Attribute Selection with Fuzzy Decision Reducts

Chris Cornelis\textsuperscript{a}, Richard Jensen\textsuperscript{b}, Germán Hurtado\textsuperscript{a,c}, Dominik Ślezak\textsuperscript{d,e}

\textsuperscript{a}Department of Applied Mathematics and Computer Science, Ghent University, Gent, Belgium (email: Chris.Cornelis@UGent.be, German.HurtadoMartin@UGent.be)
\textsuperscript{b}Department of Computer Science, Aberystwyth University, Wales, UK (email: rkj@aber.ac.uk)
\textsuperscript{c}Department of Applied Engineering Sciences, University College Ghent, Gent, Belgium
\textsuperscript{d}Institute of Mathematics, University of Warsaw, Warsaw, Poland
\textsuperscript{e}Infobright Inc., Warsaw, Poland (email: slezak@infobright.com)

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.

Key words: rough sets, fuzzy sets, attribute selection, data analysis, decision reducts.

1. Introduction

Rough set theory, initiated by Pawlak \cite{32,34} in the early eighties, 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, 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 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 \(\alpha\) (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 reducthood (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]).

\(^1\)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].
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 reducthood \( \alpha \), 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 \((X, A)\), where \(X = \{x_1, \ldots, x_n\}\) and \(A = \{a_1, \ldots, a_m\}\) are finite, non-empty sets of objects and attributes, respectively. Each \(a \in A\) corresponds to an \(X \rightarrow V_a\) mapping, in which \(V_a\) is the value set of \(a\) over \(X\). For every subset \(B\) of \(A\), the \(B\)-indiscernibility relation\(^2\) \(R_B\) is defined as

\[
R_B = \{(x, y) \in X^2 \text{ and } (\forall a \in B)(a(x) = a(y))\}
\]

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

\[
R_B \downarrow A = \{x \in X | [x]_{R_B} \subseteq A\}
\]
\[
R_B \uparrow A = \{x \in X | [x]_{R_B} \cap A \neq \emptyset\}
\]

\(^2\)When \(B = \{a\}\), i.e., \(B\) is a singleton, we will write \(R_a\) instead of \(R_{\{a\}}\).
A decision system \((X, A \cup \{d\})\) is a special kind of information system, used in the context of classification, in which \(d\) (\(d \notin A\)) is a designated attribute called the decision attribute. Its equivalence classes \([x]_{R_d}\) are called decision classes. Given \(B \subseteq A\), the \(B\)-positive region \(POS_B\) contains those objects from \(X\) for which the values of \(B\) allow to predict the decision class unequivocally:

\[
POS_B = \bigcup_{x \in X} R_B \downarrow [x]_{R_d}
\]

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 \(B\)):

\[
\gamma_B = \frac{|POS_B|}{|X|}
\]

\((X, A \cup \{d\})\) is called consistent if \(\gamma_A = 1\). A subset \(B\) of \(A\) is called a decision reduct if it satisfies \(POS_B = POS_A\), i.e., \(B\) preserves the decision making power of \(A\), and moreover it cannot be further reduced, i.e., there exists no proper subset \(B'\) of \(B\) such that \(POS_{B'} = POS_A\). 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 7 objects and 8 conditional attributes, all quantitative:

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

\(^3\)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](http://www.ics.uci.edu/~mlearn/MLRepository.html)
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

\[
\begin{array}{cccccccc}
  & a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 & d \\
 x_1 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\
x_2 & 1 & 2 & 2 & 1 & 1 & 1 & 1 & 1 \\
x_3 & 1 & 1 & 1 & 1 & 2 & 2 & 1 & 0 \\
x_4 & 1 & 2 & 1 & 1 & 1 & 2 & 0 & 1 \\
x_5 & 0 & 0 & 2 & 1 & 0 & 3 & 2 & 1 \\
x_6 & 0 & 0 & 1 & 1 & 0 & 3 & 0 & 0 \\
x_7 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
\end{array}
\]

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,

\[
POS_A = \{x_1, x_2, x_3, x_4, x_5, x_7, x_8\},
\]

