Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.

This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

International Journal of Approximate Reasoning 52 (2011) 1409–1432

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Classification systems based on rough sets under the belief function framework

Salsabil Trabelsi [a,*], Zied Elouedi [a], Pawan Lingras [b]

[a] LARODEC, University of Tunis, Institut Supérieur de Gestion de Tunis, Tunisia
[b] Saint Mary's University, Halifax, Canada

ARTICLE INFO

Article history:
Received 29 July 2010
Revised 4 August 2011
Accepted 5 August 2011
Available online 22 August 2011

Keywords:
Rough sets
Generalization distribution table
Belief functions
Uncertainty
Classification

ABSTRACT

In this paper, we present two classification approaches based on *Rough Sets* (RS) that are able to learn decision rules from uncertain data. We assume that the uncertainty exists only in the decision attribute values of the *Decision Table* (DT) and is represented by the belief functions. The first technique, named *Belief Rough Set Classifier* (BRSC), is based only on the basic concepts of the *Rough Sets* (RS). The second, called *Belief Rough Set Classifier*, is more sophisticated. It is based on *Generalization Distribution Table* (BRSC-GDT), which is a hybridization of the *Generalization Distribution Table* and the *Rough Sets* (GDT-RS). The two classifiers aim at simplifying the *Uncertain Decision Table* (UDT) in order to generate significant decision rules for classification process. Furthermore, to improve the time complexity of the construction procedure of the two classifiers, we apply a heuristic method of attribute selection based on rough sets. To evaluate the performance of each classification approach, we carry experiments on a number of standard real-world databases by artificially introducing uncertainty in the decision attribute values. In addition, we test our classifiers on a naturally uncertain web usage database. We compare our belief rough set classifiers with traditional classification methods only for the certain case. Besides, we compare the results relative to the uncertain case with those given by another similar classifier, called the *Belief Decision Tree* (BDT), which also deals with uncertain decision attribute values.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

In recent years, we have witnessed a growing interest in rough sets (RS) [23,24] and its applications. This theory has attracted attention of many researchers whose contributions have further enhanced the rough set theory. It constitutes a sound basis for decision support system and data mining. It offers solutions to the problem of discretization [32], attribute selection and decision rule generation [20,40]. Classification approach based on rough sets, called *Rough Set Classifier* (RSC), has been successfully applied in several real-world applications [1,20]. This classifier performs feature selection before generating rules. It is an efficient technique that tries to automatically produce a minimal and significant set of decision rules without many iterations. A RSC has the advantages of time complexity for learning, accuracy and size of the discovered rules. However, it is not true for other classification techniques such as *Decision Tree* [25] which needs pruning to improve the size and the accuracy of the models especially for large and uncertain databases. However, classification techniques based on standard rough sets do not perform their tasks in an environment characterized by uncertain or incomplete data. Many researchers have extended rough sets and its applications to accommodate uncertainty [14,17–19]. These extensions deal with incomplete decision tables which may be characterized by missing condition attribute values (each condition attribute is considered as one property to describe an instance or an object), but not with uncertain decision attribute. The decision attribute specifies what action should be undertaken when some conditions are satisfied. Uncertainty in the

* Corresponding author.
E-mail addresses: salsabilt@yahoo.fr (S. Trabelsi), zied.elouedi@gmx.fr (Z. Elouedi), pawan.lingras@gmail.com (P. Lingras).

0888-613X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2011.08.002

Author's personal copy

condition attribute exists in many real-world applications such as marketing, finance, management and medicine. For the latter, the diseases (classes) of some patients may be totally or partially uncertain. It is not useful to eliminate these objects from classification process because it represents a loss of important information. To overcome this shortcoming, we present two classification approaches the BRSC [35] and the BRSC-GDT [36] which are able to learn simplified decision rules from uncertain data with an objective to classify unseen objects. The uncertainty appears only in decision attributes. This kind of uncertainty can be represented by belief functions as interpreted in the *Transferable Belief Model* (TBM) – one interpretation of the belief function theory [29]. The TBM, which enables flexible representation of partial or total ignorance, has been shown to handle uncertainty in classification problems [3,6,8,11].

The *Belief Rough Set Classifier* (BRSC) is based only on the basic concepts of the rough sets (RS).

The *Belief Rough Set Classifier based on Generalization Distribution Table* (BRSC-GDT) is derived from a soft induction system called GDT-RS [9,41], which is a hybridization of the *Generalization Distribution Table* and the *Rough Set* methodology. The two classifiers aim at simplifying the *Uncertain Decision Table* (UDT) in order to generate more significant decision rules for the classification process. To improve the time complexity of the construction procedure for our classifiers, we apply a heuristic method for attribute selection also based on rough sets proposed originally in [37] as a pre-processing stage. To evaluate the performance of each classification approach, we carry experiments on standard real-world databases. The databases are artificially modified to include uncertainty in the decision attribute values. To further judge the performance and the feasibility of our two belief classification techniques, we have performed experimentation on a naturally uncertain database. This dataset was obtained from web access logs of the introductory computing science course at Saint Mary's University. Instead of using crisp assignment of a visit to one of the three clusters, a study patterns [22] associated a basic belief assignment (bba). The resulting uncertain clustering was represented using belief functions. Note that this latter database has been used in our previous work to judge the performance of our classifiers based on dynamic approach [38]. The performance criteria used in this paper for the experiments are: the classification accuracy, the size of models and the time requirement of learning. In addition, we compare our belief classification systems based on rough sets with traditional classification methods such as standard decision tree [25], K-nearest neighbors [5] and Naive Bayes classifier [16] for crisp conventional decision attributes. We also compare the results relative to the uncertain case with those obtained from a similar classifier called the *Belief Decision Tree* (BDT) [8,11,39]. The latter is a decision tree in an uncertain environment where the uncertainty is represented through the TBM and appears in the decision attribute of training objects. Inducing a BDT may lead in most cases to very large trees with bad classification accuracy and difficult comprehension. As in standard decision trees within BDT, pre-pruning [7,12] and post-pruning [33] methods have been proposed to cope with this problem.

This paper is organized as follows: Section 2 provides an overview of the basic concepts of *Rough Sets* (RS), *Generalization Distribution Table* (GDT) and the hybrid system GDT-RS. Section 3 introduces the belief function theory as understood in the *Transferable BeliefModel* (TBM). Section 4 describes our two classification approaches the BRSC and the BRSC-GDT which are based on rough sets under the belief function framework to handle the problem of uncertainty. Furthermore, we present a heuristic method of attribute selection to improve the time complexity of the construction procedure. Finally, in Section 5, we carry experiments on real world databases to evaluate our two classifiers using three evaluation criteria: the classification accuracy, the size and the time requirement.

2. Rough sets (RS), Generalization Distribution Table (GDT) and the hybrid system GDT-RS

The idea of rough sets has been introduced by Pawlak [23,24] to deal with imprecise and vague concepts. It constitutes a sound basis for *Knowledge Discovery Database* (KDD). It can be used for feature selection, discretization, data reduction and decision rule generation, etc. This theory is not competitive but complementary to other methods and can be often used jointly with other approaches [10,28]. Here, we introduce the basics of rough sets (RS), the Generalization Distribution Table (GDT) and the hybrid system GDT-RS that will be useful in this paper.

2.1. Rough sets (RS)

Let us recall some basic notions regarding rough sets [24]. A decision table (DT) is defined as A= (U, C, {d}), where $U = \{o_1, o_2, \ldots, o_n\}$ is a nonempty finite set of n objects called *the universe*, $C = \{c_1, c_2, \ldots, c_k\}$ is a finite set of k *condition* attributes and $d \notin C$ is a distinguished attribute called *decision*. The value set of d, called Θ is $\{d_1, d_2, \ldots, d_s\}$. In this paper, the notation $c_i(o_j)$ is used to represent the value of a condition attribute $c_i \in C$ for an object $o_j \in U$. We further extend these notations for a set of attributes $B \subseteq C$, by defining $B(o_j)$ to be value tuple of attributes in B for an object o_j.

2.1.1. Indiscernibility relation

Indiscernible objects may be represented several times in a DT. The objects o_i and o_j are indiscernible on a subset of attributes $B \subseteq C$, if they have the same values for each attribute in the subset B of C. Rough sets adopt the concept of indiscernibility relation [23,24] to partition the object set U into disjoint subsets (equivalence classes), denoted by U/B or Ind_B, and the equivalence class that includes o_j is denoted by $[o_j]_B$

Author's personal copy

$$[o_j]_B = \{o_i | B(o_i) = B(o_j)\} \tag{1}$$

$$Ind_B = U/B = \{[o_j]_B | o_j \in U\} \tag{2}$$

The equivalence classes based on the decision attribute is denoted by $U/\{d\}$

$$Ind_{\{d\}} = U/\{d\} = \{[o_j]_{\{d\}} | o_j \in U\} \tag{3}$$

2.1.2. Set approximation

The concept of indiscernibility relation is a natural way of reducing data, since only one element of the equivalence class is needed to represent the entire class. Subsets that are most often of interest have the same value of the outcome attribute. It may happen that a target concept cannot be defined in a crisp manner. It is here that the notion of rough sets emerges.

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

$$\underline{B}(X) = \{o_j | [o_j]_B \subseteq X\} \quad and \quad \bar{B}(X) = \{o_j | [o_j]_B \cap X \neq \emptyset\} \tag{4}$$

Objects in $\underline{B}(X)$ can be classified with certainty as members of X on the basis of knowledge in B, while objects in $\bar{B}(X)$ can only be classified as possible members of X on the basis of knowledge in B. The set $BN_B(X)$ is called the B – *boundary region* of X, and thus consists of those objects that we cannot decisively classify into X on the basis of knowledge in B:

$$BN_B(X) = \bar{B}(X) - \underline{B}(X) \tag{5}$$

A set is said to be rough if the boundary region is non-empty.

2.1.3. Dependency of attributes and positive region

Another important issue in data analysis is discovering dependencies between attributes [24]. The attribute d depends on the set of attributes C to a degree k $(0 \leq k \leq 1)$, denoted $C \Rightarrow_k \{d\}$, if

$$k = \gamma_C(\{d\}) = \frac{|Pos_C(\{d\})|}{|U|} \tag{6}$$

where

$$Pos_C(\{d\}) = \bigcup_{X \in U/\{d\}} \underline{C}(X) \tag{7}$$

$Pos_C(\{d\})$, called a positive region of the partition $U/\{d\}$ with respect to C, is the set of all elements of U that can be uniquely classified to blocks of the partition $U/\{d\}$, by means of C.

2.1.4. Reduct and core

Another way of data reduction is to keep only those attributes that preserve the positive region. There are usually several such subsets of attributes and those which are minimal are called reducts [24]. A subset $B \subseteq C$ is a reduct of C with respect to d, iff B is minimal and:

$$Pos_B(\{d\}) = Pos_C(\{d\}) \tag{8}$$

In other words, attributes that do not belong to a reduct are superfluous with regard to classification of elements of the universe. The set of all the condition attributes indispensable in the decision table (system) A with respect to d is denoted by $Core_C(\{d\})$

$$Core_C(\{d\}) = \bigcap Red_C(\{d\}) \tag{9}$$

where $Red_C(\{d\})$ is the set of all reducts of the decision system A.

Since the core is the intersection of all reducts, it is included in every reduct. Thus, in a sense, the core is the most important subset of attributes, for none of its elements can be removed without affecting the classification power of attributes.

2.1.5. Value reduct and value core

To further simplify a decision table, we can eliminate some values of an attribute from the table. We will also need a concept of value reduct and value core.

A subset $B \subseteq C$ is a value reduct of C for o_j, iff B is *minimal* for o_j and $[o_j]_C \subseteq [o_j]_{\{d\}}$ implies $[o_j]_B \subseteq [o_j]_{\{d\}}$. The set of all indispensable attributes values from C for o_j with respect to d will be called value core and will be denoted by $Core^j_C(\{d\})$ having the following property:

Author's personal copy

Table 1
Decision table.

