM, belief functions can be represented by probability functions called the pignistic probabilities denoted $BetP$ and are defined as [11]:

$$BetP(\{a\}) = \sum_{F \subseteq \Theta} \frac{\mid \{a\} \cap F \mid}{\mid F \mid} \frac{m(F)}{(1 - m(\emptyset))} \text{ for all } \; a \in \Theta \tag{9}$$

4 Heuristic for Attribute Selection Method Using Belief Discernibility Matrix

In this section, a new heuristic for simplification of partially uncertain decision system is proposed. We will remove the superfluous and redundant attributes for rules discovery without costly computations. We will keep only the features from the reduct. It is a minimal set of attributes that preserves the ability to classify as much as the entire set of condition attributes. First, we will give an overview of the proposed approach followed by experimental verification.

4.1 Uncertain Decision Table

Our uncertain decision system denoted A contains n objects $\{o_1, o_2, \ldots, o_n\}$, characterized by a set of certain condition attributes C=$\{c_1$, c_2,...,$c_k\}$ and uncertain decision attribute ud. We propose to represent the uncertainty of each object by a bba m_j expressing belief on decision defined on the frame of discernment Θ=$\{ud_1$, ud_2,...,$ud_s\}$ representing the possible values of ud.

Example: Let us use Table 1 to describe our uncertain decision system. It contains eight objects, three certain condition attributes C=$\{$Headache, Muscle-pain, Temperature$\}$ and an uncertain decision attribute ud=Flu with possible value {yes, no} representing Θ. For example, for the patient o_4, belief of 0.6 is exactly committed to the decision ud_1=yes, whereas belief of 0.4 is assigned to the entire frame of discernment Θ (ignorance).

Table 1. Uncertain decision table

U	Headache	Muscle-pain	Temperature	Flu
o_1	yes	yes	very high	$m_1(yes) = 1$
o_2	yes	yes	high	$m_2(yes) = 1$
o_3	yes	no	high	$m_3(yes) = 0.95$ $m_3(\Theta) = 0.05$
o_4	no	yes	normal	$m_4(no) = 0.6$ $m_4(\Theta) = 0.4$
o_5	no	yes	normal	$m_5(no) = 1$
o_6	no	no	high	$m_6(no) = 1$
o_7	yes	no	normal	$m_7(no) = 1$

4.2 Feature Selection with Johnson´s Heuristic Algorithm

This part describes the Johnson´s heuristic algorithm [12] to compute reducts. It sequentially selects features by finding those that are most discernible for a given decision feature. It computes a discernibility matrix M, where each cell $M_{i,j}$ of the matrix corresponding to the set of all condition attributes which discern objects o_i and o_j that do not belong to the same equivalence classes based on decision attribute d. The discernibility matrix is a symmetric n*n matrix with entries $M_{i,j}$ as given below.

$$M_{i,j} = \{c \in C | c(o_i) \neq c(o_j) \; for \, d(o_i) \neq d(o_j)\} \, \forall i, j = 1, ..., n \tag{10}$$

Given such a matrix M, for each feature, the algorithm counts the number of cells in which it appears. The feature c with the highest number of entries is selected for addition to the reduct R. Then, all the entries $M_{i,j}$ that contain c are removed and the next best feature is selected. This procedure is repeated until M is empty.

Johnson´s Reduct (U,C,{d})
Input: U: objects, C:conditional features, d: decisional features,
Output: R:reduct, R $\subseteq$ C
1. R $\leftarrow \emptyset$
2. M $\leftarrow$ ComputeDiscernibilityMatrix (U,C,{d})
3. do
4. c $\leftarrow$ SelectHighestScoringFeature (M)
5. R $\leftarrow$ R $\cup$ {c}
6. for (i=0 to $|U|$,j=0 to $|U|$)
7. $M_{i,j}$=$\emptyset$ **if** c $\in M_{i,j}$
8. C $\leftarrow$ C $\setminus$ {c}
9. Until $M_{i,j}$ =$\emptyset$ $\forall$ i,j
10. Return R