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 \((X, \mathcal{A} \cup \{d\})\) is the \(n \times n\) matrix \(O\), defined by, for \(i\) and \(j\) in \(\{1, \ldots, n\}\),

\[
O_{ij} = \begin{cases} 
\emptyset & \text{if } d(x_i) = d(x_j) \\
\{a \in \mathcal{A} | a(x_i) \neq a(x_j)\} & \text{otherwise}
\end{cases}
\]
On the other hand, the \textit{discernibility function} of \((X, \mathcal{A} \cup \{d\})\) is the \(\{0, 1\}^m \rightarrow \{0, 1\}\) mapping \(f\), defined by
\[
f(a_1^*, ..., a_m^*) = \bigwedge \left\{ \bigvee O_{ij}^* \mid 1 \leq i < j \leq n \text{ and } O_{ij} \neq \emptyset \right\}
\] (7)
in which \(O_{ij}^* = \{a^* \mid a \in O_{ij}\}\). The boolean variables \(a_1^*, ..., a_m^*\) correspond to the attributes from \(\mathcal{A}\), and we denote \(\mathcal{A}^* = \{a_1^*, ..., a_m^*\}\). If \(B \subseteq \mathcal{A}\), then the valuation function \(V_B\) corresponding to \(B\) is defined by \(V_B(a^*) = 1\) iff \(a \in B\).

This valuation can be extended to arbitrary boolean formulas, such that
\[
V_B(f(a_1^*, ..., a_m^*)) = f(V_B(a_1^*), ..., V_B(a_m^*))
\] (8)
Formula (8) expresses whether the attributes in \(B\) preserve the discernibility of \((X, \mathcal{A} \cup \{d\})\) (when its value is 1) or not (when it is 0). The discernibility function can be reduced to its disjunctive normal form, that is
\[
f(a_1^*, ..., a_m^*) = \bigwedge A_1^* \lor ... \lor \bigwedge A_p^*
\] (9)
in which \(p \geq 1\), and for all \(i\) in \(\{1, ..., p\}\) it holds that \(A_i^* \subseteq \mathcal{A}^*\), and \(A_i^* \not\subseteq A_j^*\) for \(i \neq j\). If we define \(a \in A_i^*\) iff \(a^* \in A_i^*\), then it can be shown [42] that \(A_1, ..., A_p\) constitute exactly all decision reducts of \((X, \mathcal{A} \cup \{d\})\).

\textbf{Example 2.} For the discretized decision system in Ex. 1, it can be verified that the discernibility function (after reduction) is given by
\[
f(a_1^*, ..., a_8^*) = a_2^* \lor (a_1^* \land a_7^*) \lor (a_5^* \land a_7^*) \lor (a_6^* \land a_7^*) \lor (a_7^* \land a_8^*)
\] (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 $\gamma_R = \gamma_A$. 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 $X \rightarrow [0,1]$ 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 $R_y$ defined by

$$R_y(x) = R(x, y) \quad (11)$$

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

$$R(x, x) = 1 \quad (12)$$

$$R(x, y) = R(y, x) \quad (13)$$

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

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

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

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) $T$ is any increasing, commutative and associative $[0,1]^2 \rightarrow [0,1]$ mapping satisfying $T(1, x) = x$, for all $x$ in $[0,1]$. In this paper, we use $T_M$ and $T_L$ defined by $T_M(x, y) = \min(x, y)$ and $T_L(x, y) = \max(0, x+y-1)$ (Lukasiewicz t-norm), for $x, y$ in $[0,1]$. On the other hand, an implicator is any $[0,1]^2 \rightarrow [0,1]$-mapping $I$ satisfying $I(0, 0) = 1, I(1, x) = x$, for all $x$ in $[0,1]$. Moreover we require $I$ to be decreasing in its first, and increasing in its second component. The implicators used in this paper are $I_M$ and $I_L$ defined by $I_M(x, y) = \max(1-x, y)$ (Kleene-Dienes implicator) and $I_L(x, y) = \min(1, 1-x+y)$ (Łukasiewicz implicator) for $x, y$ in $[0,1]$. 

8
2.3. Fuzzy Rough Set Theory

Research on the hybridization of fuzzy sets and rough sets emerged in the late 1980’s [12] 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 $(X, A \cup \{d\})$, let $a$ be a quantitative attribute in $A \cup \{d\}$. 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 $X$ ($\sigma_a$ 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

\(^4\)It should be mentioned that many authors impose an additional requirement of $T$-transitivity, i.e., given a $t$-norm $T$, $T(R(x,y), R(y,z)) \leq R(x,z)$ should hold for any $x$, $y$ and $z$ in $X$; $R$ is then called a fuzzy $T$-equivalence relation, or similarity relation. While $T$-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].
\( R_a(x,y) = 0 \) otherwise, we can define, for any subset \( B \) of \( A \), the fuzzy \( B \)-indiscernibility relation by

\[
R_B(x,y) = T \left( \bigwedge_{a \in B} R_a(x,y) \right)
\]

(16)

in which \( T \) 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 equation (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 \( I \) and a t-norm \( T \), they paraphrased formulas (2) and (3) to define \( R \downarrow A \) and \( R \uparrow A \) by

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

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

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 \( A \) w.r.t. their ability to maintain discernibility relative to the decision attribute and to generate adequate decision rules. Once such a measure, say \( M \), is obtained, we can associate a notion of fuzzy decision reduct with it.