U	Hair	Eyes	Height	d
o_1	Dark	Brown	Middle	d_1
o_2	Blond	Blue	Middle	d_1
o_3	Blond	Brown	Tall	d_1
o_4	Dark	Brown	Short	d_2
o_5	Blond	Blue	Middle	d_2
o_6	Dark	Brown	Tall	d_1

$$Core^j_C(\{d\}) = \bigcap Red^j_C(\{d\}) \tag{10}$$

where $Red^j_C(\{d\})$ is the family of all value reducts of C for o_j with respect to d.

Example: A small example of DT can be found in Table 1. The table includes six objects, three condition attributes (*Hair*, *Eyes*, *Height*) and one decision attribute (d) with two possible outcomes: d_1, d_2.

We may notice that objects o_2 and o_5 have exactly the same values of conditions, but they have a different outcome (different value of the decision attribute).

The relation *IND* defines three partitions of the universe.

$$IND_{\{Hair\}} = \{\{o_2, o_3, o_5\}, \{o_1, o_4, o_6\}\}$$
$$IND_{\{Hair,Eyes\}} = \{\{o_1, o_4, o_6, \}, \{o_2, o_5\}, \{o_3\}\}$$
$$IND_{\{Hair,Eyes,Height\}} = \{\{o_1\}, \{o_2, o_5\}, \{o_3\}, \{o_4\}, \{o_6\}\}$$

Let $X = \{o_j | d(o_j) = d_1\}$. In fact, the set X consists of four objects: o_1, o_2, o_3 and o_6. Now, we want to describe this set in terms of conditional attributes $C = \{Hair,\ Eyes,\ Height\}$. Using the above definitions, we obtain the following approximation regions: the $C - lower\ approximation\ \underline{C}(X) = \{o_1, o_3, o_6\}$, the $C - upper\ approximation\ \bar{C}(X) = \{o_1, o_2, o_3, o_5, o_6\}$, the $C - boundary\ region\ BN_C(X) = \{o_2, o_5\}$ and the $C - outside\ region\ U - \bar{C}(X) = \{o_4\}$. It is easy to see that the set X is rough since the boundary region is not empty.

For dependency $\{Hair, Eyes, Height\} \Rightarrow \{d\}$ we get $k = 4/6 = 2/3$, because four out of six objects can be uniquely classified as having the decision d_1 or d_2, employing attributes *Hair*, *Eyes* and *Height*, $Pos_C(A,\ \{d\}) = \{o_1, o_3, o_4, o_6\}$.

In Table 1, there are two possible reducts with respect to d, $\{Eyes, Height\}$ and $\{Hair, Height\}$, since they have the same positive region as the whole set of condition attributes C. That means that either the attribute *Eyes* or *Hair* can be eliminated from the table. The core is the attribute *Height*. It is the intersection of the two possible reducts.

$$Pos_C(A, \{d\}) = \{o_1, o_3, o_4, o_6\}$$
$$Pos_{\{Hair\}}(A, \{d\}) = \emptyset$$
$$Pos_{\{Eyes\}}(A, \{d\}) = \emptyset$$
$$Pos_{\{Height\}}(A, \{d\}) = \{o_3, o_4, o_6\}$$
$$Pos_{\{Hair,Eyes\}}(A, \{d\}) = \{o_3\}$$
$$Pos_{\{Hair,Height\}}(A, \{d\}) = \{o_1, o_3, o_4, o_6\}$$
$$Pos_{\{Eyes,Height\}}(A, \{d\}) = \{o_1, o_3, o_4, o_6\}$$

2.2. Generalization Distribution Table (GDT)

The Generalization Distribution Table (GDT) proposed by Zhong et al. in [42] is a hypothesis search space for generalization, in which the probabilistic relationship between concepts and instances over discrete domains are represented. It consists of three components:

(1) Possible instances (PI), represented at the top row of *GDT*, are defined by all possible combinations of attribute values from a database. The number of possible instances is $\prod_{i=1}^{m} n_i$, where m is the number of attributes, and n_i is the number of different attribute values in attribute i.
(2) Possible generalizations of instances (PG), represented by the left column of a *GDT*, are all possible cases of generalization for all possible instances. A wild card '*' denotes the generalization for instances. For example the generalization $*b_0c_1$ means that the attribute a is unimportant for describing a concept. The number of the possible generalization is $\prod_{i=1}^{m}(n_i + 1) - \prod_{i=1}^{m}(n_i - 1)$.

Author's personal copy

Table 2
Generalization distribution table.

PG/PI	$H_0E_0T_0$	$H_0E_0T_1$	$H_0E_0T_2$	$H_0E_1T_0$	$H_0E_1T_1$	$H_0E_1T_2$	$H_1E_0T_0$	...
$*E_0T_0$	1/2						1/2	
$*E_0T_1$		1/2						...
$*E_0T_2$			1/2					...
$*E_1T_0$				1/2				...
$*E_1T_1$					1/2			...
$*E_1T_2$						1/2		...
H_0*T_0	1/2			1/2				...
H_0*T_1		1/2			1/2			...
H_0*T_2			1/2			1/2		...
H_1*T_0							1/2	...
H_1*T_1								...
H_1*T_2								...
$H_0E_0{}^*$	1/3	1/3	1/3					...
$H_0E_1{}^*$				1/3	1/3	1/3		...
$H_1E_0{}^*$							1/3	...
$H_1E_1{}^*$								...
$**T_0$	1/4		1/4		1/4		1/4	...
$**T_1$		1/4		1/4		1/4		...
$**T_2$			1/4		1/4		1/4	...
$*E_0*$	1/6	1/6	1/6				1/6	...
$*E_1*$				1/6	1/6	1/6		...
H_0**	1/6	1/6	1/6	1/6	1/6	1/6		...
H_1**							1/6	...

(3) Probabilistic relationships between possible instances (PI) and possible generalizations (PG), represented by entries G_{ij} of a given *GDT*, are defined by means of a probabilistic distribution describing the strength of the relationship between any possible instance and any possible generalization. The prior distribution is assumed to be uniform if any prior background knowledge is not available. Certainly, we have $\sum_j G_{ij} = 1$, for any i. Thus, it is also defined by:

$$G_{ij} = p(PI_j|PG_i) = \begin{cases} \frac{1}{N_{PG_i}} & \text{if } PG_i \text{ is a generalization of } PI_j \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

where PI_j is the jth possible instance, PG_i is the ith possible generalization, and N_{PG_i} is the number of the possible instances satisfying the ith possible generalization, that is,

$$N_{PG_i} = \prod_{k \in \{l|PG_i[l]=*\}} n_k \tag{12}$$

where $PG_i[l]$ is the value of the kth attribute in the possible generalization PG_i, $PG_i[l] = *$ means that the PG_i does not contain attribute l. n_k is the number of values of the kth attribute.

Example: Table 2 describes a GDT generated from the decision table shown in Table 1. To simplify the notations in the GDT, we use H_0 to mean $Hair = Blond$ and H_1 to mean $Hair = Dark$. We also use T_0, T_1 and T_2 to mean respectively the different values of the attribute $Height = \{Short, Middle, Tall\}$.

2.3. Hybrid system GDT-RS

The GDT-RS, a soft hybrid induction system, is based on a hybridization of the *Generalization Distribution Table* (GDT) and the *Rough Sets* (RS). It is proposed for discovering classification rules from databases with noisy data [9,41]. The GDT-RS system can generate a set of decision rules with the minimal description length, having large strength and covering all instances. From the decision table (DT), we can generate decision rules expressed in the following form:

$P \rightarrow Q$ with S

- P is a conjunction of descriptors over C.
- Q denotes a concept that the rule describes.
- S is a 'measure of the strength' of the rule.

According to the GDT-RS, the strength S is equal to [9,41]:

$$S(P \rightarrow Q) = s(P) * (1 - r(P \rightarrow Q)) \tag{13}$$

Author's personal copy

where $s(P)$ is the strength of the generalization P (the condition of the rule) and r is the noise rate function. In other words, the strength of a given rule reflects incompleteness and noise.

The strength of the generalization P, $s(P)$, is defined as the sum of prior distributions between P and the observed instances satisfying P. It shows that within possible instances satisfying the generalization PG, how many of them are observed in the database. The initial value of $s(P)$ is 0. The value will be dynamically updated according to the real data in a database. If all the possible instances satisfying generalization P occur, the strength will be the maximal value, 1.

Let $P = PG$, the strength of the generalization PG, when the prior distribution is uniform, is given by:

$$s(P) = \sum_l p(PI_l|PG) = N_{ins-rel}(PG)\frac{1}{N_{PG}} \tag{14}$$

where $N_{ins-rel}(PG)$ is the number of the observed instances satisfying the generalization (PG) in the DT and N_{PG} is the number of the possible instances satisfying the generalization P which is computed using Eq. (12).

In other words, the strength of the generalization P represents explicitly the prediction for unseen instances.

On the other hand, the noise rate is given by:

$$r(P \to Q) = \frac{N_{ins-rel}(P) - N_{ins-class}(P, Q)}{N_{ins-rel}(P)} \tag{15}$$

where $N_{ins-rel}(P)$ is the number of the observed instances satisfying the generalization P, and $N_{ins-class}(P, Q)$ is the number of instances belonging to the decision Q within the instances satisfying the generalization P. It shows the quality of classification measured by the number of the instances satisfying the generalization P which cannot be classified into class Q. Furthermore, a user can specify an allowed noise rate as threshold value. Thus, the rules with larger rates than the threshold value will be deleted.

Example: We compute the strength of the decision rule described as follows: "If $Hair = Blond$ and $Eyes = Brown$ then $d = d_1$", where $Hair = Blond$ and $Eyes = Brown$ corresponds to P and $d = d_1$ corresponds to Q. P also corresponds to the possible generalization $PG = H_0E_0*$ in the GDT shown in Table 2. According to it, $\frac{1}{N_{PG}}$ is equal to $\frac{1}{3}$ and $N_{ins-rel}(PG)$ is equal to 1 according to Table 1. So, $s(P) = 1 * \frac{1}{3} = \frac{1}{3}$.

According to Table 1, $r(P \to Q) = \frac{1-1}{1} = 0$. Hence, $S(P \to Q)$=$\frac{1}{3} * (1 - 0) = \frac{1}{3}$.

3. Belief function theory

In this section, we briefly review the main concepts underlying the belief function theory as interpreted in the *Transferable Belief Model* (TBM) [29,30]. The latter is a useful model to represent quantified belief functions.

3.1. Basic concepts

Let Θ be a finite set of elementary events to a given problem, called the frame of discernment. All the subsets of Θ belong to the power set of Θ, denoted by 2^{Θ}.

The impact of a piece of evidence on the different subsets of the frame of discernment Θ is represented by a basic belief assignment (bba).

The bba is a function $m : 2^{\Theta} \to [0, 1]$ such that:

$$\sum_{E \subseteq \Theta} m(E) = 1 \tag{16}$$

The value $m(E)$, named a basic belief mass (bbm), represents the portion of belief committed exactly to the event E.

Associated with m is the belief function, denoted *bel*, corresponding to a specific bba m. The belief function *bel* assigns to every subset E of Θ the sum of masses of belief committed to every subset of E by m [26]. Contrary to the bba which expresses only the part of belief that one commits to E without being also committed to $\bar{E}$.

Author's personal copy

The belief function *bel* is defined for $E \subseteq \Theta,\ E \neq \emptyset$ as:

$$bel(E) = \sum_{\emptyset \neq F \subseteq E} m(F) \tag{17}$$

The plausibility function *pl* quantifies the maximum amount of belief that could be given to a subset E of the frame of discernment. It is equal to the sum of the bbm's for subsets F compatible with E.

