ARTICLE IN PRESS

Information Sciences xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

Attribute selection with fuzzy decision reducts

Chris Cornelis [a,*], Richard Jensen [b], Germán Hurtado [a,c], Dominik Ślęzak [d,e]

[a] Department of Applied Mathematics and Computer Science, Ghent University, Gent, Belgium
[b] Department of Computer Science, Aberystwyth University, Wales, UK
[c] Department of Applied Engineering Sciences, University College Ghent, Gent, Belgium
[d] Institute of Mathematics, University of Warsaw, Warsaw, Poland
[e] Infobright Inc., Warsaw, Poland

ARTICLE INFO

Article history:
Received 12 January 2009
Received in revised form 7 July 2009
Accepted 15 September 2009
Available online xxxx

Keywords:
Rough sets
Fuzzy sets
Attribute selection
Data analysis
Decision reducts

ABSTRACT

Rough set theory provides a methodology for data analysis based on the approximation of concepts in information systems. It revolves around the notion of discernibility: the ability to distinguish between objects, based on their attribute values. It allows to infer data dependencies that are useful in the fields of feature selection and decision model construction. In many cases, however, it is more natural, and more effective, to consider a gradual notion of discernibility. Therefore, within the context of fuzzy rough set theory, we present a generalization of the classical rough set framework for data-based attribute selection and reduction using fuzzy tolerance relations. The paper unifies existing work in this direction, and introduces the concept of fuzzy decision reducts, dependent on an increasing attribute subset measure. Experimental results demonstrate the potential of fuzzy decision reducts to discover shorter attribute subsets, leading to decision models with a better coverage and with comparable, or even higher accuracy.

1. Introduction

Rough set theory, initiated by Pawlak [32,34] in the early 80s, presents data analysts with an elegant and powerful formal framework for describing and exploiting data dependencies. In particular, it serves very well the purpose of semantics-preserving data dimensionality reduction, i.e.: to omit attributes (features) from decision systems (a particular form of representing data gathered for classification purposes) without sacrificing the ability to *discern* between objects belonging to different decision classes, or, more generally, to serve for decision models that approximate those classes well enough (see e.g. [2,11,31,42,43,56,60]). A minimal set of attributes that preserves the decision making power of the original system is called a *decision reduct*. It is worth noting that such understanding of minimality of a subset of attributes is popular also in other domains (see e.g. *Markov boundaries* in probabilistic reasoning [37]). It is also worth emphasizing that such understood approach to reducing attributes should be considered within a wider framework of feature selection methods (especially in relation to so called filter methods [3,29]), wherein the objective is to minimize the complexity of data-based decision models with no harm to their accuracy [20,25,41].

Traditionally, discernibility is modeled by an equivalence relation in the set of objects: two objects are indiscernible w.r.t. a given set of attributes B if they have the same values for all the attributes in B. Discernibility may then be used to model functional dependencies between sets of attributes, as proposed also in other fields of data analysis [15,28]. In practice, this amounts to verifying (exact) equality of values. Such understood concept of discernibility works well for most qualitative

* Corresponding author. Tel.: +32 09 264 47 72; fax: +32 09 264 49 95.
E-mail addresses: Chris.Cornelis@UGent.be (C. Cornelis), rkj@aber.ac.uk (R. Jensen), German.HurtadoMartin@UGent.be (G. Hurtado), slezak@infobright.com (D. Ślęzak).

0020-0255/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2009.09.008

data, in particular if the number of distinct values for each attribute is limited and there is no particular relationship among them. Quantitative data, however, satisfy neither of these restrictions: they involve *continuous* (i.e., real-valued) attributes like age, speed or length, and are tied to a natural scale of *closeness*, (or, e.g., ordering [18],) loosely expressing that the closer the attribute values of two objects are, the less discernible they are. While the standard methodology can be tailored to handle them, e.g. by applying discretization [31,36,44] to replace exact attribute values by interval codes, it appears more natural to consider a notion of approximate equality, or graded indiscernibility, between objects [35,39,46]. Incidentally, note that for many complex qualitative attributes, whose values can be strings, images, ...it also makes sense to consider degrees of indiscernibility. On the other hand, Stefanowski and Tsoukiás [45] argued to model missing value semantics in data by means of valued tolerance relations. In general, the notion of approximate equality can be formally modeled by means of a fuzzy relation [59] in the set of objects.

Guided by this principle, the original rough set framework for data-based attribute selection and reduction has been generalized.[1] Besides defining fuzzy rough sets (see e.g. [12,57]), the use of fuzzy (similarity) relations for deriving fuzzy decision rules has been advocated [16,17] early on. Many approaches have in common that they redefine the notion of a reduct of an information system to take into account the "fuzzy" characteristics of the data (be it objects' gradual discernibility [21,24,55], or their partial membership to the classes of a fuzzy partition [4,5,22,23,47,50,51,61]). Our approach differs from the previous research efforts by the introduction of *fuzzy decision reducts*: conceptually, an attribute subset is a fuzzy decision reduct to a degree α (a value between 0 and 1) if it preserves the predictive ability of the original decision system at least to that degree. This idea can be seen as the fuzzy-rough analogue of using approximate decision reducts [43,44,52,58] in crisp rough set analysis, where original criteria for semantics-preserving data dimensionality reduction turn out to be too restrictive for real-life data.

Just as there are numerous ways of defining decision reducts in fuzzy-rough data analysis, so there is no single way of telling how a fuzzy decision reduct should look like. In the general definition that we propose, we require an increasing $[0,1]$-valued measure, so as to guarantee that the larger an attribute subset, the higher its degree of fuzzy decision reductthood (monotonicity), which is in analogy to other approaches to define a degree of approximating decision classes [43,44]. For practical purposes, we consider various alternatives, which can be grouped along two main directions: the first direction works with an extension of the well-known positive region and dependency degree, similar to what has been proposed by Jensen and Shen in [23], while the second one is based on an extension of the discernibility function from classical rough set analysis, related to the proposal in [24]. In this sense, the present paper also provides a unified framework of fuzzy-rough (approximate) reduction strategies.

As the proposed fuzzy decision reducts are shorter than crisp ones, the reduced decision systems have less conditional attributes. As such, they yield more general classification and regression models (see also [36,52] in the context of approximate decision reducts). Naturally, this only makes sense provided the accuracy of the model does not drop too much (cf. [20,29]). Therefore, we perform a series of experiments on benchmark data sets; for data sets involving a qualitative decision attribute, we try to establish the decision class to which a test object belongs (classification), while with a quantitative decision attribute, a prediction of the exact value of the decision attribute is attempted (regression). In particular, we evaluate the impact of decreasing the degree of reductthood α, and compare it to the special situation where $\alpha = 1$, i.e., in which the corresponding crisp reduct version is recovered.

The remainder of this paper is organized as follows: after recalling some important preliminaries of rough sets, fuzzy sets and their hybridization in Section 2, in Section 3 we propose a general definition for the concept of a fuzzy decision reduct, and develop and investigate a number of concrete instances of it. In Section 4, several experiments are conducted to demonstrate the advantage of fuzzy decision reducts over crisp ones, and to compare the effectiveness of the various alternative definitions to each other. In Section 5, we conclude. Finally, we note that a preliminary version of part of the subject matter in this paper appears in [8].

2. Preliminaries

2.1. Rough set theory

2.1.1. Definitions

In rough set analysis [33], data is represented as an *information system* $(X, \mathscr{A})$, where $X = \{x_1, \ldots, x_n\}$ and $\mathscr{A} = \{a_1, \ldots, a_m\}$ are finite, non-empty sets of objects and attributes, respectively. Each a in $\mathscr{A}$ corresponds to an $X \to V_a$ mapping, in which V_a is the value set of a over X. For every subset B of $\mathscr{A}$, the B-indiscernibility relation[2] R_B is defined as

$$R_B = \{(x,y) \in X^2 \text{and} (\forall a \in B)(a(x) = a(y))\}. \tag{1}$$

Clearly, R_B is an equivalence relation. Its equivalence classes $[x]_{R_B}$ can be used to approximate concepts, i.e., subsets of the universe X. Given $A \subseteq X$, its lower and upper approximation w.r.t. R_B are defined by

$$R_B \downarrow A = \{x \in X | [x]_{R_B} \subseteq A\}, \tag{2}$$

$$R_B \uparrow A = \{x \in X | [x]_{R_B} \cap A \neq \emptyset\}. \tag{3}$$

[1] For completeness, we mention that there also exist many fuzzy feature selection methods that are not based on rough set theory, see e.g. [38,49].

[2] When $B = \{a\}$, i.e., B is a singleton, we will write R_a instead of $R_{\{a\}}$.

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

A *decision system* $(X, \mathscr{A} \cup \{d\})$ is a special kind of information system, used in the context of classification, in which d $(d \neq \mathscr{A})$ is a designated attribute called the decision attribute. Its equivalence classes $[x]_{R_d}$ are called decision classes.

Given $B \subseteq \mathscr{A}$, the B-positive region POS_B contains those objects from X for which the values of B allow to predict the decision class unequivocally:

$$POS_B = \bigcup_{x \in X} R_B \downarrow [x]_{R_d}. \tag{4}$$

Indeed, if $x \in POS_B$, it means that whenever an object has the same values as x for the attributes in B, it will also belong to the same decision class as x. The predictive ability w.r.t. d of the attributes in B is then measured by the following value (degree of dependency of d on B):

$$\gamma_B = \frac{|POS_B|}{|X|}. \tag{5}$$

$(X, \mathscr{A} \cup \{d\})$ is called *consistent* if $\gamma_{\mathscr{A}} = 1$. A subset B of $\mathscr{A}$ is called a *decision reduct* if it satisfies $POS_B = POS_{\mathscr{A}}$, i.e., B preserves the decision making power of $\mathscr{A}$, and moreover it cannot be further reduced, i.e., there exists no proper subset B' of B such that $POS_{B'} = POS_{\mathscr{A}}$. If the latter constraint is lifted, i.e., B is not necessarily minimal, we call B a decision superreduct.