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

\(^5\)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.
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 $B \subseteq A$, 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.

### 3.1. Fuzzy Positive Region

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

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

(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

$$POS_B(y) = \max_{x \in X} \inf_{z \in X} I(R_B(z, y), R_d(z, x))$$

$$= \max_{x \in R_d y} \left( \max_{z \in X} \inf_{z \in X} I(R_B(z, y), R_d(z, x)) \right)$$

$$= \max_{x \in R_d y} \max_{z \in X} I(R_B(z, y), R_d(z, x))$$

$$= \inf_{z \in X} I(R_B(z, y), R_d(z, y))$$

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

\square

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

For instance, since $\sigma_{a_4} = 16.385$ and $\sigma_{a_5} = 131.176$,

$$ POS_B(x_3) = \min(1, 1 - R_B(x, x_3), 1, 1 - R_B(x_4, x_3), 1, 1, 1) $$

$$ = 1 - \max(R_B(x_2, x_3), R_B(x_4, x_3)) $$

$$ = 1 - \max(0, R_{a_4}(x_2, x_3) + R_{a_5}(x_2, x_3) - 1, R_{a_4}(x_4, x_3) + R_{a_5}(x_4, x_3) - 1) $$

$$ = 1 - \max(0, 0.512 + 0.680 - 1, 0.817 + 0.710 - 1) = 0.473 $$

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, $POS_A = X$ 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 $(R_B \upharpoonright R_d y)(y)$, 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 7 objects, 13 conditional attributes and a quantitative decision attribute:

\(^6\)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
We use the same parameters as in Ex. 4. For $B = \{a_3\}$, we get e.g.,

\[
(R_B \downarrow R_d x_2)(x_2) = \inf_{z \in X} \mathcal{I}(R_B(z, x_2), R_d(z, x_2))
\]

\[
= \min(1, 1, 0.512, 0.707, 1, 1, 1) = 0.512
\]

but on the other hand,

\[
(R_B \downarrow R_d x_4)(x_2) = \inf_{z \in X} \mathcal{I}(R_B(z, x_2), R_d(z, x_4))
\]

\[
= \min(1, 0.707, 0.805, 1, 1, 1, 1) = 0.707
\]

As a consequence,

\[
POS_B(x_2) = \max_{x \in X} \min_{z \in X} \mathcal{I}(R_B(z, x_2), R_d(z, x))
\]

\[
= \max(0.228, 0.512, 0.512, 0.707, 0, 0, 0.455) = 0.707
\]

\[
> (R_B \downarrow R_d x_2)(x_2)
\]

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

\[
POS'_B(y) = (R_B \downarrow R_d y)(y)
\]

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.

\[ \gamma_B = \frac{|POS_B|}{|POS_A|} \text{ and } \gamma'_B = \frac{|POS'_B|}{|POS'_A|} \]

These measures 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:

\[ \delta_B = \frac{\min_{x \in X} POS_B(x)}{\min_{x \in X} POS_A(x)} \text{ and } \delta'_B = \frac{\min_{x \in X} POS'_B(x)}{\min_{x \in X} POS'_A(x)} \]

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.

\(^7\)Normalization is required in order that the measure yields a value of 1 for the whole attribute set. In this paper, we assume \(POS_A(x) > 0\) for every \(x\) in \(X\).
Proposition 2. For subsets $B_1, B_2$ of $A$,

$$B_1 \subseteq B_2 \Rightarrow \left\{ \begin{array}{l}
\gamma_{B_1} \leq \gamma_{B_2} \text{ and } \gamma'_{B_1} \leq \gamma'_{B_2} \\
\delta_{B_1} \leq \delta_{B_2} \text{ and } \delta'_{B_1} \leq \delta'_{B_2}
\end{array} \right. $$

Proposition 3. $\gamma_A = \gamma'_A = \delta_A = \delta'_A = 1$

Example 7. For the Pima decision system from Ex. 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 $\text{POS}_A = X$ (and hence also $\text{POS}'_A = X$). 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.