The plausibility function *pl* is defined as follows:

$$pl(E) = \sum_{E \cap F \neq \emptyset} m(F), \quad \forall E \subseteq \Theta \tag{18}$$

Example: Let us treat the problem of identification 'Who murdered John?' and we have three suspects. Thus, the frame of discernment related to this problem is defined as follows:

$$\Theta = \{Henry, Peter, Sara\}$$

The power set of Θ is:

$$2^{\Theta} = \{\emptyset, \{Henry\}, \{Peter\}, \{Sara\}, \{Henry, Peter\}, \{Peter, Sara\}, \{Henry, Sara\}, \{Henry, Peter, Sara\}\}$$

The bba related to a piece of evidence concerning the murderer of John is defined as follows:

$$m(\{Henry\}) = 0.1; \quad m(\{Henry, Peter\}) = 0.7; \quad m(\{Sara\}) = 0.2$$

For example, 0.1 represents the part of belief exactly supporting that the murderer is *Henry*.
The belief function *bel* corresponding to this bba m is defined as follows:

$$bel(\emptyset) = 0; \quad bel(\{Henry\}) = 0.1; \quad bel(\{Peter\}) = 0; \quad bel(\{Sara\}) = 0.2$$
$$bel(\{Peter, Sara\}) = 0.7; \quad bel(\{Henry, Peter\}) = 0.1 + 0.7 = 0.8$$
$$bel(\{Henry, Sara\}) = 0.1 + 0.2 = 0.3; \quad bel(\Theta) = 0.1 + 0.7 + 0.2 = 1$$

For example, 0.8 is the total belief committed to the proposition $\{Henry, Peter\}$.

The plausibility function *pl* corresponding to the bba m is defined as follows:

$$pl(\emptyset) = 0; \quad pl(\{Henry\}) = 0.1 + 0.7 = 0.8; \quad pl(\{Peter\}) = 0.7; \quad pl(\{Sara\}) = 0.2$$
$$pl(\{Henry, Peter\}) = 0.1 + 0.7 = 0.8; \quad pl(\{Henry, Sara\}) = 0.1 + 0.7 + 0.2 = 1$$
$$pl(\{Peter, Sara\}) = 0.7 + 0.2 = 0.9; \quad pl(\Theta) = 0.1 + 0.7 + 0.2 = 1$$

For example, 0.2 represents the maximum degree of belief that the proposition $\{Sara\}$ may have.

3.2. Combination

Handling information induced from different experts (information sources) requires an evidence gathering process in order to get the fused information. In the TBM, the basic belief assignments induced from distinct pieces of evidence can be combined using many combination rules such as the conjunctive rule and the average rule of combination.

(1) The conjunctive rule: When we know that both sources of information are fully reliable, then the bba representing the combined evidence satisfies [31]:

$$(m_1 \textcircled{\cap} m_2)(E) = \sum_{F,G \subseteq \Theta : F \cap G = E} m_1(F) m_2(G) \tag{19}$$

(2) The average rule: This combination rule is more suitable when evidence conflicts, then the bba representing the combined evidence satisfies [21]:

$$Average(m_1, m_2)(E) = \frac{1}{2}[m_1(E) + m_2(E)] \tag{20}$$

Example: Assume $\Theta = \{Henry, Peter, Sara\}$.

Author's personal copy

Let two bba's m_1 and m_2 relative to two pieces of evidence.

$$m_1(\{Henry\}) = 0.6; \quad m_1(\{Henry, Peter\}) = 0.2; \quad m_1 = (\Theta) = 0.2$$
$$m_2 = (\{Peter\}) = 0.4; \quad m_2 = (\{Henry, Peter\}) = 0.3; \quad m_2 = (\Theta) = 0.3$$

Applying the conjunctive rule of combination, we get:

$$(m_1 \textcircled{\cap} m_2)(\emptyset) = 0.24; \quad (m_1 \textcircled{\cap} m_2)(\{Henry\}) = 0.36; \quad (m_1 \textcircled{\cap} m_2)(\{Peter\}) = 0.16$$
$$(m_1 \textcircled{\cap} m_2)(\{Henry, Peter\}) = 0.18; \quad (m_1 \textcircled{\cap} m_2)(\Theta) = 0.06$$

Applying the average rule of combination, we get:

$$Average(m_1, m_2)(\{Henry\}) = 0.3; \quad Average(m_1, m_2)(\{Peter\}) = 0.2$$
$$Average(m_1, m_2)(\{Henry, Peter\}) = 0.25; \quad Average(m_1, m_2)(\Theta) = 0.25$$

3.3. Decision making

In the TBM, holding beliefs and making decisions are distinct processes. Hence, it proposes two levels:

- The credal level where beliefs are represented by belief functions.
- The pignistic level where beliefs are used to make decisions and represented by probability functions called pignistic probabilities denoted *BetP* and defined as follows [30]:

$$BetP(\{a\}) = \sum_{F \subseteq \Theta} \frac{|\{a\} \cap F|}{|F|} \frac{m(F)}{(1 - m(\emptyset))}, \quad \text{for all } a \in \Theta \tag{21}$$

Example: To make a decision, we have to compute the pignistic probability BetP corresponding to the bba, m:

$$BetP(\{Henry\}) = 0.45; \quad BetP(\{Peter\}) = 0.35; \quad BetP(\{Sara\}) = 0.2;$$

It is more probable that the murderer is *Henry*.

4. Classification systems based on rough sets under the belief function framework

In this section, we present our two classification approaches, in an uncertain context, based on rough sets. The first classification technique, called the *Belief Rough Set Classifier* (BRSC) [35], is based only on the basic concepts of the rough sets (RS) especially reduct and value reduct. The second method is based on the hybrid induction system GDT-RS [9,41] and is called *Belief Rough Set Classifier* based on *Generalization Distribution Table* (BRSC-GDT) [36]. These two classifiers aim at simplifying the uncertain decision table (UDT) and generating more significant decision rules to classify unseen objects. Please note that the uncertainty exists in the decision attribute and is handled through the TBM. To avoid the costly calculation of the reduct, we suggest a heuristic method for attribute selection based on rough sets in an uncertain context in the pre-processing stage of the both methods.

4.1. Basic concepts of rough sets under uncertainty

In order to extend the Rough Set Classifier (RSC) under the belief function framework, we need to redefine the basic concepts of rough sets in this new context. This subsection describes the modified definitions of decision table, indiscernibility relation, set approximations, positive region and dependency of attributes. These adaptations were proposed originally in [34].

4.1.1. Uncertain decision table (UDT)

Our UDT is given by $A = (U, C \cup \{ud\})$, where $U = \{o_j : 1 \leq j \leq n\}$ is characterized by a set of certain condition attributes $C = \{c_1, c_2, \ldots, c_k\}$, and an uncertain decision attribute ud. We represent the uncertainty of each object o_j by a bba m_j expressing beliefs on decisions defined on the frame of discernment $\Theta = \{ud_1, ud_2, \ldots, ud_s\}$ representing the possible values of ud. These bba's are generally given by an expert or several experts. They can also present the two extreme cases of total knowledge and total ignorance.

Example: Let us use Table 3 to describe our UDT. It contains eight objects, three certain condition attributes $C = \{Hair, Eyes, Height\}$ and an uncertain decision attribute ud with two possible values $\{ud_1, ud_2\}$ representing Θ. For example, for the

Author's personal copy

Table 3
Uncertain decision table.

U	Hair	Eyes	Height	ud
o_1	Dark	Brown	Short	$m_1(\{ud_2\}) = 0.5, m_1(\Theta) = 0.5$
o_2	Blond	Blue	Middle	$m_2(\{ud_2\}) = 1$
o_3	Blond	Brown	Short	$m_3(\{ud_1\}) = 0.7, m_3(\Theta) = 0.3$
o_4	Blond	Brown	Tall	$m_4(\{ud_1\}) = 0.95, m_4(\{ud_2\}) = 0.05$
o_5	Dark	Brown	Short	$m_5(\{ud_2\}) = 1$
o_6	Blond	Blue	Middle	$m_6(\{ud_2\}) = 0.95, m_6(\Theta) = 0.05$
o_7	Dark	Brown	Tall	$m_7(\{ud_1\}) = 1$
o_8	Dark	Brown	Middle	$m_8(\{ud_1\}) = 0.975, m_8(\Theta) = 0.025$

object o_3, 0.7 of beliefs are exactly committed to the decision ud_1, whereas 0.3 of beliefs is assigned to the whole of frame of discernment Θ (ignorance). Through a bba, we can also represent the certain case, like for the objects o_2, o_5 and o_7.

4.1.2. Tolerance relation

The indiscernibility relation U/C for a subset of condition attributes is the same as in the certain case because their values are certain. However, the indiscernibility relation for the decision attribute $U/\{ud\}$ is not the same as in the certain case. The decision value is represented by a bba. In our case, it will be called *Tolerance relation* $Tol_{\{ud\}}$. So, we need to assign each object characterized by a bba m_j to the right tolerance class X_i for a decision value ud_i. The idea is to use the distance between the two bba's m_j and a certain bba m (such that $m(\{ud_i\}) = 1$). Many distance measures between two bba's were developed. Some of them are based on pignistic transformation [2,13,43]. This kind of distances may lose information given by the initial bba's. However, the distance measures developed in [4,15] are directly defined on bba's. In our case, we choose the distance measure proposed in [4] which satisfies more properties such as non-negativity, non-degeneracy and symmetry. For every ud_i, we define a tolerance class as follows:

$$X_i = \{o_j | dist(m, m_j) < 1 - threshold\}, \ such\ that\ m(\{ud_i\}) = 1 \tag{22}$$

Besides, we define a tolerance relation as follows:

$$Tol_{\{ud\}} = U/\{ud\} = \{X_i | ud_i \in \Theta\} \tag{23}$$

where *dist* is a distance measure between two bba's.

$$dist(m_1, m_2) = \sqrt{\frac{1}{2}(\| \vec{m_1} \|^2 + \| \vec{m_2} \|^2 - 2\langle \vec{m_1}, \vec{m_2} \rangle)} \tag{24}$$

where $\langle \vec{m_1}, \vec{m_2} \rangle$ is the scalar product defined by:

$$\langle \vec{m_1}, \vec{m_2} \rangle = \sum_{i=1}^{|2^\Theta|} \sum_{j=1}^{|2^\Theta|} m_1(A_i) m_2(A_j) \frac{|A_i \cap A_j|}{|A_i \cup A_j|} \tag{25}$$

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

It should be noted here that we replace the term equivalence class from the certain decision attribute case by tolerance class for the uncertain decision attribute, because the resulting classes may overlap.

Example: Let us continue with the same example to compute the equivalence classes based on condition attributes in the same manner as in the certain case: $Ind_C = U/C = \{\{o_1, o_5\}, \{o_4\}, \{o_2, o_6\}, \{o_3\}, \{o_7\}, \{o_8\}\}$ and to compute the tolerance classes based on the uncertain decision attribute $U/\{ud\}$. If the user fixed the value of the threshold at 0.1, we would obtain the following results:

For the uncertain decision value ud_1 with $m(\{ud_1\}) = 1$: $X_1 = \{o_1, o_3, o_4, o_7, o_8\}$.
For the uncertain decision value ud_2 with $m(\{ud_2\}) = 1$: $X_2 = \{o_1, o_2, o_3, o_5, o_6\}$.
So, the tolerance classes based on ud is equal to $Tol_{\{ud\}} = U/\{ud\} = \{\{o_1, o_3, o_4, o_7, o_8\}, \{o_1, o_2, o_3, o_5, o_6\}\}$.

4.1.3. Set approximation

To compute the new *lower* and *upper* approximations for our UDT, we should follow two steps:

(1) We combine the bba's for each equivalence class from U/C using the average rule [21] as follows:

$$\bar{m}_{[o_j]_C}(E) = \frac{1}{|[o_j]_C|} \sum_{o_i \in [o_j]_C} m_i(E), \quad for\ all\ E \subseteq \Theta \tag{26}$$

Author's personal copy

Table 4
The combined bba for the subsets $\{o_1, o_5\}$ and $\{o_2, o_6\}$.

Object	$m(\{ud_1\})$	$m(\{ud_2\})$	$m(\Theta)$
o_1	0	0.5	0.5
o_5	0	1	0
$m_{1,5}$	0	0.75	0.25
o_2	0	1	0
o_6	0	0.95	0.05
$m_{2,6}$	0	0.975	0.025

Let us remember that $[o_j]_C$ is the equivalence class containing the object o_j. Note that the average rule is more suitable than the rule of combination in Eq. (19) which is proposed especially to combine different beliefs on decision for one object and not different bba's relative to different objects.