4.3 Belief Discernibility Matrix

In order to apply the previous heuristic attribute selection method to our uncertain decision table, we should adapt the discernibility matrix under the belief function framework which is originally based on certain decision attribute to be called belief discernibility matrix (M'). The instruction (2.) in the previous algorithm will be changed as follows:

2. M' $\leftarrow$ ComputeBeliefDiscernibilityMatrix (U,C,{ud})

In this case, the belief discernibility matrix will be based on a distance measure to identify the similarity or dissimilarity between decision values of the objects o_i and o_j. The idea is to use the distance measure between two bba's m_i and m_j. The threshold value is used to be more flexible. Hence, belief discernibility matrix M' is defined as follows:

$$M'_{i,j} = \{c \in C | c(o_i) \neq c(o_j) \; for \, dist(m_i, m_j) \geq threshold\} \, \forall i, j = 1, ..., n \quad (11)$$

Where *dist* is a distance measure between two bba's proposed in [2] which satisfies more properties than many other distance measures proposed in [1,3,4] as defined below.

$$dist(m_1, m_2) = \sqrt{\frac{1}{2}(\| \vec{m_1} \|^2 + \| \vec{m_2} \|^2 - 2 < \vec{m_1}, \vec{m_2} >)} \quad (12)$$

$$0 \leq dist(m_1, m_2) \leq 1 \quad (13)$$

Where $< \vec{m_1}, \vec{m_2} >$ is the scalar product defined by:

$$< \vec{m_1}, \vec{m_2} > = \sum_{i=1}^{|2^\Theta|} \sum_{j=1}^{|2^\Theta|} m_1(A_i) m_2(A_j) \frac{|A_i \cap A_j|}{|A_i \cup A_j|} \quad (14)$$

with A_i, $A_j \in 2^\Theta$ for $i, j = 1, 2, \cdots, |2^\Theta|$. $\| \vec{m_1} \|^2$ is then the square norm of $\vec{m_1}$.

Example: To apply our heuristic feature selection method to the uncertain decision system (see Table 1), we start by computing the belief discernibility matrix (see Table 2). We will use the notations H, M and T respectively for Headache, Muscle-pain and Temperature. To obtain Table 2, we have used Equation 11 with a threshold value equal to 0.1. For example, $M'_{1,3} = \emptyset$ because the two objects o_1 and o_3 have $dist(m_1, m_3) = 0.07 \ngtr 0.1$. The decision values of the two objects are considered similar. Next, we compute the reduct according to the Johnson´s heuristic algorithm. First, we find that the attribute Temperature has the highest number of entries which is equal to 11. So, we add Temperature to the reduct. Then, we remove all the cells $M'_{i,j}$ containing T. We still have only H and H, M in the matrix. So, the attribute Headache has now the highest number of entries which is equal to 2. We add the attribute Headache to the reduct. Then, we remove all the cells $M'_{i,j}$ containing H. The matrix is now empty and the process is finished.

Table 2. Belief disecrnibiliy matrix

U	o_1	o_2	o_3	o_4	o_5	o_6	o_7
o_1							
o_2							
o_3							
o_4	H,T	H,T	H,M,T				
o_5	H,T	H,T	H,M,T				
o_6	H,M,T	H,M	H	M,T			
o_7	M,T	M,T	T	H,M			

5 Belief Rough Set Classifier

Belief rough set classifier is a new classification technique proposed originally in [18]. The latter was able to learn belief decision rules for the classification process from partially uncertain data (see Table 1).

The decision rules induced from the uncertain decision table are called belief decision rules where the decision is represented by a bba.