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

$$f(a_1^*, \ldots, a_m^*)
= \bigwedge \left\{ \bigvee_{k=1}^{m} a_k^*[d(x_i) \neq d(x_j) \Rightarrow a_k(x_i) \neq a_k(x_j)] | 1 \leq i < j \leq n \right\}
= \bigwedge \left\{ \bigvee_{k=1}^{m} a_k^* [a_k(x_i) = a_k(x_j) \Rightarrow d(x_i) = d(x_j)] | 1 \leq i < j \leq n \right\}
= \bigwedge \left\{ \left[ \bigwedge_{a^*_k=1}^{m} (a_k(x_i) = a_k(x_j)) \Rightarrow d(x_i) = d(x_j) \right] | 1 \leq i < j \leq n \right\} \quad (23)$$

provided the decision system is consistent.$^8$

$^8$Recall that if $(X, A \cup \{d\})$ is inconsistent, there exist $x_i$ and $x_j$ such that $(\forall a \in A)(a(x_i) = a(x_j))$, yet $d(x_i) \neq d(x_j)$. Such $x_i$ and $x_j$ are not considered in Eq. (7), since $O_{ij} = 0$. 

15
Interpreting the connectives in Eq. (23) by a t-norm $T$ and an implicator $I$, and replacing the exact equalities by the respective approximate equalities (fuzzy indiscernibility relations), we can extend the discernibility function to a $\{0,1\}^m \rightarrow [0,1]$ mapping in the following way:

$$f(a_1^*,\ldots,a_m^*) = T(\bigwedge_{1 \leq i < j \leq n} c_{ij}(a_1^*,\ldots,a_m^*))$$

with

$$c_{ij}(a_1^*,\ldots,a_m^*) = I(T(R_{a_k}(x_i,x_j)),R_d(x_i,x_j))$$

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 $\mathcal{V}_B$ corresponding to a subset $B$ of $\mathcal{A}$, $\mathcal{V}_B(f(a_1^*,\ldots,a_m^*))$ 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:

$$f_B = \frac{\mathcal{V}_B(f(a_1^*,\ldots,a_m^*))}{\mathcal{V}_A(f(a_1^*,\ldots,a_m^*))}$$

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

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

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

$$g_B = \frac{\mathcal{V}_B(g(a_1^*,\ldots,a_m^*))}{\mathcal{V}_A(g(a_1^*,\ldots,a_m^*))}$$

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 $A$,

$$B_1 \subseteq B_2 \Rightarrow \left\{ \begin{array}{l}
  f_{B_1} \leq f_{B_2} \\
  g_{B_1} \leq g_{B_2}
\end{array} \right.$$  

Proposition 5. $f_A = g_A = 1$

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

$$
\begin{align*}
  f_B & = \frac{\nu_B(f(a_1^*, ..., a_n^*))}{\nu_A(f(a_1^*, ..., a_n^*))} = \frac{T \sum_{1 \leq i < j \leq 7} c_{ij}(0, 0, 0, 1, 0, 0, 0)}{\sum_{1 \leq i < j \leq 7} c_{ij}(1, 1, 1, 1, 1, 1, 1)} = 0.281 \\
  g_B & = \frac{\nu_B(g(a_1^*, ..., a_n^*))}{\nu_A(g(a_1^*, ..., a_n^*))} = \frac{\sum_{1 \leq i < j \leq 7} c_{ij}(0, 0, 0, 1, 0, 0, 0)}{21} = \frac{20.281}{21} = 0.966
\end{align*}
$$

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

$$
\begin{align*}
  f_B & = \frac{T \sum_{1 \leq i < j \leq 7} c_{ij}(0, 0, 1, 0, 0, 0, 0)}{\sum_{1 \leq i < j \leq 7} c_{ij}(0, 0, 1, 0, 0, 0, 0)} = \frac{0.572, 0.641, 1, 0.512, 0.707, 1, 1, 0.805, 1, 1, 1, 1, 0.094, 0.184, 0.278}{21} = 0.800
\end{align*}
$$

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., $\delta'_B \leq \gamma'_B \leq \gamma_B$ and $\delta'_B \leq \delta_B \leq \gamma_B$ 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 $T$ and implicator $I$ are used in Eqs. (16), (17), (24) and (25).
Proposition 6. If $\text{POS}'_A = X$,

$$f_B \leq \delta'_B \text{ and } \gamma'_B \leq g_B \quad (29)$$

for $B \subseteq A$. Moreover, in case $T = T_M$, $f_B = \delta'_B$, regardless of $\text{POS}'_A = X$.