Example 1. Consider the following decision system[3] with seven objects and eight conditional attributes, all quantitative:

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	d
x_1	1	101	50	15	36	24.2	0.526	26	0
x_2	8	176	90	34	300	33.7	0.467	58	1
x_3	7	150	66	42	342	34.7	0.718	42	0
x_4	7	187	68	39	304	37.7	0.254	41	1
x_5	0	100	88	60	110	46.8	0.962	31	0
x_6	0	105	64	41	142	41.5	0.173	22	0
x_7	1	95	66	13	38	19.6	0.334	25	0

The decision attribute is qualitative, and there are only two decision classes: X_0 $(= [x_1]_{R_d})$ contains all x for which $d(x) = 0$, while X_1 $(= [x_2]_{R_d})$ contains those with $d(x) = 1$. If we want to apply the standard rough set analysis approach, we first have to preprocess the system. For instance, the numerical values for the conditional attributes can be replaced by interval codes, i.e., integers recording the interval to which the actual values belong. A possible discretization is given by

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	d
x_1	0	0	0	0	0	0	2	0	0
x_2	1	2	2	1	1	1	1	1	1
x_3	1	1	1	1	1	2	2	1	0
x_4	1	2	1	1	1	2	0	1	1
x_5	0	0	2	1	0	3	2	1	0
x_6	0	0	1	1	0	3	0	0	0
x_7	0	0	1	0	0	0	1	0	0

Now we can easily calculate the positive region. For example, given $B = \{a_4, a_5\}$,

$$POS_B = \{x_1, x_5, x_6, x_7\}.$$

On the other hand,

$$POS_{\mathscr{A}} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$$

indicating that the decision system is consistent.

Decision reducts can be used to synthesize minimal decision rules: the rules result from overlaying the reducts over the original (training) decision system and reading off the values. These rules can then be used to evaluate new (training) objects with unknown decision class.

2.1.2. Finding decision reducts

Below we recall a well-known approach to generate all reducts of a decision system based on its (decision-relative) discernibility matrix and function [42]. The discernibility matrix of $(X, \mathscr{A} \cup \{d\})$ is the $n \times n$ matrix O, defined by, for i and j in $\{1, \ldots, n\}$,

[3] This is a sample taken from the Pima Indians Diabetes data set located at the UCI Machine Learning repository, available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

$$O_{ij} = \begin{cases} \emptyset & \text{if } d(x_i) = d(x_j), \\ \{a \in \mathscr{A} | a(x_i) \neq a(x_j)\} & \text{otherwise.} \end{cases} \tag{6}$$

On the other hand, the *discernibility function* of $(X, \mathscr{A} \cup \{d\})$ is the $\{0,1\}^m \to \{0,1\}$ mapping f, defined by

$$f(a_1^*, \ldots, a_m^*) = \bigwedge \left\{ \bigvee O_{ij}^* | 1 \leqslant i < j \leqslant n \text{ and } O_{ij} \neq \emptyset \right\} \tag{7}$$

in which $O_{ij}^* = \{a^* | a \in O_{ij}\}$. The boolean variables $a_1^*, \ldots, a_m^*$ correspond to the attributes from $\mathscr{A}$, and we denote $\mathscr{A}^* = \{a_1^*, \ldots, a_m^*\}$. If $B \subseteq \mathscr{A}$, then the valuation function $\mathscr{V}_B$ corresponding to B is defined by $\mathscr{V}_B(a^*) = 1$ iff $a \in B$. This valuation can be extended to arbitrary boolean formulas, such that

$$\mathscr{V}_{\mathscr{B}}(f(a_1^*, \ldots, a_m^*)) = f(\mathscr{V}_{\mathscr{B}}(a_1^*), \ldots, \mathscr{V}_{\mathscr{B}}(a_m^*)). \tag{8}$$

Formula (8) expresses whether the attributes in B preserve the discernibility of $(X, \mathscr{A} \cup \{d\})$ (when its value is 1) or not (when it is 0). The discernibility function can be reduced to its disjunctive normal form, that is

$$f(a_1^*, \ldots, a_m^*) = \bigwedge A_1^* \vee \cdots \vee \bigwedge A_p^* \tag{9}$$

in which $p \geqslant 1$, and for all i in $\{1, \ldots, p\}$ it holds that $A_i^* \subseteq \mathscr{A}^*$, and $A_i^* \subseteq A_j^*$ for $i \neq j$. If we define $a \in A_i$ iff $a^* \in A_i^*$, then it can be shown [42] that $A_1, \ldots, A_p$ constitute exactly all decision reducts of $(X, \mathscr{A} \cup \{d\})$.

Example 2. For the discretized decision system in Example 1, it can be verified that the discernibility function (after reduction) is given by

$$f(a_1^*, \ldots, a_8^*) = a_2^* \vee (a_1^* \wedge a_7^*) \vee (a_5^* \wedge a_7^*) \vee (a_6^* \wedge a_7^*) \vee (a_7^* \wedge a_8^*). \tag{10}$$

Hence, the decision reducts are $\{a_2\}, \{a_1, a_7\}, \{a_5, a_7\}, \{a_6, a_7\}$ and $\{a_7, a_8\}$.

Computing all decision reducts is an NP-complete problem [42]. In practice, however, it suffices to generate only a subset of reducts [44], or even only one of them. Also, if reducts are too time consuming to be derived, it may be acceptable to generate superreducts which are not necessarily minimal. Below we recall a version of the QuickReduct algorithm ([6,23], see also [48] for a very similar approach), which finds a single superreduct of the decision system based on the degree of dependency. We would like to emphasize, however, that there are many other algorithms [2,54], usually developed for the classical framework of rough set-based attribute reduction, which may be adapted to the needs of the approach proposed in this paper.

QuickReduct starts off with an empty set R. It computes $\gamma_{R \cup \{a_i\}}$ for each attribute a_i (i in $\{1, \ldots, m\}$); the attribute for which this value is highest (or one of them in case there are several) is selected and added to R. Then, the same process is repeated for the remaining attributes, until $\gamma_R = \gamma_{\mathscr{A}}$. By construction, when the algorithm finishes, the set R is guaranteed to equal a decision superreduct of the decision system.

Example 3. If we apply QuickReduct to the discretized decision system from Example 1, we get, after one iteration, $R = \{a_2\}$, and since $\gamma_{a_2} = 1$, the algorithm terminates.

2.2. Fuzzy set theory

Fuzzy set theory [59] allows that objects belong to a set, or couples of objects belong to a relation, to a given degree. Recall that a fuzzy set in X is an $X \to [0,1]$ mapping, while a fuzzy relation in X is a fuzzy set in $X \times X$. For all y in X, the R-foreset of y is the fuzzy set Ry defined by

$$Ry(x) = R(x, y) \tag{11}$$

for all x in X. If R is a reflexive and symmetric fuzzy relation, that is,

$$R(x, x) = 1, \tag{12}$$

$$R(x, y) = R(y, x) \tag{13}$$

hold for all x and y in X, then R is called a fuzzy tolerance relation. For a fuzzy tolerance relation R, we call Ry the fuzzy tolerance class of y.

For fuzzy sets A and B in X, $A \subseteq B \Longleftrightarrow (\forall x \in X)(A(x) \leqslant B(x))$. If X is finite, the cardinality of A is calculated by

$$|A| = \sum_{x \in X} A(x). \tag{14}$$

Fuzzy logic connectives play an important role in the development of fuzzy rough set theory. We therefore recall some important definitions. A triangular norm (t-norm for short) $\mathscr{T}$ is any increasing, commutative and associative $[0,1]^2 \to [0,1]$ mapping satisfying $\mathscr{T}(1, x) = x$, for all x in $[0,1]$. In this paper, we use $\mathscr{T}_M$ and $\mathscr{T}_L$ defined by $\mathscr{T}_M(x, y) = \min(x, y)$ and $\mathscr{T}_L(x, y) = \max(0, x + y - 1)$ (Łukasiewicz t-norm), for x, y in $[0,1]$. On the other hand, an implicator is any $[0,1]^2 \to [0,1]$-mapping $\mathscr{I}$ satisfying $\mathscr{I}(0,0) = 1, \mathscr{I}(1, x) = x$, for all x in $[0,1]$. Moreover we require $\mathscr{I}$ to be decreasing in its first, and increasing in its second component. The implicators used in this paper are $\mathscr{I}_M$ and $\mathscr{I}_L$ defined by $\mathscr{I}_M(x, y) = \max(1 - x, y)$ (Kleene–Dienes implicator) and $\mathscr{I}_L(x, y) = \min(1, 1 - x + y)$ (Łukasiewicz implicator) for x, y in $[0,1]$.

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

2.3. Fuzzy rough set theory

Research on the hybridization of fuzzy sets and rough sets emerged in the late 1980s [12,13] and has flourished recently (see e.g. [7,27,30,35]). It has focused predominantly on fuzzifying the formulas (2) and (3) for lower and upper approximation. In doing so, the following two guiding principles have been widely adopted:

- The set A may be generalized to a fuzzy set in X, allowing that objects can belong to a given concept (i.e., meet its characteristics) to varying degrees.
- Rather than assessing objects' indiscernibility, we may measure their *approximate equality*, represented by a fuzzy relation R. As a result, objects are categorized into classes, or granules, with "soft" boundaries based on their similarity to one another. As such, abrupt transitions between classes are replaced by gradual ones, allowing that an element can belong (to varying degrees) to more than one class.

Typically, we assume that R is at least a fuzzy tolerance relation[4] For our purposes, given a decision system $(X, \mathscr{A} \cup \{\mathscr{d}\})$, let a be a quantitative attribute in $\mathscr{A} \cup \{d\}$. To express the approximate equality between two objects w.r.t. a, in this paper we use the fuzzy relation R_a from [24], defined by, for x and y in X (σ_a denotes the standard deviation of a):

$$R_a(x,y) = \max\left(\min\left(\frac{a(y)-a(x)+\sigma_a}{\sigma_a}, \frac{a(x)-a(y)+\sigma_a}{\sigma_a}\right), 0\right) \tag{15}$$

Assuming that for a qualitative (i.e., nominal) attribute a, the classical way of discerning objects is used, i.e., $R_a(x,y) = 1$ if $a(x) = a(y)$ and $R_a(x,y) = 0$ otherwise, we can define, for any subset B of $\mathscr{A}$, the fuzzy B-indiscernibility relation by

$$R_B(x,y) = \mathscr{T}(\underbrace{R_a(x,y)}_{a\in B}) \tag{16}$$

in which $\mathscr{T}$ represents a t-norm. It can easily be seen that if only qualitative attributes (possibly originating from discretization) are used, then the traditional concept of B-indiscernibility relation is recovered. It should also be noted that Eq. (15) is not the only possibility to define $R_a(x,y)$ and that it is an ongoing research to adjust fuzzy relations to real-life data.