(2) We compute the new *lower* and *upper* approximations for each tolerance class X_i from $U/\{ud\}$ based on uncertain decision attribute ud_i as follows:

$$\underline{C}X_i = \{o_j | [o_j]_C \subseteq X_i \text{ and } dist(m, \bar{m}_{[o_j]_C}) \leq threshold\} \tag{27}$$

$$\bar{C}X_i = \{o_j | [o_j]_C \cap X_i \neq \emptyset\} \tag{28}$$

In the new *lower* approximation, we find all equivalence classes from U/C included in X_i where the distance [4] between the combined bba $\bar{m}_{[o_j]_C}$ and the certain bba m (such that $m(\{ud_i\}) = 1$) is less than a *threshold*. However, the *upper* approximation is computed in the same manner as in the certain case.

Note that in the case of uncertainty, the *threshold* gives more flexibility to the calculation of tolerance class and set approximations. Threshold value is fixed by the user, and for the sake of consistency, it should be the same.

Example: We continue with the same example to compute the new *lower* and *upper* approximations. After the first step, we obtain the combined bba for each equivalence class from U/C using the average rule. Table 4 presents the combined bba for the equivalence classes $[o_1]_C$ and $[o_2]_C$. *Note that to simplify the notation, we have used* $m_{1,5}$ *to mean* $\bar{m}_{[o_1]_C}$.

Next, we compute the *lower* and *upper* approximations for each tolerance class from $U/\{ud\}$. Let us remember that the value of the *threshold* is equal to 0.1. For the uncertain decision value ud_1, let $X_1 = \{o_1, o_3, o_4, o_7, o_8\}$. The subsets $\{o_4\}, \{o_7\}$ and $\{o_8\}$ are included in X_1 and the distance between their bba and the certain bba m (such that $m(\{ud_1\}) = 1$) is less than 0.1. So, they will be in the lower. $\underline{C}(X_1) = \{o_4, o_7, o_8\}$, $\bar{C}(X_1) = \{o_1, o_3, o_4, o_5, o_7, o_8\}$ and $BN_C(X_1) = \{o_1, o_3, o_5\}$.

For uncertain decision value ud_2, let $X_2 = \{o_1, o_2, o_3, o_5, o_6\}$. Only the subset $\{o_2, o_6\}$ is included in X_2 and the distance between it's bba and the certain bba m (such that $m(\{ud_2\}) = 1$) is less than 0.1. So, we obtain: $\underline{C}(X_2) = \{o_2, o_6\}$, $\bar{C}(X_2) = \{o_1, o_2, o_3, o_5, o_6\}$ and $BN_C(X_2) = \{o_1, o_3, o_5\}$.

4.1.4. Positive region and dependency of attributes

With this new *lower* approximation, we can define the new positive region denoted $UPos_C(\{ud\})$:

$$UPos_C(\{ud\}) = \bigcup_{X_i \in U/\{ud\}} \underline{C}(X_i) \tag{29}$$

The attribute ud depends on the set of attributes C to a degree k, if

$$k = \gamma_C(\{ud\}) = \frac{|UPos_C(\{ud\})|}{|U|} \tag{30}$$

Example: Let us continue with the same example, to compute the positive region and dependency of attributes of A.

$$UPos_C(\{ud\}) = \{o_2, o_4, o_6, o_7, o_8\}$$
$$\gamma_C(\{ud\}) = \tfrac{5}{8}$$

4.1.5. Belief decision rules

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 for the object o_3 from Table 3 is as follows: If *Hair* = *Blond* and *Eyes* = *Brown* and *Height* = *Short* Then $m_3(\{ud_1\}) = 0.7 \quad m_3(\Theta) = 0.3$.

Author's personal copy

4.2. The Belief Rough Set Classifier (BRSC)

Our first classification approach based under the belief function frame work is called the *Belief Rough Set Classifier* (BRSC). This technique is based only on the basic concepts of rough sets such as reduct and value reduct. The construction procedure of the BRSC follows two main phases:

(1) Simplification of the uncertain decision table.
(2) Generation of belief decision rules.

These two phases are described in the next part.

4.2.1. Simplification of the uncertain decision table

The simplification of our uncertain decision table (UDT) is the most important phase in the construction procedure of the BRSC. It consists of removing all redundant and unnecessary data for rule discovery. It generates more significant and efficient belief decision rules needed for classification process. The simplification of the UDT is done as follows:

Step 1. Elimination of the superfluous condition attributes: We remove the superfluous condition attributes that are not in reducts. This leaves us with a minimal set of attributes that preserve the ability to perform the same classification as the original set of attributes. We need to redefine the concepts of reduct and core in this new situation as follows:

Using the new formalism of positive region, we can find the reduct of C as a minimal set of attributes $B \subseteq C$ such that:

$$UPos_B(\{ud\}) = UPos_C(\{ud\}) \tag{31}$$

The core is the most important subset of attributes given by the intersection of all the reducts.

$$UCore_C(\{ud\}) = \bigcap URed_C(\{ud\}) \tag{32}$$

where $URed_C(\{ud)\}$ is the set of all the reducts of A relative to ud.

Note that calculations of reducts tend to be computationally expensive. It is one of the drawbacks of rough set methodology. It will be better to apply a heuristic method for feature selection that computes approximate reducts with acceptable time requirements (see Section 4.4).

Example: Let us continue with the same example in Table 3 to compute the possible reducts using the new formalism of positive region.

$$UPos_{\{Hair\}}(\{ud\}) = \emptyset = UPos_C(\{ud\})$$
$$UPos_{\{Eyes\}}(\{ud\}) = \emptyset = UPos_C(\{ud\})$$
$$UPos_{\{Height\}}(\{ud\}) = \{o_1, o_7\} = UPos_C(\{ud\})$$
$$UPos_{\{Hair,Eyes\}}(\{ud\}) = \{o_8\} = UPos_C(\{ud\})$$
$$UPos_{\{Hair,Height\}}(\{ud\}) = \{o_2, o_4, o_6, o_7, o_8\} = UPos_C(\{ud\})$$
$$UPos_{\{Eyes,Height\}}(\{ud\}) = \{o_2, o_4, o_6, o_7, o_8\} = UPos_C(\{ud\})$$

There are two reducts: $\{Hair, Height\}$ and $\{Eyes, Height\}$. The attribute *Height* is the only core. We have two possible solutions to simplify our uncertain decision table. If we choose the second reduct $\{Eyes, Height\}$, the simplified uncertain decision table is shown in Table 5.

Table 5
The second reduct.

U	Eyes	Height	ud
o_1	Brown	Short	$m_1(\{ud_2\}) = 0.5, m_1(\Theta) = 0.5$
o_2	Blue	Middle	$m_2(\{ud_2\}) = 1$
o_3	Brown	Short	$m_3(\{ud_1\}) = 0.7, m_3(\Theta) = 0.3$
o_4	Brown	Tall	$m_4(\{ud_1\}) = 0.95, m_4(\{ud_2\}) = 0.05$
o_5	Brown	Short	$m_5(\{ud_2\}) = 1$
o_6	Blue	Middle	$m_6(\{ud_2\}) = 0.95, m_6(\Theta) = 0.05$
o_7	Brown	Tall	$m_7(\{ud_1\}) = 1$
o_8	Brown	Middle	$m_8(\{ud_1\}) = 0.975, m_8(\Theta) = 0.025$

Author's personal copy

Table 6
Redundant objects.

U	Eyes	Height
o_1, o_3, o_5	Brown	Short
o_2, o_6	Blue	Middle
o_4, o_7	Brown	Tall
o_8	Brown	Middle

Table 7
Combined bba's of redundant objects.

U	ud
o_1, o_3, o_5	$m_{1,3,5}(\{ud_1\}) = 0.24, m_{1,3,5}(\{ud_2\}) = 0.5, m_{1,3,5}(\Theta) = 0.26$
o_2, o_6	$m_{2,6}(\{ud_2\}) = 0.975, m_{2,6}(\Theta) = 0.025$
o_4, o_7	$m_{4,7}(\{ud_1\}) = 0.975, m_{4,7}(\{ud_2\}) = 0.025$
o_8	$m_8(\{ud_1\}) = 0.975, m_8(\Theta) = 0.025$

Table 8
Simplified uncertain decision table.

U	Eyes	Height	ud
o_1, o_3, o_5	Brown	Short	$m_{1,3,5}$
o_2, o_6	Blue		$m_{2,6}$
o_4, o_7		Tall	$m_{4,7}$
o_8	Brown	Middle	m_8

Step 2. Elimination of the redundant objects: After removing the superfluous condition attributes, we will find redundant objects having the same condition attribute values. They may not have the same bba on decision attribute. So, we use their combined bba's based on the average rule as a rule of combination as follows:

$$\bar{m}_{[o_j]_B}(E) = \frac{1}{|[o_j]_C|} \sum_{o_i \in [o_j]_B} m_i(E), \ \textit{for all } E \subseteq \Theta \tag{33}$$

where B is the reduct of C with respect to ud.

Example: After removing the superfluous condition attributes and the redundant objects for the uncertain decision table, we obtain Tables 6 and 7. Let us remember that we have used the notation $m_{1,3,5}$ to mean $\bar{m}_{[o_1]_B}$ with B equal to {Eyes, Height}.

Step 3. Elimination of the superfluous condition attribute values: To further simplify the uncertain decision table (UDT), we can eliminate some attribute values. We also need to redefine the concept of value reduct and value core. For each belief decision rule of the form: **If** $C(o_j)$ **then** m_j.

For all $B \subset C$, Let $X = \{o_k | B(o_j) = B(o_k) \ and \ j \neq k\}$
If $X = \emptyset$ **then** $B(o_j)$ is a value reduct of o_j.
Else If $Max(dist(m_j, \bar{m}_{[o_k]_C})) \leq$ threshold **then** $B(o_j)$ is a value reduct of o_j.

The set of all indispensable attributes values from C for o_j with respect to ud will be called value core and will be denoted by $UCore^j_C(\{ud\})$.

$$UCore^j_C(\{ud\}) = \bigcap URed^j_C(\{ud\}) \tag{34}$$

where $URed^j_C(\{ud\})$ is the family of all reduct value of C for o_j with respect to ud.

Example: We compute the value reduct of each decision rule from Table 7 with $threshold = 0.1$. We take one decision rule from the table: If $Eyes = Brown$ and $Height = Tall$ then $m_{4,7}$.

Let us split the rule into:

(1) If $Eyes = Brown$ then $m_{4,7} : X = \{o_1, o_3, o_5, o_8\}$ and $\text{Max}(\text{dist}(m_{4,7}, m_{1,3,5}), \text{dist}(m_1, m_8)) > 0.1$
(2) If $Height = Tall$ then $m_{4,7} : X = \emptyset$

So, $Height = Tall$ is the only value reduct for the first rule. If we compute the value reduct for all the belief decision rules, we obtain Table 8.

4.2.2. Generation of 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. From one UDT, we can find

Author's personal copy

many simplified solutions. The number of solutions, denoted *nb*, is computed as follows:

$$nb = |URed_C(\{ud\})| * \prod_{j=1}^{n} |URed_C^j(\{ud\})| \tag{35}$$

Example: Table 8 gives one solution of rule classification described as follows:

If *Eyes* = *Brown* and *Height* = *Short* then $m_{1,3,5}(\{ud_1\}) = 0.24,\ m_{1,3,5}(\{ud_2\}) = 0.5,\ m_{1,3,5}(\Theta) = 0.26$.
If *Eyes* = *Blue* then $m_{2,6}(\{ud_2\}) = 0.975,\ m_{2,6}(\Theta) = 0.025$.
If *Height* = *Tall* then $m_{4,7}(\{ud_1\}) = 0.975,\ m_{4,7}(\{ud_2\}) = 0.025$.
If *Eyes* = *Brown* and *Height* = *Middle* then $m_8(\{ud_1\}) = 0.975, m_8(\Theta) = 0.025$.

The generated belief decision rules are used to classify unseen objects. To simplify this task, we can compute the pignistic probability from m_j for each belief decision rule using Eq. (21).