Proof.

$$f_B = \frac{\mathcal{V}_B(f(a_1^*, \ldots, a_m^*))}{\mathcal{V}_A(f(a_1^*, \ldots, a_m^*))} \leq \mathcal{V}_B(f(a_1^*, \ldots, a_m^*))$$

$$= T(I(R_B(x_i, x_j), R_d(x_i, x_j)))$$

$$\leq \min_{1 \leq i < j \leq n} I(R_B(x_i, x_j), R_d(x_i, x_j)) = \min_{x, y \in X} I(R_B(x, y), R_d(x, y))$$

$$= \min_{y \in X} (R_B \downarrow R_d y)(y) \leq \min_{y \in X} \text{POS}'_B(y) = \delta'_B$$

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

$$\gamma'_B = \sum_{y \in X} (R_B \downarrow R_d y)(y)$$

$$= \sum_{y \in X} \inf_{x \in X} I(R_B(x, y), R_d(x, y))$$

$$= \sum_{1 \leq j \leq n} \inf_{x \in X} I(R_B(x, x_j), R_d(x, x_j))$$

$$\leq 2 \sum_{1 \leq i < j \leq n} I(R_B(x_i, x_j), R_d(x_i, x_j))$$

$$\leq \frac{2 \sum_{1 \leq i < j \leq n} I(R_B(x_i, x_j), R_d(x_i, x_j))}{n(n - 1)} = g_B$$

The above proof shows that $f$ and $\delta$ are essentially built upon the same idea, with some variations due to the parameter choice, and also reveals the essential difference between $\gamma$ and $g$: while the former looks at the lowest value of the formula $I(R_B(x, y), R_d(x, y))$ 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 \( \gamma/\gamma' \)-decision reduct.

**Proposition 7.** If \( POS_A' = X \),

\[ g_B = 1 \Rightarrow (\gamma_B' = 1 \text{ and } \gamma_B = 1) \]  \hspace{1cm} (30)

for any \( B \subseteq A \).

**Proof.**

\[
g_B = 1 \Rightarrow \frac{2 \sum_{1 \leq i < j \leq n} I(R_B(x_i, x_j), R_d(x_i, x_j))}{n(n-1)} = 1
\]

\[
\Rightarrow (\forall 1 \leq i < j \leq n)(I(R_B(x_i, x_j), R_d(x_i, x_j)) = 1)
\]

\[
\Rightarrow (\forall 1 \leq j \leq n)(\inf_{x \in X} I(R_B(x, x_j), R_d(x, x_j)) = 1)
\]

\[
\Rightarrow \gamma_B' = 1
\]

Since \( \gamma_B' \leq \gamma_B \), \( \gamma_B = 1 \) holds as well. \( \square \)

**Example 9.** Some of the relationships discussed above are illustrated in the following table, which contains the values obtained with \( T_L \) and \( I_L \) in the previous sections (\( B = \{a_4, a_5\} \) for pima, \( B = \{a_3\} \) for housing), along with those obtained with \( T_M \) and \( I_M \):

<table>
<thead>
<tr>
<th>Data set</th>
<th>Connectives</th>
<th>( \gamma_B )</th>
<th>( \gamma_B' )</th>
<th>( \delta_B )</th>
<th>( \delta_B' )</th>
<th>( f_B )</th>
<th>( g_B )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pima</td>
<td>( T_M, I_M )</td>
<td>0.72</td>
<td>0.72</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>( T_L, I_L )</td>
<td>0.82</td>
<td>0.82</td>
<td>0.47</td>
<td>0.47</td>
<td>0.28</td>
<td>0.97</td>
</tr>
<tr>
<td>Housing</td>
<td>( T_M, I_M )</td>
<td>0.43</td>
<td>0.38</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>( T_L, I_L )</td>
<td>0.44</td>
<td>0.38</td>
<td>0.09</td>
<td>0.09</td>
<td>0.00</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Note that there is no general pattern in the relationship between the results obtained with \( T_M \) and \( I_M \) and those with \( T_L \) and \( I_L \); 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 \((X, \mathcal{A} \cup \{d\})\), a measure \(\mathcal{M}\) as in Definition 1, and a threshold \(\alpha\) \((0 < \alpha \leq 1)\), we run a 10-fold cross validation experiment. In each iteration, we apply an adapted version of the \textsc{QuickReduct} heuristic, shown in Figure 1, to the training data to obtain a fuzzy \(\mathcal{M}\)-decision superreduct to degree \(\alpha\). All measures use Eq. (15) for evaluating attribute level discernibility, and, unless explicitly stated otherwise, the Lukasiewicz connectives \(T_L\) and \(I_L\) 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,

\(^9\)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 \textsc{QuickReduct} heuristic, have been implemented in Weka [53]. The program can be downloaded from http://users.aber.ac.uk/rkj/book/programs.php
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 Conclusion).