Example: *The belief decision rule relative the object o_3 from the Table 1 is as follows: If $Headache = yes$ and $Muscle - pain = no$ and $Temperature = high$ Then $m_3(yes) = 0.95 \quad m_3(\Theta) = 0.05$.*

To create a belief rough set classifier, we need to simplify the uncertain decision system to generate the more significant belief decision rules by means of the following steps:

1. **Step 1. Eliminate the superfluous condition attributes:** We remove the superfluous condition attributes that are not in reduct. We can apply in this step our proposed heuristic for feature selection.

2. **Step 2. Eliminate the redundant objects:** After removing the superfluous condition attributes, we will find redundant objects. They may not have the same bba on decision attribute. So, we use their combined bbas using a rule of combination.
3. **Step 3. Eliminate the superfluous condition attribute values:** In this step, we compute the reduct value for each belief decision rule R_j of the form: **If** $C(o_j)$ **then** m_j.
4. **Step 4. Generate belief decision rules:** After the simplification of the uncertain decision table, we can generate shorter and significant belief decision rules. With simplification, we can improve the time and the performance of classification of unseen objects.

Once the belief rough set classifier is constructed, the following procedure will be the classification of unseen instances. Our method is able to ensure the standard classification where each attribute value of the new instance to classify is assumed to be exact and certain. We search among all belief decision rules which one corresponds to the unseen object. The new instance's decision will be defined by a bba. In order to make a decision and to get the probability of each singular decision, we apply the pignistic transformation using eqn. (9). We can take only the most probable decision.

6 Experimental Results

In our experiments, several tests were performed on real-world databases obtained from the U.C.I. repository[1] to evaluate the proposed heuristic feature selection method in comparison with exhaustive search proposed originally in [17]. A brief description of the databases is presented in Table 3.

The comparison is based on two evaluation criteria: the time requirement (the number of seconds needed to find the reduct) and the classification accuracy (Percent of Correct Classification (PCC)). To compute PCC, we apply our two methods in the first step of the belief rough set classifier described in Section 5. The belief rough set classifier generated decision rules that were used for classification. These databases were artificially modified in order to include uncertainty in decision attribute. We took different degrees of uncertainty based on increasing values of probabilities P used to transform the actual decision value d_i of each object o_j to a bba $m_j(\{d_i\}) = 1 - P$ and $m_j(\Theta) = P$. A larger P gives a larger degree of uncertainty. Each database is divided into ten parts. Nine parts are used as the training set, the last is used as the testing set. The procedure is repeated ten times, each time another part is chosen as the testing set. This method, called a cross-validation, permits a more reliable estimation of the evaluation criterion. In this paper, we perform ten-fold cross-validation tests with different data splits and we report the average of the evaluation criteria.

[1] `http://www.ics.uci.edu/~mlearn/MLRepository.html`

Table 3. Description of databases

Databases	#instances	#attributes	#decision values
W. Breast Cancer	690	8	2
Balance Scale	625	4	3
C. Voting records	497	16	2
Zoo	101	17	7
Nursery	12960	8	3
Solar Flares	1389	10	2
Lung Cancer	32	56	3
Hyes-Roth	160	5	3
Car Evaluation	1728	6	4
Lymphography	148	18	4
Spect Heart	267	22	2
Tic-Tac-Toe Endgame	958	9	2

Table 4. Experimentation results

Databases	Exhaustive Mean Time (s)	Heuristic Mean Time (s)	Gain in Time (s)	Gain in Time (%)	Exhaustive Mean PCC (%)	Heuristic Mean PCC (%)	Loss of PCC (%)
W. Breast Cancer	154	74	80	52%	86.60	85.46	1.14
Balance Scale	129	51	78	60%	83.43	83.13	0.3
C. Voting records	110	73	37	34%	98.78	98.10	0.68
Zoo	101	46	55	54%	96.41	95.70	0.72
Nursery	380	209	171	45%	96.49	95.65	0.84
Solar Flares	157	113	44	28%	88.69	88.18	0.51
Lung Cancer	48	37	11	23%	75.91	75.69	0.21
Hyes-Roth	91	46	45	49%	97.46	97.23	0.23
Car Evaluation	178	149	29	16%	84.40	83.84	0.56
Lymphography	102	79	23	23%	83.20	82.01	0.95
Spect Heart	109	87	22	20%	85.40	84.70	0.69
Tic-Tac-Toe Endgame	139	109	30	22%	86.37	85.95	0.42
Mean	141	89	52	37%	88.59	85.95	0.6