For the lower and upper approximation of a fuzzy set A in X by means of a fuzzy tolerance relation R, we adopt the definitions proposed by Radzikowska and Kerre in [40]: given an implicator $\mathscr{I}$ and a t-norm $\mathscr{T}$, they paraphrased formulas (2) and (3) to define[5] $R \downarrow A$ and $R \uparrow A$ by

$$(R \downarrow A)(y) = \inf_{x\in X} \mathscr{I}(R(x,y), A(x)), \tag{17}$$

$$(R \uparrow A)(y) = \sup_{x\in X} \mathscr{T}(R(x,y), A(x)) \tag{18}$$

for all y in X.

3. Fuzzy-rough attribute reduction

In this section, we extend the framework for rough set analysis described in Section 2.1 using concepts of fuzzy set theory, to deal with quantitative attributes more appropriately. In order to do so, we introduce a number of increasing, $[0,1]$-valued measures to evaluate subsets of $\mathscr{A}$ w.r.t. their ability to maintain discernibility relative to the decision attribute and to generate adequate decision rules. Once such a measure, say $\mathscr{M}$, is obtained, we can associate a notion of fuzzy decision reduct with it.

Definition 1 (*Fuzzy $\mathscr{M}$-decision reduct*). Let $\mathscr{M}$ be a monotonic $\mathscr{P}(\mathscr{A}) \to [0,1]$ mapping such that $\mathscr{M}(\mathscr{A}) = 1, B \subseteq \mathscr{A}$ and $0 < \alpha \leqslant 1$. B is called a fuzzy $\mathscr{M}$-decision superreduct to degree α if $\mathscr{M}(B) \geqslant \alpha$. It is called a fuzzy $\mathscr{M}$-decision reduct to degree α if moreover for all $B' \subset B, \mathscr{M}(B') < \alpha$.

Below, we outline two important approaches to obtain such fuzzy decision reducts. Specifically, in Section 3.1, we extend the notion of positive region, while in Section 3.2 we introduce a fuzzy discernibility function. In Section 3.3, we investigate the relationships between these approaches. Throughout this section, we assume that R_B, the fuzzy relation that provides the means to evaluate to what extent objects are indiscernible w.r.t. the attributes of $B \subseteq \mathscr{A}$, is defined by Eq. (16). On the other hand, as already noted before, it is not the only possibility to introduce fuzzy relations for real-life data.

[4] It should be mentioned that many authors impose an additional requirement of $\mathscr{T}$-transitivity, i.e., given a t-norm $\mathscr{T}$.

$$\mathscr{T}(R(x,y), R(y,z)) \leqslant R(x,z)$$

should hold for any x, y and z in X; R is then called a fuzzy $\mathscr{T}$-equivalence relation, or similarity relation. While $\mathscr{T}$-equivalence relations naturally extend the transitivity of their classical counterparts, they may exhibit some undesirable effects, which were pointed out e.g. in [9,10].

[5] Note that when X is finite (as will always be the case in the context of decision systems), inf and sup can be replaced with min and max, respectively. We will use both notations interchangeably in this paper.

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

3.1. Fuzzy positive region

Using fuzzy B-indiscernibility relations, we can define the fuzzy B-positive region by, for y in U,

$$POS_B(y) = \left(\bigcup_{x\in X} R_B \downarrow R_d x\right)(y). \tag{19}$$

This means that the fuzzy positive region is a fuzzy set in X, to which an object y belongs to the extent that its R_B-foreset is included into *at least one* of the decision classes. The following proposition shows that when the decision attribute d is qualitative, only the decision class that y belongs to needs to be inspected.

Proposition 1. *For $y \in X$, if R_d is a crisp relation,*

$$POS_B(y) = (R_B \downarrow R_d y)(y).$$

Proof. We find

$$\begin{aligned} POS_B(y) &= \max_{x\in X} \inf_{z\in X} \mathscr{I}(R_B(z,y), R_d(z,x)) = \max\left(\max_{x\in R_d y} \inf_{z\in X} \mathscr{I}(R_B(z,y), R_d(z,x)), \max_{x\notin R_d y} \inf_{z\in X} \mathscr{I}(R_B(z,y), R_d(z,x))\right) \\ &= \max\left(\max_{x\in R_d y} \inf_{z\in X} \mathscr{I}(R_B(z,y), R_d(z,x)), 0\right) = \inf_{z\in X} \mathscr{I}(R_B(z,y), R_d(z,y)) = (R_B \downarrow R_d y)(y), \end{aligned}$$

where we used $R_B(y,y) = 1, \mathscr{I}(1,0) = 0$ and the fact that $R_d(z,x) = R_d(z,y)$ when $x \in R_d y$. □

Example 4. Let us come back to the undiscretized decision system in Ex. 1. Using Eqs. (15) and (16) with $\mathscr{T} = \mathscr{T}_L$ to compute approximate equality, and $\mathscr{I} = \mathscr{I}_L$ in (19), we can calculate the fuzzy positive region for $B = \{a_4, a_5\}$. For instance, since $\sigma_{a_4} = 16.385$ and $\sigma_{a_5} = 131.176$,

$$\begin{aligned} POS_B(x_3) &= \inf_{x\in X} \mathscr{I}(R_B(x,x_3), R_d x_3(x)) = \inf_{x\in X} \mathscr{I}(R_B(x,x_3), X_0(x)) = \min(1, 1 - R_B(x_2,x_3), 1, 1 - R_B(x_4,x_3), 1, 1, 1) \\ &= 1 - \max(R_B(x_2,x_3), R_B(x_4,x_3)) = 1 - \max(0, R_{a_4}(x_2,x_3) + R_{a_5}(x_2,x_3) - 1, R_{a_4}(x_4,x_3) + R_{a_5}(x_4,x_3) - 1) \\ &= 1 - \max(0, 0.512 + 0.680 - 1, 0.817 + 0.710 - 1) = 0.473 \end{aligned}$$

The complete result is

$$POS_B = \{(x_1,1), (x_2,0.808), (x_3,0.473), (x_4,0.473), (x_5,1), (x_6,1), (x_7,1)\}.$$

Compare this with Ex. 1, where POS_B was computed for the discretized system: the fuzzy positive region allows gradual membership values, and hence is able to express that e.g. x_2 is a less problematic object than x_3 and x_4. Finally, it can also be verified that, with the given parameters, $POS_{\mathscr{A}} = X$ still holds.

Now assume that d is quantitative. In this case, to each object x in X, a fuzzy tolerance class $R_d x$ is associated, and for different objects these classes may be partially overlapping. Unfortunately, in this case Proposition 1 no longer holds: $POS_B(y)$ is at least equal to $(R_B \downarrow R_d y)(y)$, but because of the partial overlapping between decision classes, it is possible that a higher value is obtained for $x \neq y$, as the following example illustrates.

Example 5. Consider the following decision system[6] with seven objects, 13 conditional attributes and a quantitative decision attribute:

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	d
x_1	0.088	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
x_2	3.321	0.0	19.58	1	0.871	5.403	100.0	1.3216	5	403	14.7	396.9	26.82	13.4
x_3	2.149	0.0	19.58	0	0.871	5.709	98.5	1.6232	5	403	14.7	261.9	15.79	19.4
x_4	1.414	0.0	19.58	1	0.871	6.129	96.0	1.7494	5	403	14.7	321.0	15.12	17.0
x_5	0.084	45.0	3.44	0	0.437	7.185	38.9	4.5667	5	398	15.2	396.9	5.39	34.9
x_6	0.035	95.0	2.68	0	0.416	7.853	33.2	5.1180	4	224	14.7	392.8	3.81	48.5
x_7	0.106	30.0	4.93	0	0.428	6.095	65.1	6.3361	6	300	16.6	394.6	12.40	20.1

We use the same parameters as in Ex. 4. For $B = \{a_3\}$, we get e.g.,

$$(R_B \downarrow R_d x_2)(x_2) = \inf_{z\in X} \mathscr{I}(R_B(z,x_2), R_d(z,x_2)) = \min(1,1,0.512,0.707,1,1,1) = 0.512$$

[6] This is a sample taken from the Boston Housing data set located at the UCI Machine Learning repository, available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

but on the other hand,

$$(R_B \downarrow R_d x_4)(x_2) = \inf_{z \in X} \mathscr{I}(R_B(z,x_2), R_d(z,x_4)) = \min(1, 0.707, 0.805, 1, 1, 1, 1) = 0.707.$$

As a consequence,

$$POS_B(x_2) = \max_{x \in X} \inf_{z \in X} \mathscr{I}(R_B(z,x_2), R_d(z,x)) = \max(0.228, 0.512, 0.512, 0.707, 0, 0, 0.455) = 0.707 > (R_B \downarrow R_d x_2)(x_2).$$

While formula (19) provides the most faithful way to define the fuzzy positive region, it is not the most practically useful one in this case, since the computational complexity is high (cubic in the number of objects for computing the entire positive region). Therefore we may opt to replace it by

$$POS'_B(y) = (R_B \downarrow R_d y)(y), \tag{20}$$

which results in smaller positive regions (as shown above), that are easier to compute (quadratic complexity in the number of objects for computing the entire positive region).

Example 6. Continuing Ex. 5, it holds that

$$POS_B = \{(x_1, 0.571), (x_2, 0.512), (x_3, 0.512), (x_4, 0.707), (x_5, 0.094), (x_6, 0.094), (x_7, 0.184)\},$$

when (19) is used, and

$$POS'_B = \{(x_1, 0.572), (x_2, 0.707), (x_3, 0.707), (x_4, 0.707), (x_5, 0.094), (x_6, 0.094), (x_7, 0.209)\},$$

when (20) is used.