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 at-
Figure 1: Modified QuickReduct to obtain a fuzzy $M$-decision superreduct to degree $\alpha$.

tribute 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 Lukasiewicz connectives, except $f$, which was found to perform better in combination with $T_M$ and $I_M$. Note that, by Proposition 6, this implementation of $f$ coincides with the corresponding $\delta$ measure based on $T_M$ and $I_M$. 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.

The results are shown in Table 2 and Table 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 $\gamma$ and $\gamma'$ 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 $\delta$ and $f$ are generally longer, without necessarily being better than their $\gamma$ 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, 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 $\delta$ 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 ($\gamma$ 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 $\gamma'$ and $\delta'$ are in general competitive with those of the more complex measures $\gamma$ and $\delta$, which justifies their simplification.
<table>
<thead>
<tr>
<th>Data set</th>
<th>Unred.</th>
<th>( \alpha = 1 )</th>
<th>( \alpha = .95 )</th>
<th>( \delta )</th>
<th>( \alpha = 1 )</th>
<th>( \alpha = .95 )</th>
<th>( f )</th>
<th>( \alpha = 1 )</th>
<th>( \alpha = .95 )</th>
<th>( g )</th>
<th>( \alpha = 1 )</th>
<th>( \alpha = .9999 )</th>
<th>CFS</th>
<th>WSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEVELAND</td>
<td>51.56</td>
<td>49.23</td>
<td>45.85</td>
<td>53.62</td>
<td>53.95</td>
<td>49.11</td>
<td>49.45</td>
<td>49.23</td>
<td>50.54</td>
<td>53.20</td>
<td>49.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLASS</td>
<td>70.52</td>
<td>70.52</td>
<td>70.00</td>
<td>70.06</td>
<td>70.04</td>
<td>73.33</td>
<td>73.33</td>
<td>69.59</td>
<td>70.06</td>
<td>78.92</td>
<td>69.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IONOSPHERE</td>
<td>86.09</td>
<td>89.13</td>
<td>86.52</td>
<td>84.78</td>
<td>85.65</td>
<td>84.35</td>
<td>84.35</td>
<td>88.26</td>
<td>88.26</td>
<td>87.39</td>
<td>83.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIMA</td>
<td>73.45</td>
<td>71.15</td>
<td>75.24</td>
<td>72.93</td>
<td>73.70</td>
<td>72.44</td>
<td>72.19</td>
<td>71.65</td>
<td>72.20</td>
<td>69.41</td>
<td>68.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WATER 2</td>
<td>84.10</td>
<td>81.54</td>
<td>82.56</td>
<td>80.00</td>
<td>80.00</td>
<td>83.33</td>
<td>83.08</td>
<td>82.56</td>
<td>79.74</td>
<td>86.15</td>
<td>77.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WATER 3</td>
<td>81.54</td>
<td>78.97</td>
<td>77.18</td>
<td>78.21</td>
<td>78.21</td>
<td>79.74</td>
<td>79.49</td>
<td>80.77</td>
<td>79.74</td>
<td>83.85</td>
<td>75.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WEB</td>
<td>40.95</td>
<td>42.95</td>
<td>42.86</td>
<td>37.52</td>
<td>37.52</td>
<td>35.48</td>
<td>34.81</td>
<td>34.90</td>
<td>32.24</td>
<td>53.71</td>
<td>47.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WINE</td>
<td>94.97</td>
<td>92.71</td>
<td>91.57</td>
<td>91.57</td>
<td>97.22</td>
<td>96.63</td>
<td>91.05</td>
<td>93.86</td>
<td>96.08</td>
<td>94.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLEVELAND</td>
<td>13</td>
<td>7.10</td>
<td>5.00</td>
<td>8.80</td>
<td>8.70</td>
<td>11.50</td>
<td>11.40</td>
<td>7.70</td>
<td>6.00</td>
<td>6.90</td>
<td>4.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLASS</td>
<td>9</td>
<td>9.00</td>
<td>7.00</td>
<td>8.00</td>
<td>7.70</td>
<td>6.20</td>
<td>5.80</td>
<td>8.20</td>
<td>8.00</td>
<td>6.30</td>
<td>5.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IONOSPHERE</td>
<td>34</td>
<td>7.10</td>
<td>5.10</td>
<td>9.20</td>
<td>9.00</td>
<td>12.10</td>
<td>11.60</td>
<td>6.90</td>
<td>5.90</td>
<td>10.80</td>
<td>4.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIMA</td>
<td>8</td>
<td>7.50</td>
<td>5.00</td>
<td>7.70</td>
<td>7.00</td>
<td>6.20</td>
<td>6.10</td>
<td>7.60</td>
<td>5.00</td>
<td>4.30</td>
<td>2.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WATER 2</td>
<td>38</td>
<td>6.00</td>
<td>4.00</td>
<td>6.00</td>
<td>6.00</td>
<td>20.20</td>
<td>18.20</td>
<td>6.00</td>
<td>4.90</td>
<td>9.10</td>
<td>3.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WATER 3</td>
<td>38</td>
<td>6.00</td>
<td>4.20</td>
<td>6.20</td>
<td>6.20</td>
<td>22.00</td>
<td>20.90</td>
<td>5.90</td>
<td>5.00</td>
<td>11.10</td>
<td>3.70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WEB</td>
<td>2556</td>
<td>18.40</td>
<td>14.90</td>
<td>30.10</td>
<td>30.10</td>
<td>42.90</td>
<td>41.40</td>
<td>16.00</td>
<td>15.00</td>
<td>57.80</td>
<td>8.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WINE</td>
<td>13</td>
<td>5.00</td>
<td>3.90</td>
<td>4.90</td>
<td>4.90</td>
<td>10.50</td>
<td>9.50</td>
<td>4.80</td>
<td>4.00</td>
<td>10.90</td>
<td>4.60</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Regression results