The time complexity of the algorithm is $O(n * m^2 * Nr_{max})$, where n is the number of instances in a given database, m stands for the number of attributes, Nr_{max} is the maximal number of reducts. We suggest a heuristic method for attribute selection to avoid the costly calculation of the reducts (see Section 4.4).

4.3. The Belief Rough Set Classifier based on Generalization Distribution Table

In this subsection, we present our second approach for discovering a set of classification rules from uncertain data. Like the BRSC, the uncertainty appears only in decision attribute values and is handled by the TBM. This method is based on the hybrid system GDT-RS originally developed in [9,41]. Our solution namely BRSC-GDT [36] can generate from uncertain databases a set of rules with the minimal description length, having large strength and covering all instances. After the simplification of the uncertain decision table, the BRSC generates all possible decision rules. However, in the simplification of the uncertain decision tables, the BRSC-GDT eliminates the contradictory decision rules. Besides, if one decision rule has more than one possibility of simplification, it keeps only the decision rule with best strength. The main steps used to build BRSC-GDT and to discover the set of classification rules from UDT under the belief function framework are described as follows:

Step 1. Creation of the GDT: This step can be omitted because the prior distribution of a generalization can be calculated using eqns. (11) and (12). Note that the standard GDT depends only on condition attributes and not in decision attribute values. In the uncertain context, our GDT still has the same structure as in [42].

Step 2. Definition of the compound object: Objects from U having the same condition attribute values are considered as one object over C, called the compound object o'_j and defined as follows:

$$o'_j = \{o_i | C(o_i) = C(o_j)\} \tag{36}$$

For objects composing each compound object, combine their bba's using the average rule as follows:

$$m'_j(E) = \frac{1}{|o'_j|} \sum_{o_j \in o'_j} m_j(E), \forall E \subseteq \Theta \tag{37}$$

Let us remember that the average rule [21] is more suitable to combine these bba's than the rule of combination in Eq. (19) which is proposed especially to combine different beliefs on decision for one object and not different bba's for different objects.

The definition of the compound objects is a way of reducing the uncertain decision table.

Example: Let us continue with the same example in Table 3. By applying *the step 2*, we obtain the following tables:

Step 3. Elimination of the contradictory compound objects: For any compound object o'_j from U, compute $r_{ud_i}(o'_j)$ which represents a noise rate for each decision value ud_i, If there exists a ud_i such that $r_{ud_i}(o'_j) = min\{r_{ud_{i'}}(o'_j) | ud_{i'} \in \Theta\} < T_{noise}$ (threshold value), then we assign the decision value ud_i to the object o_j. If there is no $ud_i \in \Theta$ such that $r_{ud_i}(o'_j) < T_{noise}$, we treat the compound object o'_j as a contradictory one, and set the decision value of o'_j to $\perp$ (uncertain).

The noise rate is calculated originally using Eq. (15). This is not appropriate in our uncertain context since the decision value is represented by a bba. So, we propose the computing of the noise rate based on a distance measure [4] as follows:

$$r_{ud_i}(o'_j) = dist(m'_j, m), \text{ such that } m(\{ud_i\}) = 1 \tag{38}$$

Author's personal copy

Table 9
The compound objects.

U	Hair	Eyes	Height
$o'_1(o_1, o_5)$	Dark	Brown	Short
$o'_2(o_2, o_6)$	Blond	Blue	Middle
o'_3	Blond	Brown	Short
o'_4	Blond	Brown	Tall
o'_7	Dark	Brown	Tall
o'_8	Dark	Brown	Middle

Table 10
The combined bba's.

U	ud
o'_1	$m'_1(\{ud_2\}) = 0.75, m'_1(\Theta) = 0.25$
o'_2	$m'_2(\{ud_2\}) = 0.975, m'_2(\Theta) = 0.025$
o'_3	$m'_3(\{ud_2\}) = 0.7, m'_3(\Theta) = 0.3$
o'_4	$m'_4(\{ud_1\}) = 0.95, m'_4(\{ud_2\}) = 0.05$
o'_7	$m'_7(\{ud_1\}) = 1$
o'_8	$m'_8(\{ud_1\}) = 0.975, m'_8(\Theta) = 0.025$

Table 11
Contradictory and not contradictory compound objects.

U	Hair	Eyes	Height	ud
o'_1	Dark	Brown	Short	$\perp$
o'_2	Blond	Blue	Middle	ud_2
o'_3	Blond	Brown	Short	$\perp$
o'_4	Blond	Brown	Tall	ud_1
o'_7	Dark	Brown	Tall	ud_1
o'_8	Dark	Brown	Middle	ud_1

Table 12
Discernibility vector for o'_4.

U'	$o'_1(\perp)$	$o'_2(ud_2)$	$o'_3(\perp)$	$o'_4(ud_1)$	$o'_7(ud_1)$	$o'_8(ud_1)$
$o'_4(ud_1)$	Hair, Height	Eyes, Height	Height	$\emptyset$	$\emptyset$	$\emptyset$

The idea is to use the distance between two bba's m'_j and a certain bba m (such that $m(\{ud_i\}) = 1$). In this manner, we can check that the decisions of all instances in a compound object are close to a certain case. In such a case, the compound object is considered as not contradictory.

Elimination of the contradictory compound objects is another way of reducing the uncertain decision table.

Example: We fixed the threshold value of T_{noise} at 0.05. For the compound object o'_1, the noise rate for the decision value ud_1 is equal to $r_{ud_1}(o'_1) = 0.63$ and the noise rate for the decision value ud_2 is equal to $r_{ud_2}(o'_1) = 0.36$. So, o'_1 is a contradictory compound object. By applying the *step* 3 to Tables 9 and 10, we obtain Table 11.

Step 4. Minimal description length of decision rule: The value reduct of a decision rule for a compound object gives a minimal description length of the decision rule without a lose of information. Therefore, from U', we select the set of all the non-contradictory compound objects. For each non-contradictory compound object o'_j we then create a discernibility vector (the row or the column with respect to o'_j in the discernibility matrix [27]) for o'_j.

The discernibility matrix of A is a symmetric $n * n$ matrix. Each entry a_{ij} consists of the set of attributes upon which objects o_i and o_j differ.

$$a_{ij} = \{c \in C | c(o_i) \neq c(o_j)\} \ and \ ud(o_i) \neq ud(o_j) \ for \ i, j = 1, \ldots, n \tag{39}$$

Next, we compute all the value reducts for the compound object o'_j by using the discernibility function $f_A(o'_j)$. It is a boolean function of k boolean variables corresponding to the k condition attributes defined as below:

$$f_A(o'_j) = \bigwedge \left\{ \bigvee a_{ij} | 1 \leq i \leq n, a_{ij} \neq \emptyset \right\} \tag{40}$$

The set of all prime implicants of $f_A(o'_j)$ determines the sets of all reduct values of the compound object o'_j.

Example: According to Table 11, the discernibility vector for the compound object o'_4 is as follows:

Author's personal copy

Table 13
Decision rules.

U'	Rules	Strengths
o'_2	If $Eyes = Blue$ then ud_2	0.5
o'_4, o'_7	If $Height = Tall$ then ud_1	0.5
o'_8	If $Eyes = Brown$ and $Height = Middle$ then ud_1	0.5

We obtain from Table 12 three reduct values namely {Hair, Height}, {Eyes, Height} and {Height} by applying the indiscernibility function : $f_A(o'_4)$=$(Hair \vee Height) \wedge (Eyes \vee Height) \wedge (Height) = (Height) \vee (Hair \wedge Height) \vee (Eyes \wedge Height)$. So, we have three possibilities to simplify the decision rule for the compound object o'_4.

Step 5. Selection of the best rules: We start by constructing the simplified decision rules obtained from the value reducts for the each compound object o'_j from U', and revising their strength using Eq. (13). Then, we select only the best rule for o'_j having the maximum strength.

Example: The following rules are acquired for the compound object o'_4 according to the same example :

(1) If $Height = Tall$ then ud_1 with $S = 2 * \frac{1}{4} = 0.5$.
(2) If $Hair = Blond$ and $Height = Tall$ then ud_1 with $S = 1 * \frac{1}{2} = 0.5$.
(3) If $Eyes = Brown$ and $Height = Tall$ then ud_1 with $S = 1 * \frac{1}{2} = 0.5$.

The rule If $Height = Tall$ then ud_1 is selected for the compound object o'_4 due to its strength and description length.

Step 6. Stopping criterion: Let $U' = U' - \{o'_j\}$. *If* $U' \neq \emptyset$, then go back to *Step* 4. Otherwise, STOP.

Example: As a result, we obtain a set of decision rules able to classify unseen objects shown in Table 13.

The time complexity of the algorithm is $O(k * n^2 Nr_{max})$, where n is the number of instances in a given database, k stands for the number of attributes, Nr_{max} is the maximal number of reducts for instances. We again suggest a heuristic method for attribute selection in pre-processing stage before using our BRSC-GDT (see Section 4.4).

4.4. Heuristic method for attribute selection

Finding all possible reducts from uncertain decision table (UDT) is an NP-hard problem. To solve this problem, we have proposed a heuristic attribute selection method based on rough sets that is able to select the relevant features from our uncertain data without costly calculations [37]. The advantages of this heuristic is that it is fast and generates only one reduct. This heuristic can be applied in the pre-processing stage of the two classifiers BRSC and BRSC-GDT.

4.4.1. Principle

Our solution is based on the algorithm by Zhong et al. [40] which uses the attributes from the core as an initial attribute subset. Next, it selects attributes one by one from unselected ones using certain strategies, and adds them to the attribute subset until a reduct approximation is obtained. A threshold value is required as a stopping criterion to determine when a reduct candidate is "near enough" to being a reduct. On each iteration, those objects that are consistent with the current reduct candidate are removed.

4.4.2. Notations

The following notations are used to introduce the algorithm:

- R: the set of selected condition attributes,
- P: the set of unselected condition attributes,
- ε: reduct threshold.
- v_c : the cardinality of the positive region when the attribute c is added to the set of selected condition attributes R. In other words, v_c reflects the consistent objects that can be formed when the attribute c is added to the set of selected condition attributes R.
- m_c : the largest equivalence class obtained from the positive region when the attribute c is added to the set of selected condition attributes R and is based on condition and decision attributes $(R \cup \{c\} \cup \{ud\})$. In other words, m_c represents the redundant and the consistent objects that can be formed when the attribute c is added to the set of selected condition attributes R.
- x_c : the largest equivalence class obtained from the set of objects U and is based on the set of selected condition attributes R when the attribute c is added to it. In other words, m_c reflects the redundant objects not necessary consistent that can be formed when the attribute c is added to the set of selected condition attributes R.

Author's personal copy

Table 14
Initial state.

U	Height	ud
o_1	Short	$m_1(\{ud_2\}) = 0.5, m_1(\Theta) = 0.5$
o_2	Middle	$m_2(\{ud_2\}) = 1$
o_3	Short	$m_3(\{ud_1\}) = 0.7, m_3(\Theta) = 0.3$
o_5	Short	$m_5(\{ud_2\}) = 1$
o_6	Middle	$m_6(\{ud_2\}) = 0.95, m_6(\Theta) = 0.05$
o_8	Middle	$m_8(\{ud_1\}) = 0.975, m_8(\Theta) = 0.025$

4.4.3. Algorithm

Initial state
(1) R $\leftarrow UCore_C(\{ud\})$ // calculation of the core
$P \leftarrow C - UCore_C(\{ud\})$
(2) **while** $(\gamma_R(\{ud\}) < \varepsilon)$
(3) $U \leftarrow U - UPos_R(\{ud\})$ //optimization
(4) $\forall\, c \in P$
(5) $v_c = |UPos_{R\cup\{c\}}(\{ud\})|$
(6) $m_c = max - size(UPos_{R\cup\{c\}}(\{ud\})/R \cup \{c\} \cup \{ud\})$
(7) $x_c = max - size(U/R \cup \{c\})$
(8) Choose c with largest $v_c * m_c * x_c$
(9) $R \leftarrow R \cup \{c\}$
(10) P$\leftarrow$ P-{c}
(11) **return** R

The strategy for attribute selection used in this algorithm can be described as follows: select a given attribute c, if by adding it to the subset R of attributes, the cardinality of $UPos_{R\cup\{c\}}(\{ud\}))$ increases faster and the $max - size(UPos_{R\cup\{c\}}(\{ud\})/R \cup \{c\} \cup \{ud\}))$ and $max - size(U/R \cup \{c\})$ are larger than by adding any other attribute. These discussed conditions can be competitive. So, we choose in our quality criterion the result of multiplication.