Once we have fixed the fuzzy positive region, we can define an increasing $[0,1]$-valued measure to implement a corresponding notion of fuzzy decision reducts. The most obvious way is to introduce a normalized[7] extension of the degree of dependency, i.e.

$$\gamma_B = \frac{|POS_B|}{|POS_{\mathscr{A}}|} \quad \text{and} \quad \gamma'_B = \frac{|POS'_B|}{|POS'_{\mathscr{A}}|}. \tag{21}$$

These measure resemble the one introduced by Jensen and Shen in [23]. Rather than considering an average of the membership degrees to the B-positive region, it is also possible to focus on the most problematic element. This is reflected by the following measures:

$$\delta_B = \frac{\min_{x \in X} POS_B(x)}{\min_{x \in X} POS_{\mathscr{A}}(x)} \quad \text{and} \quad \delta'_B = \frac{\min_{x \in X} POS'_B(x)}{\min_{x \in X} POS'_{\mathscr{A}}(x)}. \tag{22}$$

They reflect the extent to which *all* objects can still be classified correctly, when only the attributes in B are considered. Their use is inspired by the fact that in standard rough set theory, the property of being a (super)reduct is also determined by the worst object. The following easily verified propositions show that these measures can indeed be used to define fuzzy decision reducts.

Proposition 2. For subsets B_1, B_2 of $\mathscr{A}$,

$$B_1 \subseteq B_2 \Rightarrow \begin{cases} \gamma_{B_1} \leqslant \gamma_{B_2} & \text{and} \quad \gamma'_{B_1} \leqslant \gamma'_{B_2}, \\ \delta_{B_1} \leqslant \delta_{B_2} & \text{and} \quad \delta'_{B_1} \leqslant \delta'_{B_2}. \end{cases}$$

Proposition 3. $\gamma_{\mathscr{A}} = \gamma'_{\mathscr{A}} = \delta_{\mathscr{A}} = \delta'_{\mathscr{A}} = 1$.

Example 7. For the Pima decision system from Example 4, it is easy to verify that for $B = \{a_4, a_5\}, \gamma_B = 0.822$, while $\delta_B = 0.473$. For the Housing decision system from Examples 5 and 6, note that $POS_{\mathscr{A}} = X$ (and hence also $POS'_{\mathscr{A}} = X$). Given $B = \{a_3\}, \gamma_B = 0.441$, while $\gamma'_B = 0.382$. On the other hand, $\delta_B = \delta'_B = 0.094$.

3.2. Fuzzy discernibility function

The fuzzy tolerance relations that represent objects' approximate equality can be used to redefine the discernibility function (7) as an $\{0,1\}^m \to [0,1]$ mapping, such that, for each combination of conditional attributes, a value between 0 and 1 is obtained, indicating how well these attributes maintain the discernibility, relative to the decision attribute, between all objects.

[7] Normalization is required in order that the measure yields a value of 1 for the whole attribute set. In this paper, we assume $POS_{\mathscr{A}}(x) > 0$ for every x in X.

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

In order to obtain such a generalization, first note that Eq. (7) can be rewritten as

$$\begin{aligned} f(a_1^*,\ldots,a_m^*) &= \bigwedge\left\{\bigvee_{k=1}^{m} a_k^*[d(x_i)\neq d(x_j) \Rightarrow a_k(x_i)\neq a_k(x_j)] | 1 \leqslant i < j \leqslant n\right\} \\ &= \bigwedge\left\{\bigvee_{k=1}^{m} a_k^*[a_k(x_i) = a_k(x_j) \Rightarrow d(x_i) = d(x_j)] | 1 \leqslant i < j \leqslant n\right\} \\ &= \bigwedge\left\{\left[\bigwedge_{a_k^*=1} (a_k(x_i) = a_k(x_j))\right] \Rightarrow d(x_i) = d(x_j) | 1 \leqslant i < j \leqslant n\right\} \end{aligned} \tag{23}$$

provided the decision system is consistent.[8]

Interpreting the connectives in Eq. (23) by a t-norm $\mathscr{T}$ and an implicator $\mathscr{I}$, and replacing the exact equalities by the respective approximate equalities (fuzzy indiscernibility relations), we can extend the discernibility function to a $\{0,1\}^m \to [0,1]$ mapping in the following way:

$$f(a_1^*,\ldots,a_m^*) = \mathscr{T}\left(\underbrace{c_{ij}(a_1^*,\ldots,a_m^*)}_{1\leqslant i<j\leqslant n}\right), \tag{24}$$

with

$$c_{ij}(a_1^*,\ldots,a_m^*) = \mathscr{I}\left(\mathscr{T}\left(\underbrace{R_{a_k}(x_i,x_j)}_{a_k^*=1}\right), R_d(x_i,x_j)\right). \tag{25}$$

By the definition of an implicator, this means that the degree to which an attribute a_k serves to distinguish between objects x_i and x_j increases as their approximate equality $R_{a_k}(x_i,x_j)$ w.r.t. a_k decreases, and their approximate equality $R_d(x_i,x_j)$ w.r.t. d increases. If R_{a_k} and R_d are crisp, the traditional format (23) is regained (again, assuming consistency).

Referring again to the valuation $\mathscr{V}_B$ corresponding to a subset B of $\mathscr{A}$, $\mathscr{V}_B(f(a_1^*,\ldots,a_m^*))$ is now a value between 0 and 1 that expresses the degree to which, for all object pairs, different values in attributes of B correspond to different values of d. Based on this, we introduce the following normalized subset evaluation measure:

$$f_B = \frac{\mathscr{V}_B(f(a_1^*,\ldots,a_m^*))}{\mathscr{V}_{\mathscr{A}}(f(a_1^*,\ldots,a_m^*))}. \tag{26}$$

Alternatively, rather than taking a minimum operation in Eq. (24), one can also consider the average over all object pairs, i.e.,

$$g(a_1^*,\ldots,a_m^*) = \frac{2\cdot\sum_{1\leqslant i<j\leqslant n} c_{ij}(a_1^*,\ldots,a_m^*)}{n(n-1)} \tag{27}$$

This formula exhibits a less rigid behaviour than Eq. (24), which yields 0 as soon as one of the c_{ij} equals 0. Analogously to f_B, the associated measure is given by

$$g_B = \frac{\mathscr{V}_B(g(a_1^*,\ldots,a_m^*))}{\mathscr{V}_{\mathscr{A}}(g(a_1^*,\ldots,a_m^*))}. \tag{28}$$

The following two propositions express that the measures we have defined are monotonic, and that they assume the value 1 when all the attributes are considered, which makes it possible to consider fuzzy f- and g-decision reducts.

Proposition 4. *For subsets B_1, B_2 of $\mathscr{A}$,*

$$B_1 \subseteq B_2 \Rightarrow \begin{cases} f_{B_1} \leqslant f_{B_2}, \\ g_{B_1} \leqslant g_{B_2}. \end{cases}$$

Proposition 5. $f_{\mathscr{A}} = g_{\mathscr{A}} = 1$.

Example 8. We first take up the undiscretized Pima decision system from Example 1. Using Eq. (15) to compute approximate equality, and $\mathscr{T} = \mathscr{T}_L, \mathscr{I} = \mathscr{I}_L$ in Eqs. (24) and (25),

$$\begin{aligned} f_B &= \frac{\mathscr{V}_B(f(a_1^*,\ldots,a_m^*))}{\mathscr{V}_{\mathscr{A}}(f(a_1^*,\ldots,a_m^*))} = \frac{\mathscr{T}_{1\leqslant i<j\leqslant 7} c_{ij}(0,0,0,1,1,0,0,0)}{\mathscr{T}_{1\leqslant i<j\leqslant 7} c_{ij}(1,1,1,1,1,1,1,1)} \\ &= \frac{\mathscr{T}(1,1,1,1,1,1,0.808,1,1,1,1,0.473,1,1,1,1,1,1,1,1,1)}{1} = 0.281, \end{aligned}$$

[8] Recall that if $(X, \mathscr{A}\cup\{d\})$ is inconsistent, there exist x_i and x_j such that $(\forall a\in\mathscr{A})(a(x_i)=a(x_j))$, yet $d(x_i)\neq d(x_j)$. Such x_i and x_j are not considered in Eq. (7), since $O_{ij}=\emptyset$.

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

$$g_B = \frac{\mathscr{V}_B(g(a_1^*,\ldots,a_m^*))}{\mathscr{V}_{\mathscr{A}}(g(a_1^*,\ldots,a_m^*))} = \frac{\sum_{1\leqslant i<j\leqslant 7} c_{ij}(0,0,0,1,1,0,0,0)}{21} = \frac{20.281}{21} = 0.966.$$

Next, consider again the Housing data set from Example 5. Using the same parameters as above for $B = \{a_3\}$, we obtain

$$f_B = \mathscr{T}_{1\leqslant i<j\leqslant 7} c_{ij}(0,0,1,0,0,0,0,0) = \mathscr{T}(1,1,1,0.572,0.641,1,0.512,0.707,1,1,1,0.805,1,1,1,1,1,1,0.094,0.184,0.278) = 0,$$
$$g_B = \frac{\sum_{1\leqslant i<j\leqslant 7} c_{ij}(0,0,1,0,0,0,0,0)}{21} = \frac{16.792}{21} = 0.800.$$

3.3. Relationships between fuzzy decision reducts

As we have shown, the evaluation measures $\gamma, \gamma', \delta, \delta', f$ and g introduced in the previous subsections all give rise to corresponding fuzzy decision reducts. However, not all of them are independent: e.g., $\delta'_B \leqslant \gamma'_B \leqslant \gamma_B$ and $\delta'_B \leqslant \delta_B \leqslant \gamma_B$ always hold, and $\gamma_B = \gamma'_B$ and $\delta_B = \delta'_B$ when the decision attribute is qualitative. Moreover, a number of interesting relationships hold between the approaches based on the fuzzy positive region and those based on the fuzzy discernibility function, which are summed up by the following propositions; we assume that the same t-norm $\mathscr{T}$ and implicator $\mathscr{I}$ are used in Eqs. (16), (17), (24) and (25).

Proposition 6. If $POS'_{\mathscr{A}} = X$,

$$f_B \leqslant \delta'_B \quad \text{and} \quad \gamma'_B \leqslant g_B \tag{29}$$

for $B \subseteq \mathscr{A}$. Moreover, in case $\mathscr{T} = \mathscr{T}_M, f_B = \delta'_B$, regardless of $POS'_{\mathscr{A}} = X$.