In Table 4, we report the average time requirement and the classification accuracy for different degrees of uncertainty. From this table, we see that the proposed heuristic feature selection method is faster than the exhaustive search method for attribute selection. It is true for all the databases (see Figure 2). For example, the time requirement for Balance Scale database goes from 129 seconds to 51 seconds. Gain of speed is equal to 78 seconds or 60%. The average gain in speed relative to all databases is 52 seconds or 37%. This gain in computational speed came at very little loss of accuracy (see Figure 1). For example, the classification accuracy for Lung Cancer database goes from 75.91% to 75.69%. The loss of accuracy is equal to 0.21%. The average loss of accuracy relative to all databases is also very small and is equal to (0.6%).

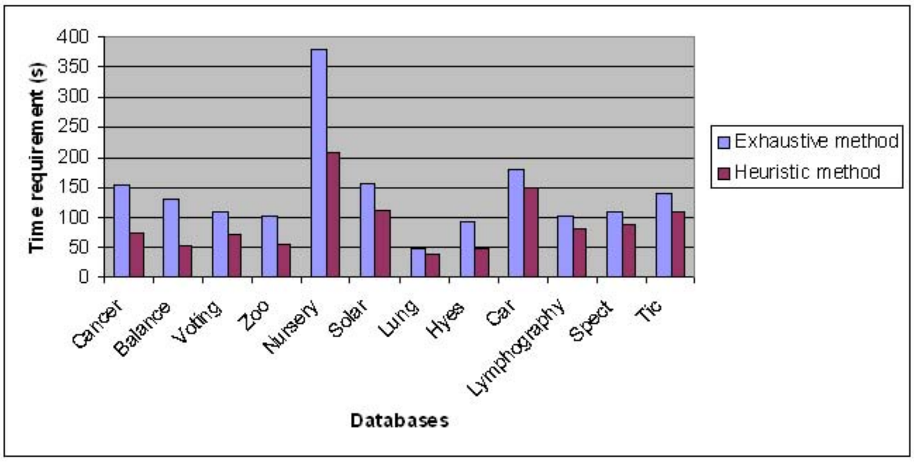

Fig. 1. Classification accuracy for exhaustive and heuristic methods

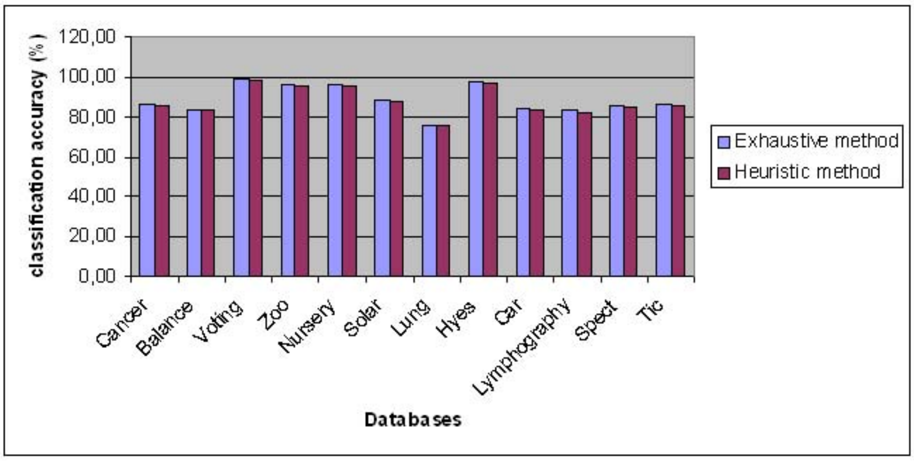

Fig. 2. Time requirement for exhaustive and heuristic methods

7 Conclusion and Future Work

In this paper, we have proposed a heuristic attribute selection method to remove the superfluous attributes from uncertain decision table in less time with minimal loss of classification accuracy. We handle uncertainty in decision attributes using the belief function. The proposed heuristic method was compared with the exhaustive search for optimal set of reduced features. The average gain in computational speed for twelve standard databases was 37%, while the loss of classification accuracy was mere 0.6%.

References

1. Bauer, M.: Approximations algorithm and decision making in the Dempster-Shafer theory of evidence - an empirical study. IJAR 17(2-3), 217–237 (1997)
2. Bosse, E., Jousseleme, A.L., Grenier, D.: A new distance between two bodies of evidence. Information Fusion 2, 91–101 (2001)
3. Elouedi, Z., Mellouli, K., Smets, P.: Assessing sensor reliability for multisensor data fusion within the transferable belief model. IEEE Trans. Syst. Man Cyben. 34(1), 782–787 (2004)
4. Fixen, D., Mahler, R.P.S.: The modified Dempster-Shafer approach to classification. IEEE Trans. Syst. Man Cybern. 27(1), 96–104 (1997)