Example: Let us continue with the same example in Table 3 to compute the reduct using the heuristic. We start by computing the core (the set of indispensable condition attributes) as follows:

Remove the *hair* attribute from the condition attributes:
$UPos_{\{Eyes,Height\}}(\{ud\}) = \{o_1, o_2, o_6, o_7, o_8\} = UPos_C(\{ud\})$
So, the *hair* attribute is not indispensable.

Remove the *eyes* attribute from the condition attributes:
$UPos_{\{Hair,Height\}}(\{ud\}) = \{o_1, o_2, o_6, o_7, o_8\} = UPos_C(\{ud\})$
So, the *eyes* attribute is not indispensable.

Remove the *height* attribute from the condition attributes:
$UPos_{\{Hair,Eyes\}}(\{ud\}) = \{o_8\} \neq UPos_C(\{ud\})$
So, the *height* attribute is indispensable.

Only the *height* attribute is indispensable with respect to *ud*. Hence, it is the core.

We have, in the initial state:
$R = UCore_C(\{ud\}) = \{Height\}$
$P = C\text{-}\ UCore_C(\{ud\}) = \{Hair, Eyes\}$

Setting reduct threshold: $\varepsilon = \gamma_C(\{ud\}) = 5/8$, the termination condition will be $\gamma_R(\{ud\}) \geq 5/8$. Since $\gamma_R(\{ud\}) = 2/8 < 5/8$, R is not a reduct, we must continue adding other condition attributes to R until a reduct is obtained.

The positive region of the attribute $\{Height\}$ with respect to {ud}: $UPos_{\{Height\}}(\{ud\}) = \{o_4, o_7\}$. From U, the consistent objects $\{o_4, o_7\}$ should be removed. The initial state is shown in Table 14 with $U = \{o_1, o_2, o_3, o_5, o_6, o_8\}$.

Next, we have two candidates *Hair* and *Eyes*. Tables 15 and 16 give the results of adding *Hair* and *Eyes* to R, respectively.

From these tables and according to the selection strategy, we compute the v_c, m_c and x_c for each candidate which are *Hair* and *Eyes* attributes as follows:

Author's personal copy

Table 15
Selecting the *Hair* attribute.

U	Hair	Height	ud
o_1	Dark	Short	$m_1(\{ud_2\}) = 0.5, m_1(\Theta) = 0.5$
o_2	Blond	Middle	$m_2(\{ud_2\}) = 1$
o_3	Blond	Short	$m_3(\{ud_1\}) = 0.7, m_3(\Theta) = 0.3$
o_5	Dark	Short	$m_5(\{ud_2\}) = 1$
o_6	Blond	Middle	$m_6(\{ud_2\}) = 0.95, m_6(\Theta) = 0.05$
o_8	Dark	Middle	$m_8(\{ud_1\}) = 0.975, m_8(\Theta) = 0.025$

Table 16
Selecting the *Eyes* attribute.

U	Eyes	Height	ud
o_1	Brown	Short	$m_1(\{ud_2\}) = 0.5, m_1(\Theta) = 0.5$
o_2	Blue	Middle	$m_2(\{ud_2\}) = 1$
o_3	Brown	Short	$m_3(\{ud_1\}) = 0.7, m_3(\Theta) = 0.3$
o_5	Brown	Short	$m_5(\{ud_2\}) = 1$
o_6	Blue	Middle	$m_6(\{ud_2\}) = 0.95, m_6(\Theta) = 0.05$
o_8	Brown	Middle	$m_8(\{ud_1\}) = 0.975, m_8(\Theta) = 0.025$

$$UPos_{\{Hair,Height\}}(\{ud\}) = \{o_2, o_6, o_8\}$$

$$UPos_{\{Eyes,Height\}}(\{ud\}) = \{o_2, o_6, o_8\}$$

$$v_{Hair} = |UPOS_{\{Hair,Height\}}(\{ud\})| = 3$$

$$v_{Eyes} = |UPOS_{\{Eyes,Height\}}(\{ud\})| = 3$$

The two candidates have the same v_c which is equal to 3. In other words, by adding *Hair* or *Eyes* to R, the objects o_2, o_6 and o_8 become consistent.

Then, we should check the value of m_c

$$UPos_{\{Hair,Height\}}(\{ud\})/\{Hair, Height, \{ud\}\} = \{\{o_2, o_6\}, \{o_8\}\}$$

$$UPos_{\{Eyes,Height\}}(\{ud\})/\{Eyes, Height, \{ud\}\} = \{\{o_2, o_6\}, \{o_8\}\}$$

$$m_{Hair} = 2, \text{ which is relative to the } \{o_2, o_6\}$$

$$m_{Eyes} = 2, \text{ which is relative to the } \{o_2, o_6\}.$$

The two candidates have the same m_c which is equal to 2. In other words, by adding *Hair* or *Eyes* to R, the objects o_2 and o_6 become consistent and redundant.

Finally, we should check the value of x_c

$$U/\{Hair, Height\} = \{\{o_2, o_6\}, \{o_3\}, \{o_4, o_5\}, \{o_8\}\}$$

$$U/\{Eyes, Height\} = \{\{o_2, o_6\}, \{o_3, o_4, o_5\}, \{o_8\}\}$$

$$x_{Hair} = 2, \text{ which is relative to the } \{o_2, o_6\} \text{ or } \{o_4, o_5\}$$

$$x_{Eyes} = 3, \text{ which is relative to the } \{o_3, o_4, o_5\}$$

The two candidates have not the same x_c. Since, the maximal set is in $U/\{Eyes, Height\}$. So, according to our selection strategy, *Eyes* should be selected first.

After adding *Eyes* to R, $\gamma_R(\{ud\})$= 5/8 $\geq$ 5/8. The process is finished. Thus, the selected attribute subset is $\{Eyes, Height\}$.

5. Experimentation

In our experiments, we have performed several tests on real-world databases to evaluate our two proposed classifiers the *Belief Rough Set Classifier* (BRSC) and the *Belief Rough Set Classifier* based on *Generalization Distribution Table* (BRSC-GDT) using our heuristic method for feature selection.

Author's personal copy

Table 17
Description of databases.

Database	#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

5.1. Experimental databases

- In the first set of experiments, we tested our classifiers on standard real-world databases obtained from the U.C.I. repository.[1] A brief description of these databases is presented in Table 17. These databases were artificially modified in order to include uncertainty in the decision attribute. We took different degrees of uncertainty (Low, Middle and High) based on increasingly 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 $m_j(\Theta) = P$.

  – Low degree of uncertainty: $0 < P <= 0.3$.
  – Middle degree of uncertainty: $0.3 < P <= 0.6$.
  – High degree of uncertainty: $0.6 < P <= 1$.

  A larger P gives a larger degree of uncertainty. Besides, we take in our experiments the case of no uncertainty in the decision attribute (certain case) in the chosen databases.
- In the second set of experiments, we tested our classifier on a naturally uncertain web usage database. The latter was obtained from web access logs of the introductory computing science course at Saint Mary's University. The course is "Introduction to Computing Science and Programming" offered in the first term of the first year. The initial number of students in the course was 180. The number reduced over the course of the semester to 140 students. The students in the course come from a wide variety of backgrounds, such as computing science major hopefuls, students taking the course as a required science course, and students taking the course as a science or general elective. As is common in a first year course, students' attitudes towards the course also vary a great deal. Lingras and West [22] showed that visits from students attending the first course could fall into one of the following three categories (decision values):

  – Studious: These visitors download the current set of notes. Since they download a limited/current set of notes, they probably study class-notes on a regular basis.
  – Crammers: These visitors download a large set of notes. This indicates that they have stayed away from the class-notes for a long period of time. They are planning for pretest cramming.
  – Workers: These visitors are mostly working on class or lab assignments or accessing the discussion board.

  The web logs were preprocessed to create an appropriate representation of each user, corresponding to a visit. The abstract representation of a web user is a critical step that requires a good knowledge of the application domain. Based on some observations, it was decided to use the following attributes for representing each visitor [22]:

  – On campus/Off campus access.
  – Day time/Night time access: 8 a.m. to 8 p.m. were considered to be the daytime.
  – Access during lab/class days or non-lab/class days: All the labs and classes were held on Tuesdays and Thursdays. The visitors on these days are more likely to be workers.
  – Number of hits.
  – Number of class-notes downloads.

  The first three attributes had binary values of 0 or 1. The last two variables represent the number of hits and number of class-notes, and were integer values. Total visits were 23754. The visits where no class-notes were downloaded were eliminated, since these visits correspond to either casual visitors or workers. Elimination of outliers and visits from the search engines further reduced the sizes of the data set to 7965. Instead of representing an object as belonging to a cluster ud_i (decision value), we also associated a degree of belief in the object belonging to the cluster ud_i. Let $\Theta = \{ud_i | 1 \leq i \leq k\}$ be the set of all the clusters (the possible decision values). We will use the cluster label ud_i and the cluster center vector interchangeably. If K-means algorithm identified the object o_j as belonging to cluster ud_i,

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

Author's personal copy

then the bba m_j for the object was calculated as follows:

$$m_j(\{ud_i\}) = \frac{sim(o_j, ud_i)}{\sum_{j=1}^{k} sim(o_j, ud_i)}, m_j(\Theta) = 1 - m_j(\{ud_i\}) \tag{41}$$

where $sim(o_j, ud_i) = 1/d(o_j, ud_i)$ represents the similarity between object o_j and ud_i and is calculated as the inverse of the distance d between them. The latter is a standard Euclidean distance. It should be noted here that the bba described above is just one of the possibilities. It is possible to have a more elaborate bba by using a threshold to identify more than one possible decision class for an object.

In our experiments, we will test this web usage database based on two cases:

- Certain case: each object o_j is belonging to one of the three clusters (Studious, Crammers, Workers).
- Uncertain case: instead of using crisp assignment, it is also associated a degree of belief m_j to the object belonging to the cluster ud_i.

5.2. Evaluation criteria

The relevant criteria used to judge the performance of our proposed classifiers are as follows:

(1) The time requirement represents the number of seconds needed for the construction procedure.
(2) The size represents the number of belief decision rules generated from the classifier.
(3) The classification accuracy represents the percent of correct classification (PCC) of the objects belonging to testing set. For each testing instance, we determine its corresponding rule and we look for the most probable decision corresponding to this rule using the pignistic probability computed from its bba.

Each database is divided into ten equal 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 ten fold cross-validation, provides a more reliable estimation of the evaluation criteria.

5.3. Experimental results from U.C.I repository databases

5.3.1. Experimental results with the certain case

To judge the performance of our rough set classifiers, we have first compared their results with those given by traditional classification methods such as C4.5, K-nearest neighbors (KNN) and Naive Bayes classifier for the certain case (no uncertainty in the decision attribute) and only for the classification accuracy criterion (the chosen classification methods are not similar). Table 18 gives the different results obtained from BRSC, BRSC-GDT, C4.5, KNN and Naive Bayes for unmodified U.C.I repository databases. Note that relative to the KNN, we have taken the best results from three different values of K=(1,3,5,7) which represents the number of nearest neighbors.

From this table, we can conclude that using our classification systems based on rough sets give more accurate models than using the chosen traditional methods for the chosen U.C.I repository databases. The improvement is especially notable when compared with the KNN. For example, the PCC for the W. Breast Cancer using KNN is equal to 79.33% and equal to 84.51% using BRSC. BRSC and BRSC-GDT also perform better than C4.5 and Naive Bayes Classifier. For example, the PCC for the Car Evaluation data base using Naive Bayes Classifier is equal to 80.19% and equal to 81.51% using BRSC-GDT.