Proof

$$f_B = \frac{\mathscr{V}_B(f(a_1^*,\ldots,a_m^*))}{\mathscr{V}_{\mathscr{A}}(f(a_1^*,\ldots,a_m^*))} \leqslant \mathscr{V}_B(f(a_1^*,\ldots,a_m^*)) = \mathscr{T}(\underbrace{\mathscr{I}(R_B(x_i,x_j),R_d(x_i,x_j))}_{1\leqslant i<j\leqslant n}) \leqslant \min_{1\leqslant i<j\leqslant n} \mathscr{I}(R_B(x_i,x_j),R_d(x_i,x_j))$$
$$= \min_{x,y\in X} \mathscr{I}(R_B(x,y),R_d(x,y)) = \min_{y\in X}(R_B \downarrow R_d y)(y) = \min_{y\in X} POS'_B(y) = \delta'_B,$$

where we have used the fact that min is the largest t-norm, and that it is symmetric and idempotent. When $\mathscr{T} = \mathscr{T}_M$, it is clear that $\mathscr{V}_B(f(a_1^*,\ldots,a_m^*)) = \min_{y\in X} POS'_B(y)$ for each $B \subseteq \mathscr{A}$, hence it also holds that $\mathscr{V}_{\mathscr{A}}(f(a_1^*,\ldots,a_m^*)) = \min_{y\in X} POS'_{\mathscr{A}}(y)$. This completes the proof of $f_B = \delta'_B$. To see that $\gamma'_B \leqslant g_B$,

$$\gamma'_B = \frac{\sum_{y\in X}(R_B \downarrow R_d y)(y)}{n} = \frac{\sum_{y\in X} \inf_{x\in X} \mathscr{I}(R_B(x,y),R_d(x,y))}{n} = \frac{\sum_{1\leqslant j\leqslant n} \inf_{x\in X} \mathscr{I}(R_B(x,x_j),R_d(x,x_j))}{n}$$
$$\leqslant \frac{2\sum_{1\leqslant i<j\leqslant n} \mathscr{I}(R_B(x_i,x_j),R_d(x_i,x_j))}{n(n-1)} = g_B \qquad \square$$

The above proof shows that f and δ are essentially built upon the same idea, with some variations due to the parameter choice, and also reveals the essential difference between γ and g: while the former looks at the lowest value of the formula $\mathscr{I}(R_B(x,y),R_d(x,y))$ for each y (reflecting to what extent there exists an x that has similar values for all the attributes in B, but a different decision), and averages over these values, the latter evaluates all pairwise evaluations of this formula.

The following proposition shows that, for consistent data, a crisp g-decision reduct is always a crisp γ/γ'-decision reduct.

Proposition 7. If $POS'_{\mathscr{A}} = X$,

$$g_B = 1 \Rightarrow (\gamma'_B = 1 \text{ and } \gamma_B = 1) \tag{30}$$

for any $B \subseteq \mathscr{A}$.

Proof

$$g_B = 1 \Rightarrow \frac{2\sum_{1\leqslant i<j\leqslant n} \mathscr{I}(R_B(x_i,x_j),R_d(x_i,x_j))}{n(n-1)} = 1$$
$$\Rightarrow (\forall 1 \leqslant i < j \leqslant n)(\mathscr{I}(R_B(x_i,x_j),R_d(x_i,x_j)) = 1)$$
$$\Rightarrow (\forall 1 \leqslant j \leqslant n)\left(\inf_{x\in X} \mathscr{I}(R_B(x,x_j),R_d(x,x_j)) = 1\right)$$
$$\Rightarrow \gamma'_B = 1.$$

Since $\gamma'_B \leqslant \gamma_B, \gamma_B = 1$ holds as well. $\square$

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

Example 9. Some of the relationships discussed above are illustrated in the following table, which contains the values obtained with $\mathscr{T}_L$ and $\mathscr{I}_L$ in the previous sections ($B = \{a_4, a_5\}$ for pima, $B = \{a_3\}$ for housing), along with those obtained with $\mathscr{T}_M$ and $\mathscr{I}_M$:

Data set	Connectives	γ_B	γ'_B	δ_B	δ'_B	f_B	g_B
Pima	$\mathscr{T}_M, \mathscr{I}_M$	0.72	0.72	0.29	0.29	0.29	0.94
	$\mathscr{T}_L, \mathscr{I}_L$	0.82	0.82	0.47	0.47	0.28	0.97
Housing	$\mathscr{T}_M, \mathscr{I}_M$	0.43	0.38	0.09	0.09	0.09	0.79
	$\mathscr{T}_L, \mathscr{I}_L$	0.44	0.38	0.09	0.09	0	0.80

Note that there is no general pattern in the relationship between the results obtained with $\mathscr{T}_M$ and $\mathscr{I}_M$ and those with $\mathscr{T}_L$ and $\mathscr{I}_L$; in this particular example, the latter connectives result in a higher evaluation for all measures, except for f, but this does not hold in general.

4. Experimental analysis

To evaluate the use of the various fuzzy decision reduct instances that we have introduced in this paper, we have run a series of classification and regression experiments on a number of benchmark data sets whose characteristics are summarized in Table 1; $|V_d|$ denotes the number of decision classes (only for data sets with a qualitative decision attribute). Like the PIMA and HOUSING samples used as running examples in the previous sections, several of them are taken from the UCI Machine Learning repository. WATER 2 and WATER 3 are derived from UCI's water quality dataset, with the decision feature values collapsed to 2 or 3 classes representing the overall state of the system behaviour. The WEB dataset is from [22], where the task was to classify web pages based on their content into one of several predefined categories. The ALGAE data sets[9] are provided by ERUDIT [14] and describe measurements of river samples for each of seven different species of alga, including river size, flow rate and chemical concentrations. The decision feature is the corresponding concentration of the particular alga. Finally, the CPU dataset is taken from Delve,[10] where the regression task is to predict the portion of time that CPUs run in user mode based on a number of computer system activity measures.

The general setup of these experiments is as follows: given a decision system $(X, \mathscr{A} \cup \{d\})$, a measure $\mathscr{M}$ as in Definition 1, and a threshold α ($0 < \alpha \leqslant 1$), we run a 10-fold cross validation experiment. In each iteration, we apply an adapted version of the QUICKREDUCT heuristic, shown in Fig. 1, to the training data to obtain a fuzzy $\mathscr{M}$-decision superreduct to degree α. All measures use Eq. (15) for evaluating attribute level discernibility, and, unless explicitly stated otherwise, the Łukasiewicz connectives $\mathscr{T}_L$ and $\mathscr{I}_L$ are used throughout the experiments.[11]

The quality of the obtained attribute subset B is then evaluated as the classification accuracy obtained by running a fixed classifier (respectively, the root mean square error obtained by running a fixed regression method, in case the decision attribute is quantitative) on the reduced test data. In our experiments, we have used the very simple K-nearest neighbour classifier [1], implemented in Weka [53] as IBK, with default parameters ($K = 1$, no distance weighting). This means that the method uses Euclidean distance to compute the closest neighbour in the training data, and outputs this object's decision as its prediction. The reason for using such a basic classifier like IBK is that we want to evaluate the intrinsic quality of the selected subsets of attributes, influenced by the choice of parameters in the definition of a fuzzy reduct, in isolation from the gain related to application of more advanced models. On the other hand, we obviously assume a usage of more advanced classifiers in real life applications, once we collect more experience with the phase of fuzzy reduct-based feature selection at this level of our research. We also intend to investigate other distance measures in IBK, possibly better adjusted to the way of searching for optimal fuzzy reducts (see also Section 5).

4.1. Cross-comparison between different measures

In the first set of experiments, for each of the data sets in Table 1 (excluding CPU and SPAMBASE which are used in the detailed analysis later), we compared the IBK classification/regression performance on the full attribute set to that obtained on versions reduced according to the different strategies in this paper. For each of the fuzzy-rough measures introduced, we ran QUICKREDUCT once with $\alpha = 1$, and a second time with a fixed $\alpha < 1$; in particular, a value of $\alpha = 0.95$ was deemed a suitable overall choice for most measures, except for g, which requires a much higher threshold, and for which $\alpha = 0.9999$ was selected. All measures use Łukasiewicz connectives, except f, which was found to perform better in combination with $\mathscr{T}_M$ and $\mathscr{I}_M$. Note that, by Proposition 6, this implementation of f coincides with the corresponding δ measure based on $\mathscr{T}_M$ and $\mathscr{I}_M$. In order to compare how well the methods perform against the state-of-the-art, correlation-based feature selection (CFS) [19] and Kohavi's wrapper subset evaluator (WSE) [26] were also run on the same data folds.

[9] See http://archive.ics.uci.edu/ml/datasets/Coil+1999+Competition+Data.

[10] Data for Evaluating Learning in Valid Experiments, see http://www.cs.toronto.edu/~delve/. The considered regression task corresponds to the CPU prototask on the COMP-ACTIV database.

[11] All evaluation measures described in this paper, along with the adapted QUICKREDUCT heuristic, have been implemented in Weka [53]. The program can be downloaded from http://users.aber.ac.uk/rkj/book/programs.php.

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

Table 1
Data set characteristics.

Data set	n	m	$\lvert V_d \rvert$	Origin
CLEVELAND	297	13	5	UCI
GLASS	270	13	7	UCI
IONOSPHERE	230	34	2	UCI
PIMA	392	8	2	UCI
SPAMBASE	4601	57	2	UCI
WATER 2	390	38	2	Adapted from UCI
WATER 3	390	38	3	Adapted from UCI
WEB	149	2556	5	[22]
WINE	178	13	3	UCI
ALGAE A → F	187	11	Quantitative	ERUDIT [14]
CPU	8192	21	Quantitative	Delve
HOUSING	506	13	Quantitative	UCI

(1) $B \leftarrow \{\}$
(2) **do**
(3) $\quad T \leftarrow B$, best $\leftarrow -1$
(4) $\quad$ **foreach** $a \in (\mathcal{A} \setminus B)$
(5) $\quad\quad$ **if** $\mathcal{M}(B \cup \{a\}) >$ best
(6) $\quad\quad\quad T \leftarrow B \cup \{a\}$, best $\leftarrow M(B \cup \{a\})$
(7) $\quad B \leftarrow T$
(8) **until** $\mathcal{M}(B) \geq \alpha$
(9) **return** B

Fig. 1. Modified QUICKREDUCT to obtain a fuzzy $\mathcal{M}$-decision superreduct to degree α.

Table 2
Classification results.