<table>
<thead>
<tr>
<th>Data set</th>
<th>Orig.</th>
<th>$\gamma$</th>
<th>$\gamma'$</th>
<th>$\delta$</th>
<th>$\delta'$</th>
<th>$f$</th>
<th>$g$</th>
<th>CFS</th>
<th>WSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha = 1$</td>
<td>$\alpha = 0.95$</td>
<td>$\alpha = 1$</td>
<td>$\alpha = 0.95$</td>
<td>$\alpha = 1$</td>
<td>$\alpha = 0.95$</td>
<td>$\alpha = 1$</td>
<td>$\alpha = 0.95$</td>
<td>$\alpha = 1$</td>
</tr>
<tr>
<td>Algae B</td>
<td>17.4</td>
<td>15.46</td>
<td>12.95</td>
<td>15.84</td>
<td>13.63</td>
<td>14.57</td>
<td>14.3</td>
<td>16.11</td>
<td>14.05</td>
</tr>
<tr>
<td>Algae D</td>
<td>4.72</td>
<td>3.52</td>
<td>3.83</td>
<td>3.52</td>
<td>4.43</td>
<td>3.6</td>
<td>3.57</td>
<td>3.68</td>
<td>3.68</td>
</tr>
<tr>
<td>Algae G</td>
<td>6.42</td>
<td>5.41</td>
<td>5.84</td>
<td>5.43</td>
<td>5.09</td>
<td>5.79</td>
<td>5.85</td>
<td>5.69</td>
<td>5.85</td>
</tr>
<tr>
<td>Housing</td>
<td>4.6</td>
<td>3.93</td>
<td>4.64</td>
<td>3.93</td>
<td>4.64</td>
<td>4.03</td>
<td>4.1</td>
<td>4.76</td>
<td>4.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data set</th>
<th>Orig.</th>
<th>$\gamma$</th>
<th>$\gamma'$</th>
<th>$\delta$</th>
<th>$\delta'$</th>
<th>$f$</th>
<th>$g$</th>
<th>CFS</th>
<th>WSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha = 1$</td>
<td>$\alpha = 0.95$</td>
<td>$\alpha = 1$</td>
<td>$\alpha = 0.95$</td>
<td>$\alpha = 1$</td>
<td>$\alpha = 0.95$</td>
<td>$\alpha = 1$</td>
<td>$\alpha = 0.95$</td>
<td>$\alpha = 1$</td>
</tr>
<tr>
<td>Algae A</td>
<td>11</td>
<td>10.7</td>
<td>5.9</td>
<td>10.7</td>
<td>5.8</td>
<td>9.8</td>
<td>8.7</td>
<td>9.8</td>
<td>8.7</td>
</tr>
<tr>
<td>Algae B</td>
<td>11</td>
<td>9.7</td>
<td>5.3</td>
<td>10.1</td>
<td>5.3</td>
<td>9.2</td>
<td>7.6</td>
<td>10.2</td>
<td>9.2</td>
</tr>
<tr>
<td>Algae C</td>
<td>11</td>
<td>10.8</td>
<td>5.4</td>
<td>10.8</td>
<td>6.2</td>
<td>9.1</td>
<td>7.3</td>
<td>10</td>
<td>8.6</td>
</tr>
<tr>
<td>Algae D</td>
<td>11</td>
<td>9.8</td>
<td>4.3</td>
<td>9.8</td>
<td>5.3</td>
<td>9.2</td>
<td>7.8</td>
<td>9.4</td>
<td>8.1</td>
</tr>
<tr>
<td>Algae E</td>
<td>11</td>
<td>9.4</td>
<td>5.1</td>
<td>9.5</td>
<td>5.1</td>
<td>9.6</td>
<td>8.3</td>
<td>9.3</td>
<td>8.5</td>
</tr>
<tr>
<td>Algae F</td>
<td>11</td>
<td>7.5</td>
<td>5.2</td>
<td>7.5</td>
<td>5.2</td>
<td>7.8</td>
<td>7.5</td>
<td>8</td>
<td>7.6</td>
</tr>
<tr>
<td>Algae G</td>
<td>11</td>
<td>8.2</td>
<td>4.9</td>
<td>8.6</td>
<td>5.5</td>
<td>8.9</td>
<td>7.6</td>
<td>8.8</td>
<td>8.3</td>
</tr>
<tr>
<td>Housing</td>
<td>13</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>6.9</td>
<td>6.3</td>
<td>5.5</td>
<td>5.2</td>
</tr>
</tbody>
</table>