5.3.2. Experimental results with the uncertain case

To further evaluate the performance of our two proposed classifiers, we compare our results with those given by another similar classifier relative the uncertain case (Low, Middle and High) and based on the three evaluation criteria (Time requirement, Size and PCC). We choose the Belief Decision Tree (BDT) approach developed in [11]. The latter is a decision tree in an uncertain environment where the uncertainty is represented through the TBM. The uncertainty appears in the decision attribute of training objects. Hence, the comparison between the two classifiers is possible. From the two procedures of building BDT, we choose for the comparison the conjunctive approach which is conceptually much closer to the TBM itself. The pruning step is necessary for improving the size and the accuracy of the BDT. In [33], we have concluded that the post-pruning is better than pre-pruning [12] in terms of size and accuracy. For this reason, we focus only on the post-pruned BDT with complexity equal to $O(k * n^2 * log\ n)$, where n is the number of instances in a given database and k stands for the number of attributes.

Author's personal copy

Table 18
Comparison with traditional methods based on PCC (%).

Database	BRSC	BRSC-GDT	C4.5	KNN	Naive Bayes
W. Breast Cancer	84.51	83.71	82.11	79.33	82.03
Balance Scale	83.89	81.53	80.46	78.07	80.61
C. Voting records	98.63	98.52	95.45	94.76	95.48
Zoo	94.11	93.64	91.01	89.61	91.89
Nursery	96.45	96.17	94.03	90.32	94.03
Solar Flares	89.74	89.71	87.67	83.66	87.64
Lung Cancer	76.69	75.58	72.25	69.35	72.36
Hyes-Roth	97.78	97.55	94.32	89.11	94.26
Car Evaluation	83.86	81.51	80.71	74.39	80.19
Lymphography	84.89	84.36	82.70	76.35	82.50
Spect Heart	87.78	87.49	84.38	81.17	84.35
Tic-Tac-Toe Endgame	87.05	86.35	84.81	82.26	84.39

Table 19
The experimental results for the time requirement (s).

Bases	Pruned BDT	BRSC	BRSC-GDT
W. Breast Cancer	66	41	156
Balance Scale	42	35	139
C. Voting records	68	27	117
Zoo	39	34	103
Nursery	191	127	386
Solar Flares	108	104	160
Lung Cancer	29	15	56
Hyes-Roth	34	28	93
Car Evaluation	133	105	189
Lymphography	65	43	108
Spect Heart	71	52	111
Tic-Tac-Toe Endgame	107	98	149

Table 20
The experimental results for the size relative to the BRSC.

Database	Low Unc	Middle Unc	High Unc	Mean
W. Breast Cancer	49	50	53	51
Balance Scale	61	61	61	61
C. Voting records	50	51	56	52
Zoo	41	41	42	41
Nursery	225	232	248	235
Solar Flares	125	127	128	127
Lung Cancer	23	23	29	25
Hyes-Roth	44	45	46	45
Car Evaluation	174	175	175	175
Lymphography	57	58	58	58
Spect Heart	51	55	57	54
Tic-Tac-Toe Endgame	122	125	127	125

The time requirement: In this part, we detail the experimental results for the first evaluation criterion, i.e. the time requirements for the construction of the BRSC, the BRSC-GDT and the pruned BDT. Table 19 gives results from all the databases for the time requirements for the three classifiers. Note that the time requirement is almost the same for different degrees of uncertainty. Table 19 shows that the two classification approaches based on rough sets, i.e. the BRSC and the BRSC-GDT, are faster than the pruned BDT. Especially, for large databases such as *Nursery database*, the time requirement goes down from 386s for pruned BDT to 191s with BRSC and to 127s with BRSC-GDT. Furthermore, we also conclude from Table 19 that the BRSC-GDT approach is slightly faster than the BRSC.

The size: We move to the second evaluation criterion representing the number of the learned uncertain decision rules generated from our three classifiers: the BRSC, the BRSC-GDT and the pruned BDT in conjunctive approach. Tables 20, 21 and 22 summarize the results for the size relative to the chosen databases for the three uncertain cases (Low Unc, Middle Unc, High Unc) with their mean. From these tables, we conclude that the pruned BDT gives more combined decision rules for all databases than the BRSC. For example, the number of decision rules induced from *Nursery database* is 206 using BDT and 235 using BRSC. However, this increase of size becomes smaller when the uncertainty increases. For example, the number of decision rules induced from *Nursery database* in low uncertainty case is 197 using BDT and 225 using BRSC. However, with high uncertainty, the size of model goes from 222 using BDT to 228 using BRSC. We also conclude that the BRSC-GDT gives size of models smaller than the pruned BDT and the BRSC for all the databases and for all the degrees of uncertainty. For example, the number of uncertain decision rules induced from *W. Breast Cancer database* goes from 47 using pruned BDT and 51 using BRSC to 41 using BRSC-GDT. Finally, we can also conclude that the size increases when the uncertainty increases. It is true for all databases and for all classifiers.

Author's personal copy

Table 21
The experimental results for the size relative to the BRSC-GDT.

Database	Low Unc	Middle Unc	High Unc	Mean
W. Breast Cancer	38	40	44	41
Balance Scale	49	53	55	52
C. Voting records	38	44	40	41
Zoo	27	29	32	29
Nursery	186	198	199	194
Solar Flares	109	114	123	115
Lung Cancer	21	22	22	22
Hyes-Roth	27	31	36	31
Car Evaluation	157	159	161	159
Lymphography	44	44	43	44
Spect Heart	43	43	45	44
Tic-Tac-Toe Endgame	104	111	116	110

Table 22
The experimental results for the size relative to the pruned BDT.

Database	Low Unc	Middle Unc	High Unc	Mean
W. Breast Cancer	43	47	52	47
Balance Scale	50	55	56	54
C. Voting records	38	43	52	44
Zoo	28	34	40	34
Nursery	197	199	222	206
Solar Flares	117	128	130	125
Lung Cancer	23	23	27	24
Hyes-Roth	33	34	37	35
Car Evaluation	170	170	173	171
Lymphography	49	50	55	51
Spect Heart	43	47	51	47
Tic-Tac-Toe Endgame	109	124	127	120

Table 23
The experimental results for the PCC (%) relative to the BRSC.

Database	Low Unc	Middle Unc	High Unc	Mean
W. Breast Cancer	84.41	84.01	83.67	84.03
Balance Scale	83.77	83.56	83.47	83.6
C. Voting records	98.58	98.27	98.01	98.28
Zoo	94.05	94.01	93.61	93.89
Nursery	96.34	96.03	95.72	96.03
Solar Flares	89.68	89.64	89.66	89.66
Lung Cancer	76.5	76.25	76.35	76.36
Hyes-Roth	97.65	97.32	97.11	97.36
Car Evaluation	83.79	83.71	82.09	83.19
Lymphography	84.75	84.70	85.04	84.50
Spect Heart	87.56	87.33	87.12	87.33
Tic-Tac-Toe Endgame	86.98	86.85	86.26	86.69

The classification accuracy: Finally, Tables 23, 24 and 25 detail the results for the more important evaluation criterion representing the classification accuracy for each classifier relative to the three degrees of uncertainty with their mean. From these tables, we conclude that our two classifiers based on rough sets called the BRSC and the BRSC-GDT, give better PCC than the pruned BDT for all the databases and for all the degrees of uncertainty. For *Solar Flares database*, the mean PCC obtained from the BRSC and the BRSC-GDT are respectively 89.66% and 86.62%. However, the mean PCC for the pruned BDT is 85.63%. On the other hand, we conclude that the BRSC gives better PCC than the BRSC-GDT for all the databases and for all the degrees of uncertainty. For example, the mean PCC obtained from *Balance database* is 83.6% using BRSC and 80.56% using BRSC-GDT. Finally, we can also conclude that the PCC decreases when the uncertainty increases. It is true for all databases and for all classifiers.

5.3.3. Experimental results with the noisy case

The main goal of the experiments in the previous sections was to show that with the same level or degree of the uncertainty, our two belief rough set classification systems are better in classification performance than the conventional classification methods as well as the pruned belief decision tree which is a similar classifier and handles uncertainty in the decision attribute. However, we conclude that the classification using the crisp decision value (no uncertainty) is better than introducing uncertainty in the decision values using belief functions. The uncertainty in this case was created artificially and randomly. Hence, the previous results are sufficient to show the utility of using the belief functions to represent the uncertainty in the condition attribute but they do not show that accommodation of uncertainty in the decision attributes can lead to better performance than using crisp ones.

Author's personal copy

Table 24
The experimental results for the PCC (%) relative to the BRSC-GDT.

Database	Low Unc	Middle Unc	High Unc	Mean
W. Breast Cancer	83.77	83.48	83.05	83.43
Balance Scale	81.46	80.21	80.03	80.56
C. Voting records	98.44	98.16	97.92	98.17
Zoo	93.52	93.47	92.87	93.28
Nursery	96.06	95.81	95.27	95.71
Solar Flares	89.67	89.61	89.56	89.61
Lung Cancer	75.50	75.50	66.33	72.44
Hyes-Roth	97.46	97.11	96.75	97.10
Car Evaluation	81.46	81.01	81.17	81.21
Lymphography	84.24	84.03	83.67	83.98
Spect Heart	87.34	87.28	87.07	87.23
Tic-Tac-Toe Endgame	86.26	86.21	86.18	86.21

Table 25
The experimental results for the PCC (%) relative to the pruned BDT.

Database	Low Unc	Middle Unc	High Unc	Mean
W. Breast Cancer	83.46	83.01	82.17	82.88
Balance Scale	78.15	77.83	77.76	77.91
C. Voting records	98.28	97.76	97.71	97.91
Zoo	91.94	91.36	91.41	91.57
Nursery	95.84	95.13	95.11	95.36
Solar Flares	85.78	85.61	85.46	85.61
Lung Cancer	74.63	74.36	74.07	74.35
Hyes-Roth	83.66	83.31	82.14	83.03
Car Evaluation	73.49	73.11	72.97	73.19
Lymphography	79.25	78.97	78.94	79.05
Spect Heart	83.46	83.01	82.17	82.88
Tic-Tac-Toe Endgame	83.91	83.75	83.42	83.69

Table 26
The experimental results for the PCC (%) by introducing noise.

Database	Crisp noise BRSC	Unc noise BRSC	Crisp noise BRSC-GDT	Unc noise BRSC-GDT	Crisp noise Pruned BDT	Unc noise Pruned BDT
W. Breast Cancer	84.44	85.87	83.97	85.31	83.73	84.94
Balance Scale	83.36	84.89	82.85	84.63	82.76	84.35
C. Voting records	97.46	98.90	97.24	98.68	97.17	98.61
Zoo	94.49	96.23	94.02	95.63	93.68	95.08
Nursery	95.74	97.06	95.62	96.91	95.33	96.53
Solar Flares	89.26	90.78	89.19	90.70	88.87	90.19
Lung Cancer	76.45	77.97	75.88	77.38	75.76	77.12
Hyes-Roth	96.62	97.91	96.13	97.84	96.11	97.60
Car Evaluation	82.77	84.29	82.72	84.13	82.39	83.78
Lymphography	84.59	85.94	84.23	85.76	83.82	85.38
Spect Heart	86.59	87.97	86.34	87.81	86.19	87.49
Tic-Tac-Toe Endgame	86.77	88.10	86.55	87.94	86.44	87.83

Therefore, we add noise in the decision attribute for all the records in each database. We use 20% probability for randomly changing the class (decision attribute value) as follows: we generate a random number between 0 to 9. If the number is 8 or 9, then we change the value of the decision attribute (if the original class were 0 change to 1 and vice versa). We obtain in this case databases with crisp noisy decision attributes.

For these records, we use the same level of noise (20% probability) as a metaknowledge for fixing the bba's of the decision attribute to be $m(\{d_i\}) = 0.8$ and $m(\Theta) = 0.2$. (where d_i is the real decision value). This provides us databases with uncertain noisy decision attributes.