Data set	Unred.	γ		δ		f		g		CFS	WSE
		$\alpha = 1$	$\alpha = .95$	$\alpha = 1$	$\alpha = .95$	$\alpha = 1$	$\alpha = .95$	$\alpha = 1$	$\alpha = .9999$		
CLEVELAND	51.56	49.23	45.85	53.62	53.95	49.11	49.45	49.23	50.54	53.20	49.83
GLASS	70.52	70.52	70.00	70.06	70.04	73.33	73.33	69.59	70.06	78.92	69.50
IONOSPHERE	86.09	89.13	86.52	84.78	85.65	84.35	84.35	88.26	88.26	87.39	83.48
PIMA	73.45	71.15	75.24	72.93	73.70	72.44	72.19	71.65	72.20	69.41	68.13
WATER 2	84.10	81.54	82.56	80.00	80.00	83.33	83.08	82.56	79.74	86.15	77.18
WATER 3	81.54	78.97	77.18	78.21	78.21	79.74	79.49	80.77	79.74	83.85	75.64
WEB	40.95	42.95	42.86	37.52	37.52	35.48	34.81	34.90	32.24	53.71	47.05
WINE	94.97	92.71	91.57	91.57	91.57	97.22	96.63	91.05	93.86	96.08	94.38
CLEVELAND	13	7.70	5.00	8.80	8.70	11.50	11.40	7.70	6.00	6.90	4.50
GLASS	9	9.00	7.00	8.00	7.70	6.20	5.80	8.20	8.00	6.30	5.30
IONOSPHERE	34	7.10	5.10	9.20	9.00	12.10	11.60	6.90	5.90	10.80	4.10
PIMA	8	7.50	5.00	7.70	7.00	6.20	6.10	7.60	5.00	4.30	2.90
WATER 2	38	6.00	4.00	6.00	6.00	20.20	18.20	6.00	4.90	9.10	3.20
WATER 3	38	6.00	4.20	6.20	6.20	22.00	20.90	5.90	5.00	11.10	3.70
WEB	2556	18.40	14.90	30.10	30.10	42.90	41.40	16.00	15.00	57.80	8.50
WINE	13	5.00	3.90	4.90	4.90	10.50	9.50	4.80	4.00	10.90	4.60

The results are shown in Tables 2 and 3. When interpreting these results, one should always keep in mind the trade-off between accuracy (RMSE) and attribute subset size: a higher accuracy (lower RSME) is of course desirable, but so is a smaller subset size, i.e., the less conditional attributes there are in the reduced data set, the stronger its generalization capacity. Like this, it is clear that on ALGAE A, the overall best result is obtained for γ and γ' with $\alpha = 0.95$, since they have the highest accuracy and the lowest average attribute subset size across folds. On the other hand, for WATER 3, the accuracy for g and $\alpha = 0.9999$ is similar to that obtained on the unreduced data set and the one reduced with f, but the number of used conditional attributes is considerably less, arguably making the reduction by g the better option here.

The selected subsets for δ and f are generally longer, without necessarily being better than their γ and g counterparts, some notable exceptions notwithstanding, like the f results on GLASS and PIMA. The problems are especially visible on some of the larger data sets, like SPAMBASE and WEB, which are either poorly reduced, or hardly reduced at all. This behaviour is,

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

Table 3
Regression results.

Data set	Orig.	γ		γ'		δ		δ'		f		g		CFS	WSE
		$\alpha = 1$	$\alpha = .95$	$\alpha = 1$	$\alpha = .95$	$\alpha = 1$	$\alpha = .95$	$\alpha = 1$	$\alpha = .95$	$\alpha = 1$	$\alpha = .95$	$\alpha = 1$	$\alpha = .9999$		
ALGAE A	24.04	23.96	19.74	23.96	19.84	23.47	22.9	23.47	22.9	24.04	24.04	23.96	22.29	21.22	24.29
ALGAE B	17.4	15.46	12.95	15.84	13.63	14.57	14.3	16.11	16.15	14.06	14.05	15.33	14.76	12.04	13.05
ALGAE C	9.16	8.93	9.57	8.93	8.8	9.11	9.3	8.61	9.07	9.16	9.16	8.93	8.65	9.91	9.11
ALGAE D	4.72	3.52	3.83	3.52	4.43	3.6	3.57	3.68	3.68	4.47	4.44	3.39	3.55	3.26	3.59
ALGAE E	9.06	9.48	9.95	9.48	9.65	9.4	9.45	9.52	9.77	10.26	10.39	9.23	9.38	9.47	8.9
ALGAE F	13.3	12.63	14.24	13.16	14.85	12.91	13.02	12.92	12.74	12.98	13.01	13.73	13.87	15.9	13.44
ALGAE G	6.42	5.41	5.84	5.43	5.09	5.79	5.85	5.69	5.85	6.46	6.46	5.76	5.95	6.03	5.52
HOUSING	4.6	3.93	4.64	3.93	4.64	4.03	4.1	4.76	4.79	4.51	4.6	4.53	4.64	4.88	4.33
ALGAE A	11	10.7	5.9	10.7	5.8	9.8	8.7	9.8	8.7	11	11	10.7	7.4	7.7	5.6
ALGAE B	11	9.7	5	10.1	5	9.2	7.6	10.2	9.2	7.7	7.5	7.5	5.9	3.6	3.3
ALGAE C	11	10.8	5.4	10.8	6.2	9.1	7.3	10	8.6	11	11	10.8	7.3	2.4	5.7
ALGAE D	11	9.8	4.3	9.8	5	9.2	7.8	9.4	8.1	10.3	8.6	9.4	5.4	5.3	4.6
ALGAE E	11	9.4	5	9.5	5	9.6	8.3	9.3	8.5	8.6	8.3	7.7	5.9	4.1	5.6
ALGAE F	11	7.5	5	7.5	5	7.8	7.5	8	7.6	8.8	8.5	7.7	5.6	3.7	4.3
ALGAE G	11	8.2	4.9	8.6	5	8.9	7.6	8.8	8	9	8.8	7.8	5.9	1.8	1.6
HOUSING	13	8	5	8	5	6.9	6.3	5.5	5.2	12.1	11.7	11.5	5	4	8.6

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

to a large extent, due to the strictness of these measures: because they focus on the worst object in the data set, they tend to have zero values very often, especially for small attribute subsets. This affects QUICKREDUCT's operation adversely; when all of the considered subsets in a given iteration evaluate to 0, the heuristic is forced to select one without any information about its true merit. For the data sets with a quantitative decision, the results are more balanced, with less of the negative effects plaguing the operation of δ and f. Note also that in several cases (e.g. ALGAE A/B/C), the latter three measures manage little or no reduction when $\alpha = 1$ is selected, but yield good results when a slightly smaller threshold is selected, illustrating the use of fuzzy decision reducts.

A paired t-test was used to determine the statistical significance of the results at the 0.05 level. From this it was determined that, generally speaking, all the measures performed dimensionality reduction with no significant drop in accuracy (increase in RMSE). For the classification results, there is only one case where a fuzzy rough measure performs significantly worse than that of the unreduced data approach (g, with $\alpha = 0.9999$ for the WEB dataset). For the remaining measures, the performance cannot be said to be better or worse from a statistical viewpoint, even though a high proportion of features have been removed via these methods. This is also reflected in the regression results, where all measures perform equivalently to or better than the unreduced approach. Eight methods (γ with $\alpha = 1, \gamma$ with $\alpha = 0.95, \gamma'$ with $\alpha = 1, \gamma'$ with $\alpha = 0.95, \delta'$ with $\alpha = 1, g$ with $\alpha = 1, g$ with $\alpha = 0.9999$ and CFS) produce results that are statistically better than the unreduced approach for at least one dataset. Considering the measures themselves, when using fuzzy decision reducts ($\alpha < 1$) the resulting performance is almost always statistically equivalent to that of the corresponding crisp decision reduct methods ($\alpha = 1$). Again, this is achieved with substantially smaller subsets.

Finally, from the complexity point of view, it is interesting that regression results obtained with γ' and δ' are in general competitive with those of the more complex measures γ and δ, which justifies their simplification.

4.2. Detailed analysis on SPAMBASE and CPU

From the above results, it is clear that the selection of an adequate α threshold is not only dependent on the measure used but also on the data set. On the other hand, it may be argued that – just like membership degrees in a fuzzy set – the exact values of the measures are less important than the *partial ordering* they induce on attribute subsets. In particular, in keeping with the nature of QUICKREDUCT, we can rank individual attributes in the order in which they are added by this heuristic, and after each iteration evaluate the performance of the attribute subset constructed thus far. We have done this for SPAMBASE and CPU, using the measures γ and g. The results are listed in Tables 4 and 5. Each row in these tables records the size of the current subset, the attribute selected by QUICKREDUCT, the value of the measure and the accuracy (RMSE) of the reduced decision system.

In particular, the reduction of SPAMBASE by means of γ and g both resulted in 45 attributes being added before a corresponding crisp decision superreduct was obtained; attributes 3, 10, 14, 16, 28, 29, 31, 33, 37, 39, 42 and 46 do not belong to the final result for either of them. The order in which the attributes are added does differ, however. As can be seen in Fig. 2a, g's order of selection is markedly better, especially in the 10–25 subset size region, when the corresponding subset outperforms the one obtained with γ by several percents.

Table 4
Spambase results.