5. Modrzejewski, M.: Feature selection using rough sets theory. In: Proceedings of the 11th International Conference on Machine Learning, pp. 213–226 (1993)
6. Pawlak, Z.: Rough Sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
7. Pawlak, Z., Zdzislaw, A.: Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishing, Dordrecht (1991) ISBN 0-7923-1472-7
8. Pawlak, Z., Rauszer, C.M.: Dependency of attributes in Information systems. Bull. Polish Acad. Sci., Math. 33, 551–559 (1985)
9. Rauszer, C.M.: Reducts in Information systems. Fundamenta Informaticae (1990)
10. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton (1976)
11. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence 66, 191–236 (1994)
12. Johnson, D.S.: Approximation algorithms for combinatorial problems. Journal of Computer and System Sciences 9, 256–278 (1974)
13. Smets, P.: The transferable belief model for quantified belief representation. In: Gabbay, D.M., Smets, P. (eds.) Handbook of Defeasible Reasoning and Uncertainty Management Systems, vol. 1, pp. 207–301. Kluwer, Doordrecht (1998)
14. Tanaka, H., Ishibuchi, H., Matuda, N.: Reduction of information system based on rough sets and its application to fuzzy expert system. Advancement of Fuzzy Theory and Systems in china and Japan (1990)
15. Tessem, B.: Approximations for efficient computation in the theory of evidence. Artif. Intell 61(2), 315–329 (1993)
16. Trabelsi, S., Elouedi, Z.: Heuristic method for attribute selection from partially uncertain data using rough sets. International Journal of General Systems 39(3), 271–290 (2010)
17. Trabelsi, S., Elouedi, Z.: Attribute selection from partially uncertain data using rough sets. In: The Third International Conference on Modeling Simulation, and Applied Optimization, UAE, January 20-22 (2009)
18. Trabelsi, S., Elouedi, Z., Lingras, P.: Belief Rough Set Classifier. In: Gao, Y., Japkowicz, N. (eds.) AI 2009. LNCS (LNAI), vol. 5549, pp. 257–261. Springer, Heidelberg (2009)
19. Wroblewski, J.: Finding minimal reducts using genetic algorithms. In: Proceedings of the 2nd Annual Joint Conference on Information Sciences, pp. 186–189 (1995)
20. Zhong, N., Dong, J.Z., Ohsuga, S.: Using Rough Sets with Heuristics for Feature Selection. Journal of Intelligent Information Systems 16(3), 199–214 (2001)