The objective is to compare the results obtained from two cases:

(1) Database with crisp noisy decision attribute
(2) Database with uncertain noisy decision attribute

Table 26 shows the experimental results obtained from BRSC, BRSC-GDT and pruned BDT for both the cases: "crisp noisy decision" and "uncertain noisy decision". We find that the classification accuracy for the uncertain case is much better than for the crisp case. It is true for all databases and for all classifiers. For example, the PCC for the BRSC obtained from Zoo database goes from 94.49% with the crisp noisy case to 96.23% with uncertain noisy case.

We can conclude that the classification accuracy is higher for databases with some partially uncertain decision attributes represented using belief functions than those databases with crisp noisy decision attribute.

Author's personal copy

Table 27
Experimental results for the web usage database.

Approaches	PCC (%)		Size		Time requirement
	Certain case	Uncer case	Certain case	Uncer case	(s)
pruned BDT	84.07	85.12	37	39	188
BRSC	85.16	89.63	41	46	157
BRSC-GDT	84.92	88.46	32	35	143

5.4. Experimental results for the web usage database

In this subsection, we apply our classification approaches based on rough sets under belief functions (BRSC, BRSC-GDT) on a real-world web usage mining database. We compare the results with those obtained from the pruned BDT.

Table 27 summarizes the results for the three evaluation criteria by applying our rough set based approaches to learn decision rules from the web usage mining in comparison to those obtained from the BDT.

- For the classification accuracy criterion, PCC for precise clustering (certain case) is less than the PCC based on the degrees of belief in cluster membership (uncertain case). For example, the PCC of the BRSC is equal to 85.24 % with certain case. However, the PCC goes up to 89.63% for uncertain case. We also conclude that the PCC for our rough set based approaches is better than the PCC for the pruned BDT.
- For the *size criterion*, we find that the sizes of the generated models for all the approaches in the certain case is less than those for uncertain case. For example, the size of the BRSC-GDT goes up from 32 with certain case to 35 with uncertain case. We also conclude that the size for the BRSC-GDT is smaller than the size for the post pruned BDT.
- The *time requirement* is almost the same for the certain and the uncertain case. Our rough set based classification approaches are faster than the post-pruned BDT. For example, the time requirement for the BRSC-GDT is equal to 143 seconds. However, the time requirement for the pruned BDT is 188 seconds.

Based on these observations, we conclude that our classification approaches based on rough sets may be better for the web usage mining databases, where cluster assignments are characterized by belief functions.

6. Summary and conclusions

In this paper, we have presented two new classification approaches based on rough sets called BRSC and BRSC-GDT under the belief function framework. Both are shown to simplify the uncertain decision table and to generate more efficient decision rules for the classification process. The uncertainty appears only in decision attribute values which is represented through the belief functions. We have also proposed a heuristic for the attribute selection to avoid costly calculations of reduct. To judge the performance of our belief rough set classifiers, we have first compared them with some traditional classification methods such as decision tree, K-nearest neighbors and Naive Bayes classifier only from crisp conventional decision attributes. The results are very interesting end encouraging. To further evaluate our solutions in the uncertain case, we have compared the results with those given by *Belief Decision Tree* (BDT) after post-pruning.

According to three evaluation criteria, we conclude that the BRSC and the BRSC-GDT are more accurate than the pruned BDT. This increased accuracy can be easily explained. Our classification techniques based on rough sets try to reduce the UDT without affecting the classification task. Our experiments also show that the BRSC is more accurate than the BRSC-GDT. We further conclude that the post-pruned BDT gives more combined decision rules than the BRSC. This is due to the pruning which can reduce the size of the BDT. However, the model for the BRSC-GDT is smaller than the pruned BDT and the BRSC. This performance gain is due to the fact that the BRSC-GDT selects only the best and the non-contradictory decision rules. Finally, we also conclude that our new classification approaches based on rough sets are faster than the post-pruned BDT. This positive result is due to the heuristic method used in the construction of the BRSC and the BRSC-GDT which produces only one reduct from our UDT. On the other hand, the pruning increases the time requirement needed to build the BDT. Furthermore, the BRSC-GDT, which can avoid many iterations, is slightly faster than the BRSC. Hence, we can summarize our conclusions as follows:

(1) Use BRSC for more accurate decision making.
(2) Use BRSC-GDT for quick decision making.
(3) Use BRSC-GDT for quick model building.

As a future work, we plan to investigate uncertainty in condition attribute values.

Author's personal copy

Acknowledgements

The authors would like to thank the anonymous reviewers for their detailed comments, criticisms, and suggestions that have helped greatly in improving the quality of this paper.

References

[1] A. An, C. Chan, N. Shan, N. Cercone, W. Ziarko, Applying knowledge discovery to predict water-supply consumption, IEEE Expert 12 (4) (1997) 72–78.
[2] M. Bauer, Approximations algorithm and decision making in the Dempster–Shafer theory of evidence – an empirical study, IJAR 17 (2–3) (1997) 217–237.
[3] S. Ben Hariz, Z. Elouedi, K. Mellouli, Clustering approach using belief function theory, in: 12th International Conference, AIMSA 2006, Varna, Bulgaria, September 12–15, LNAI 4183, 2006, pp. 162–171.
[4] E. Bosse, A.L. Jousseleme, D. Grenier, A new distance between two bodies of evidence, Information Fusion 2 (2001) 91–101.
[5] B.V. Dasarathy, Nearest Neighbour (NN) Norms: NN Pattern Classification Techniques, IEEE Computer Society Press, Los Alamitos, CA, 1991.
[6] T. Denoeux, A *k*-nearest neighbor classification rule based on Dempster–Shafer theory, IEEE Transactions on Systems, Man and Cybernetics 25 (5) (1995) 804–813.
[7] T. Denoeux, M. Skarstien-Bajanger, Induction of decision trees form partially classified data using belief functions, in: IEEE International Conference on Systems, Nashville, USA, vol. 4, 2000, pp. 2923–2928.
[8] T. Denoeux, A neural network classifier based on Dempster–Shafer theory, IEEE Transactions on Systems, Man and Cybernetics, Part A 30 (2) (2000) 131–150.
[9] J.Z. Dong, N. Zhong, S. Ohsuga, Probabilistic rough induction: the GDS-RS methodology and algorithms, in: Z.W. Ras, A. Skowron (Eds.), Foundations of Intelligent Systems, Springer, Berlin, 1999, pp. 621–629.
[10] D. Dubois, Putting rough sets and fuzzy sets together, in: R. Slowinski (Ed.), Intelligent Decision Support, Kluwer, Dordrecht, 1992, pp. 203–232.
[11] Z. Elouedi, K. Mellouli, P. Smets, Belief decision trees: theoretical foundations, International Journal of Approximate Reasoning 28 (2–3) (2001) 91–124.
[12] Z. Elouedi, K. Mellouli, P. Smets, A pre-pruning method in belief decision trees, in: The Ninth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Annecy, France, vol. 1, 2002, pp. 579–586.
[13] Z. Elouedi, K. Mellouli, P. Smets, Assessing sensor reliability for multisensor data fusion within the transferable belief model, IEEE Transactions on Systems, Man and Cybernetics 34 (1) (2004) 782–787.
[14] T.F. Fan, C.J. Liau, D.R. Liu, Dominance-based fuzzy rough set analysis of uncertain and possibilistic data tables, International Journal of Approximate Reasoning 52 (9) (2011) 1283–1297.
[15] D. Fixen, R.P.S. Mahler, The modified Dempster–Shafer approach to classification, IEEE Transactions on Systems, Man and Cybernetics A 27 (1) (1997) 96–104.
[16] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Machine Learning 29 (1997) 131–163.
[17] J.W. Grzymala-Busse, Rough set strategies to data with missing attribute values, in: Workshop Notes, Foundations and New Directions of Data Mining, The 3rd International Conference on Data Mining, Melbourne, FL, 2003, pp. 56–63.
[18] J.W. Grzymala-Busse, S. Siddhaye, Rough set approaches to rule induction from incomplete data, in: Proceedings of the IPMU'2004, Perugia, Italy, July 4–9, vol. 2, 2004, pp. 923–930.
[19] T.P. Hong, L.H. Tseng, B.C. Chien, Learning fuzzy rules from incomplete quantitative data by rough sets, IEEE, 2002, pp. 1438–1442.
[20] P. Maji, S. Paul, Rough set based maximum relevance-maximum significance criterion and Gene selection from microarray data, International Journal of Approximate Reasoning 52 (3) (2011) 408–426.
[21] C.K. Murphy, Combining belief functions when evidence conflicts, Decision Support Systems 29 (2000) 1–9.
[22] P. Lingras, C. West, Interval set clustering of web users with rough *k*-means, Journal of Intelligent Information Systems 23 (1) (2004) 5–16.
[23] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11 (5) (1982) 341–356.
[24] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishing, Dordrecht, 1991.
[25] J.R. Quinlan, C4.5: Programs for Machine learning, Morgan Kaufman, San Mateo, CA, 1993.
[26] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, NJ, 1976.
[27] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, in: R. Slowinski (Ed.), Intelligent Decision Support, Kluwer Academic Publishers, Boston, MA, 1992, pp. 331–362.
[28] A. Skowron, J.W. Grzymala-Busse, From rough set theory to evidence theory, in: Advances in the Dempster–Shafer Theory of Evidence, New York, 1994, pp. 193–236.
[29] P. Smets, R. Kennes, The transferable belief model, Artificial Intelligence 66 (2) (1994) 191–234.
[30] P. Smets, The transferable belief model for quantified belief representation, in: D.M. Gabbay, P. Smets (Eds.), Handbook of Defeasible Reasoning and Uncertainty Management Systems, vol. 1, Kluwer, Dordrecht, The Netherlands, 1998, pp. 207–301.
[31] P. Smets, Application of belief transferable belief model to diagnostic problems, International Journal of Intelligent Systems 13 (2–3) (1998) 127–157.
[32] D. Tian, X.J. Zeng, J. Keane, Core-generating approximate minimum entropy discretization for rough set feature selection in pattern classification, International Journal of Approximate Reasoning 52 (6) (2011) 863–880.
[33] S. Trabelsi, Z. Elouedi, K. Mellouli, Pruning belief decision tree methods in averaging and conjunctive approaches, International Journal of Approximate Reasoning 46 (2007) 91–124.
[34] S. Trabelsi, Z. Elouedi, Learning decision rules from uncertain data using rough sets, in: The 8th International FLINS Conference on Computational Intelligence in Decision and Control, Madrid, Spain, September 21–24, World Scientific, 2008, pp. 114–119.
[35] S. Trabelsi, Z. Elouedi, P. Lingras, Belief rough set classifier, Canadian AI 2009, LNAI 5549, 2009, pp. 257–261.
[36] S. Trabelsi, Z. Elouedi, P. Lingras, Rule discovery process based on rough sets under the belief function framework, IPMU 2010, LNAI 6178, 2010, pp. 726–736.
[37] S. Trabelsi, Z. Elouedi, Heuristic method for attribute selection from partially uncertain data using rough sets, International Journal of General Systems 39 (3) (2010) 271–290.
[38] S. Trabelsi, Z. Elouedi, P. Lingras, Belief rough set classification for web mining based on dynamic core, in: The Tenth International Conference on Intelligent Systems Design and Applications (ISDA), Cairo, Egypt, November 29, December 1, 2010, pp. 403–408.
[39] P. Vannoorenberghe, On aggregating belief decision trees, Information Fusion 5 (2) (2004) 179–188.
[40] N. Zhong, J.Z. Dong, S. Ohsuga, Using rough sets with heuristics for feature selection, Journal of Intelligent Information Systems 16 (3) (2001) 199–214.
[41] N. Zhong, J.Z. Dong, S. Ohsuga, Data mining: a probabilistic rough set approach, Rough Sets in Knowledge Discovery 2 (1998) 127–146.
[42] N. Zhong, J.Z. Dong, S. Ohsuga, Using generalization distribution tables as a hypotheses search space for generalization, in: Proceedings of Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery (RSFD-96), 1996, pp. 396–403.
[43] L.M. Zouhal, T. Denoeux, An evidence-theory k-NN rule with parameter optimization, IEEE Transactions on Systems, Man and Cybernetics C 28 (2) (1998) 263–271.