Size	Att.	γ	Acc. (%)	Att.	g	Acc. (%)	Size	Att.	γ	Acc. (%)	Att.	g	Acc. (%)
1	26	0.05	58.42	18	0.87	67.40	24	54	0.98	87.63	43	0.9999	90.33
2	24	0.10	64.92	20	0.95	74.90	25	6	0.98	89.02	54	0.9999	90.42
3	11	0.16	70.98	49	0.98	76.27	26	0	0.99	88.83	50	0.9999	90.24
4	18	0.26	76.09	56	0.991	79.27	27	48	0.992	89.05	48	0.9999	90.18
5	20	0.41	79.63	51	0.995	82.42	28	7	0.993	89.18	13	0.9999	90.07
6	56	0.55	83.22	11	0.997	82.46	29	23	0.994	89.48	2	0.9999	89.59
7	49	0.67	82.74	52	0.999	84.57	30	17	0.995	89.31	36	0.9999	89.63
8	51	0.77	83.37	26	0.998	84.87	31	34	0.995	89.35	47	0.9999	89.65
9	4	0.82	84.94	24	0.9991	85.81	32	47	0.996	89.39	25	0.9999	89.74
10	44	0.87	85.37	4	0.9993	86.16	33	25	0.996	89.51	17	0.9999	89.72
11	45	0.9	86.33	15	0.9995	87.39	34	22	0.997	90.13	34	0.9999	89.74
12	2	0.91	85.70	6	0.9996	88.79	35	5	0.997	89.76	53	0.9999	89.81
13	1	0.93	85.11	9	0.9997	88.72	36	21	0.998	89.89	5	0.9999	89.65
14	15	0.94	85.70	44	0.9997	89.05	37	53	0.998	89.92	12	0.9999	89.65
15	32	0.95	85.87	7	0.9998	89.55	38	8	0.998	89.76	30	0.9999	89.69
16	52	0.95	86.76	0	0.9998	89.39	39	38	0.998	89.78	40	0.9999	89.81
17	9	0.96	87.26	55	0.9998	89.35	40	19	0.999	90.07	22	0.9999	90.18
18	41	0.97	87.68	1	0.9999	89.59	41	40	0.999	90.13	38	0.9999	90.11
19	55	0.97	87.83	45	0.9999	89.46	42	12	0.999	89.98	21	0.9999	90.22
20	43	0.97	87.63	19	0.9999	89.76	43	30	0.999	90.07	35	0.9999	90.31
21	50	0.98	87.50	23	0.9999	90.07	44	35	0.999	90.22	8	0.9999	90.20
22	36	0.98	87.46	32	0.999	90.26	45	27	1	90.42	27	1	90.39
23	13	0.98	87.70	41	0.999	90.52							

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

Table 5
CPU results.

Size	Att.	γ	RMSE	Att.	g	RMSE
1	20	0.06	10.20	2	0.92	22.30
2	17	0.48	5.18	17	0.98	22.53
3	2	0.76	4.05	20	0.999	4.05
4	7	0.91	3.9	8	0.9998	3.78
5	8	0.97	3.75	19	0.9999	3.65
6	19	0.99	3.66	7	0.9999	3.66
7	4	0.99	3.6	4	0.9999	3.60
8	3	0.998	3.6	3	0.9999	3.60
9	0	0.999	3.52	0	0.9999	3.52
10	14	0.9995	3.39	14	0.9999	3.39
11	16	0.9996	3.35	15	0.9999	3.40
12	15	0.9997	3.34	16	0.9999	3.34
13	1	0.9998	3.38	10	0.9999	3.48
14	10	0.9998	3.53	1	0.9999	3.52
15	18	0.9999	3.52	18	0.9999	3.52
16	6	0.9999	3.43	13	0.9999	3.52
17	13	0.9999	3.44	6	0.9999	3.44
18	5	0.9999	3.4	9	0.9999	3.53
19	9	0.9999	3.51	5	0.9999	3.51
20	11	1	3.49	11	1	3.49

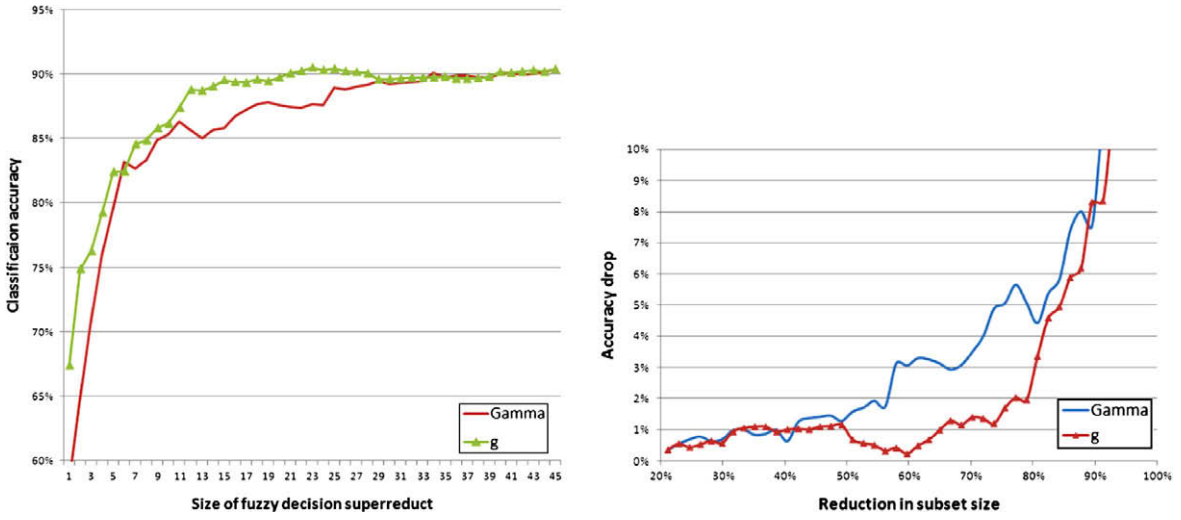

Fig. 2. SPAMBASE results: (a) classification accuracy versus subset size and (b) accuracy drop versus subset size reduction.

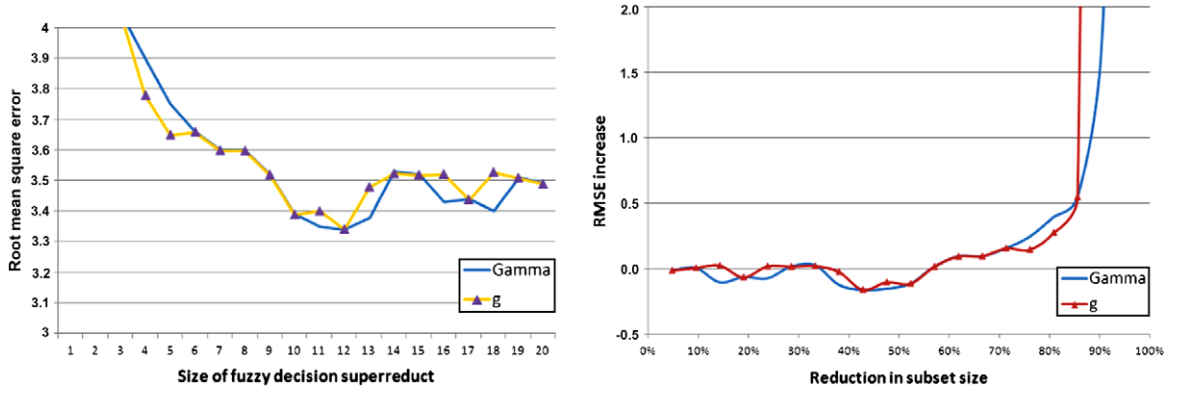

Fig. 3. CPU results: (a) RMSE versus subset size (b) RMSE increase versus subset size reduction.

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

The most important observation about these results, however, is the obvious benefit fuzzy decision reducts have over crisp ones. For instance, a crisp γ-decision reduct contains 45 out of 57 conditional attributes and achieves a classification accuracy of 90.42% (compared to 90.76% obtained for the full data set), in other words, a 21% reduction in the number of attributes versus a 0.34% drop in accuracy; compare this to the 23-attribute subset obtained by g which manages a 60% reduction at an even smaller 0.24% accuracy loss. Moreover, as seen in Fig. 2b, if a 1% accuracy drop is permissible, fuzzy γ-decision reducts manage to reduce the subset size by over 40%, while with g a reduction of the data set by more than 63% is possible.

Table 4 also reveals that the selection of $\alpha = 0.95$ for γ and $\alpha = 0.9999$ in the experiment of Section 4.1 was probably a bit too low; this again stresses the relative, rather than absolute, importance of this threshold, which should always be tuned in relation to the size of the obtained subset.

The results for CPU, given in Table 5 and Fig. 3, show a largely similar picture. In this case, the advantage of fuzzy decision reducts is even more evident: γ and g both yield a 20-attribute crisp decision reduct (5% size reduction, only attribute 12 is not selected), but each of them can get a better RMSE using only 10 attributes (52% size reduction). The difference between the γ and g results is smaller than in the SPAMBASE experiment, with a small advantage of γ over g.

5. Conclusion and future work

In this paper, we have introduced a framework for fuzzy-rough set based feature selection, built up around the formal notion of a fuzzy decision reduct. By expressing that an attribute subset should retain the quality of the full feature set up to a certain extent only, we are able to generate shorter attribute subsets, without paying a corresponding price in subset quality (evaluated by means of the corresponding classification accuracy or RMSE).

At the same time, we have provided a comprehensive typology of subset evaluation measures that can be used to define fuzzy decision reducts, and that take into account the gradual nature of objects' discernibility. We have shown that, while these measures come in various different shapes, with many variations possible due to the choice of connectives and other characteristics like how to define the positive region (e.g., γ versus γ'), a common thread running through all of them is the question whether objects that have (sufficiently) similar conditional attributes, also have (sufficiently) similar decisions. The main differences between the proposed measures lie in the strictness with which they enforce this criterion: δ and f focus on its worst single violation within the data, γ makes an average assessment of individual objects' performance, while g simply averages over all pairwise evaluations of the criterion.

Our experiments clearly endorse the benefit of using fuzzy decision reducts, showing a greater flexibility and better potential to produce good-sized, high-quality attribute subsets than the crisp decision reducts that have been used so far in fuzzy-rough data analysis. At the same time, these experiments also raise the challenge of measure selection and parameter optimization. While some generic guidelines can be given and some general observations apply (like the fact that the δ/f measure is typically too strict for realistic data), different data sets require different parameter settings for optimal performance.

On the other hand, this unpredictability may also be due in part to the gap that still exists between the attribute reduction procedure and its evaluation by means of classification or regression; an interesting proposition, therefore, would be to adapt the IBK classifier such that it uses the same approximate equality/distance metric (viz. based on Eqs. (15) and (16)) as in our approaches, instead of the currently used Euclidean distance.

Finally, in view of the different behaviour of different types of measures, it may be worthwhile to combine their characteristics into aggregated measures (e.g., weighted averages), or to allow the heuristic to use different measures at different iterations (e.g., QUICKREDUCT could start by adding attributes based on γ, then at some point shift to δ to fine-tune the result).

Acknowledgments

Chris Cornelis would like to thank the Research Foundation – Flanders for funding his research. Dominik Ślęzak was partially supported by the Grant Nos. N516 368334 and N516 077837 from the Ministry of Science and Higher Education of the Republic of Poland.

References

[1] D. Aha, Instance-based learning algorithm, Machine Learning 6 (1991) 37–66.
[2] J.G. Bazan, H.S. Nguyen, S.H. Nguyen, P. Synak, J. Wróblewski, Rough set algorithms in classification problem, Rough Set Methods and Applications. New Developments in Knowledge Discovery in Information Systems. Studies in Fuzziness and Soft Computing, vol. 56, Physica-Verlag, 2000, pp. 49–88.
[3] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.
[4] D. Chen, E. Tsang, S. Zhao, An approach of attributes reduction based on fuzzy rough sets, in: Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics, 2007, pp. 486–491.
[5] D. Chen, E. Tsang, S. Zhao, Attribute reduction based on fuzzy rough sets, in: Proc. Int. Conf. on Rough Sets and Intelligent Systems Paradigms, 2007, pp. 73–89.
[6] A. Chouchoulas, Q. Shen, Rough set-aided keyword reduction for text categorisation, Applied Artificial Intelligence 15 (9) (2001) 843–873.
[7] C. Cornelis, M. De Cock, A.M. Radzikowska, Fuzzy rough sets: from theory into practice, in: W. Pedrycz, A. Skowron, V. Kreinovich (Eds.), Handbook of Granular Computing, John Wiley and Sons, 2008, pp. 533–552.

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008

[8] C. Cornelis, G. Hurtado Martín, R. Jensen, D. Ślęzak, Feature selection with fuzzy decision reducts, in: Proc. 3rd Int. Conf. on Rough Sets and Knowledge Technology (RSKT2008), 2008, pp. 284–291.
[9] M. De Cock, E.E. Kerre, On (un)suitable fuzzy relations to model approximate equality, Fuzzy Sets and Systems 133 (2) (2003) 137–153.
[10] M. De Cock, C. Cornelis, E.E. Kerre, Fuzzy rough sets: the forgotten step, IEEE Transactions on Fuzzy Systems 15 (1) (2007) 137–153.
[11] D. Chen, C.Z. Wang, Q.H. Hu, A new approach to attribute reduction of consistent and inconsistent covering decision systems with covering rough sets, Information Sciences 17 (1) (2007) 3500–3518.
[12] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, International Journal of General Systems 17 (1990) 191–209.
[13] D. Dubois, H. Prade, Putting rough sets and fuzzy sets together, in: S.Y. Huang, (Ed.), Intelligent Decision Support, 1992, pp. 203–232.
[14] European Network for Fuzzy Logic and Uncertainty Modelling in Information Technology (ERUDIT), Protecting rivers and streams by monitoring chemical concentrations and algae communities, Computational Intelligence and Learning (CoIL) Competition, 1999.
[15] B. Ganter, R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer, 1998.
[16] S. Greco, B. Matarazzo, R. Slowiński, Fuzzy similarity relation as a basis for rough approximations, in: Proc. 1st Int. Conf. on Rough Sets and Current Trends in Computing (RSCTC1998), 1998, pp. 283–289.
[17] S. Greco, B. Matarazzo, R. Slowiński, Rough set processing of vague information using fuzzy similarity relations, in: C.S. Calude, G. Paun (Eds.), Finite Versus Infinite – Contributions to an Eternal Dilemma, Springer-Verlag, 2000, pp. 149–173.
[18] S. Greco, B. Matarazzo, R. Slowiński, Rough approximation by dominance relations, International Journal of Intelligence System 17 (2002) 153–171.
[19] M.A. Hall, Correlation-based feature selection for machine learning. PhD thesis, The University of Waikato, 1999.
[20] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer, 2009.
[21] Q.H. Hu, X.Z. Xie, D.R. Yu, Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognition 40 (12) (2007) 3509–3521.
[22] R. Jensen, Q. Shen, Fuzzy-rough attribute reduction with application to web categorization, Fuzzy Sets and Systems 141 (3) (2004) 469–485.
[23] R. Jensen, Q. Shen, Fuzzy-rough sets assisted attribute selection, IEEE Transactions on Fuzzy Systems 15 (1) (2007) 73–89.
[24] R. Jensen, Q. Shen, New approaches to fuzzy-rough feature selection, IEEE Transactions on Fuzzy Systems 17 (4) (2009) 824–838.
[25] I. Kononenko, E. Simec, M. Robnik-Sikonja, Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF, Applied Intelligence 7 (1) (1997) 39–55.
[26] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1–2) (1997) 273–324.
[27] P. Lingras, R. Jensen, Survey of rough and fuzzy hybridization, in: Proc. 16th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'07), 2007, pp. 1–6.
[28] D. Maier, The Theory of Relational Databases, Computer Science Press, 1983.
[29] O.Z. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, Springer Science & Business, 2005.
[30] J.S. Mi, Y. Leung, H.Y. Zhao, T. Feng, Generalized fuzzy rough sets determined by a triangular norm, Information Sciences 178 (16) (2008) 3203–3213.
[31] H.S. Nguyen, Approximate boolean reasoning: foundations and applications in data mining, Transactions on Rough Sets V, Lecture Notes in Computer Science 4100, Springer, 2006, pp. 334–506.
[32] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11 (5) (1982) 341–356.
[33] Z. Pawlak, Rough Sets – Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, Netherlands, 1991.
[34] Z. Pawlak, A. Skowron, Rudiments of rough sets, Information Sciences 177 (2007) 3–27.
[35] Z. Pawlak, A. Skowron, Rough sets: some extensions, Information Sciences 177 (2007) 28–40.
[36] Z. Pawlak, A. Skowron, Rough sets and boolean reasoning, Information Sciences 177 (2007) 41–73.
[37] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufman, 1988.
[38] W. Pedrycz, G. Vukovich, Feature analysis through information granulation and fuzzy sets, Pattern Recognition 35 (4) (2002) 825–834.
[39] L. Polkowski, Rough mereology as a link between rough and fuzzy set theories, Transactions on Rough Sets II, Lecture Notes in Computer Science 3135, Springer, 2004, pp. 253–277.
[40] A.M. Radzikowska, E.E. Kerre, A comparative study of fuzzy rough sets, Fuzzy Sets and Systems 126 (2002) 137–156.
[41] J. Rissanen, Minimum-Description-Length Principle, Encyclopedia of Statistical Sciences, Wiley, 1985. pp. 523–527.
[42] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, in: R. Słowiński (Ed.), Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory, Kluwer Academic Publishers, Dordrecht, Netherlands, 1992, pp. 331–362.
[43] D. Ślęzak, Various approaches to reasoning with frequency based decision reducts, Rough Set Methods and Applications, New Developments in Knowledge Discovery in Information Systems, Studies in Fuzziness and Soft Computing, vol. 56, Physica-Verlag, 2000, pp. 235–288.
[44] D. Ślęzak, Degrees of conditional (in)dependence: a framework for approximate Bayesian networks and examples related to the rough set-based feature selection, Information Sciences 179 (2009) 197–209.
[45] J. Stefanowski, A. Tsoukiás, Incomplete information tables and rough classification, Computational Intelligence Journal 17 (3) (2001) 545–566.
[46] J. Stepaniuk, Tolerance Information Granules, Monitoring, Security, and Rescue Techniques in Multiagent Systems. Advances in Soft Computing, Springer, 2005, pp. 305–316.
[47] B. Sun, Z. Gong, D. Chen, Fuzzy rough set theory for the interval-valued fuzzy information systems, Information Sciences 178 (13) (2008) 2794–2815.
[48] J. Teghem, M. Benjelloun, Some experiments to compare rough sets theory and ordinal statistical methods, in: R. Slowiński (Ed.), Intelligent Decision Support, Handbook of Applications and Advances of the Rough Sets Theory, Kluwer Academic Publishers, 1992, pp. 267–284.
[49] E.C.C. Tsang, D.S. Yeung, X.Y. Wang, OFFSS: optimal fuzzy-valued feature subset selection, IEEE Transactions on Fuzzy Systems 11 (2) (2003) 202–213.
[50] E.C.C. Tsang, D.G. Chen, D.S. Yeung, X.Z. Wang, J.W. T Lee, Attributes reduction using fuzzy rough sets, IEEE Transactions on Fuzzy Systems 16 (5) (2008) 1130–1141.
[51] X. Wang, X.E. Tsang, S. Zhao, D. Chen, D. Yeung, Learning fuzzy rules from fuzzy samples based on rough set technique, Information Sciences 177 (20) (2007) 4493–4514.
[52] S. Widz, D. Ślęzak, Approximation degrees in decision reduct-based MRI segmentation, in: Proc. Int. Conf. on Frontiers in the Convergence of Bioscience and Information Technologies (FBIT'07), 2007, pp. 431–436.
[53] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, second ed., Morgan Kaufmann, San Francisco, 2005.
[54] J. Wróblewski, Theoretical foundations of order-based genetic algorithms, Fundamenta Informaticae 28 (1996) 423–430.
[55] M. Yang, S. Chen, X. Yang, A novel approach of rough set-based attribute reduction using fuzzy discernibility matrix, in: Proc. 4th Int. Conf. on Fuzzy Systems and Knowledge Discovery, 2007, pp. 96–101.
[56] X. Yang, J. Yang, C. Wu, D. Yu, Dominance-based rough set approach and knowledge reductions in incomplete ordered information system, Information Sciences 178 (4) (2008) 1219–1234.
[57] Y. Yao, Combination of rough and fuzzy sets based on α-level sets, in: T.Y. Lin, N. Cercone (Eds.), Rough Sets and Data Mining: Analysis for Imprecise Data, Kluwer Academic Publishers, 1997, pp. 301–321.
[58] Y. Yao, Y. Zhao, Attribute reduction in decision-theoretic rough set models, Information Sciences 178 (17) (2008) 3356–3373.
[59] L.A. Zadeh, Fuzzy sets, Information and Control 8 (1965) 338–353.
[60] Y. Zhao, Y. Yao, F. Luo, Data analysis based on discernibility and indiscernibility, Information Sciences 177 (22) (2007) 4959–4976.
[61] S. Zhao, E.C.C. Tsang, On fuzzy approximation operators in attribute reduction with fuzzy rough sets, Information Sciences 178 (16) (2007) 3163–3176.

Please cite this article in press as: C. Cornelis et al., Attribute selection with fuzzy decision reducts, Inform. Sci. (2009), doi:10.1016/j.ins.2009.09.008