|          |        |        |        |        |        |      |      |      |      |       |       |       |       |
4.2. Detailed analysis on Spambase and CPU

From the above results, it is clear that the selection of an adequate $\alpha$ 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 $\gamma$ 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 $\gamma$ 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 Figure 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 $\gamma$ by several percents.

Figure 2: Spambase results: a) Classification accuracy versus subset size b) Accuracy drop versus subset size reduction
<table>
<thead>
<tr>
<th>size</th>
<th>att.</th>
<th>γ</th>
<th>acc.</th>
<th>att.</th>
<th>g</th>
<th>acc.</th>
<th>att.</th>
<th>γ</th>
<th>acc.</th>
<th>att.</th>
<th>g</th>
<th>acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26</td>
<td>0.05</td>
<td>58.42%</td>
<td>18</td>
<td>0.87</td>
<td>67.40%</td>
<td>24</td>
<td>54</td>
<td>0.98</td>
<td>87.63%</td>
<td>43</td>
<td>0.999</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>0.10</td>
<td>64.92%</td>
<td>20</td>
<td>0.95</td>
<td>74.90%</td>
<td>25</td>
<td>6</td>
<td>0.98</td>
<td>89.02%</td>
<td>54</td>
<td>0.999</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>0.16</td>
<td>70.98%</td>
<td>49</td>
<td>0.98</td>
<td>76.27%</td>
<td>26</td>
<td>0</td>
<td>0.99</td>
<td>88.83%</td>
<td>50</td>
<td>0.999</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>0.26</td>
<td>76.09%</td>
<td>56</td>
<td>0.991</td>
<td>79.27%</td>
<td>27</td>
<td>48</td>
<td>0.992</td>
<td>89.05%</td>
<td>48</td>
<td>0.999</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0.41</td>
<td>79.63%</td>
<td>51</td>
<td>0.995</td>
<td>82.42%</td>
<td>28</td>
<td>7</td>
<td>0.993</td>
<td>89.18%</td>
<td>13</td>
<td>0.999</td>
</tr>
<tr>
<td>6</td>
<td>56</td>
<td>0.55</td>
<td>83.22%</td>
<td>11</td>
<td>0.997</td>
<td>82.46%</td>
<td>29</td>
<td>23</td>
<td>0.994</td>
<td>89.48%</td>
<td>2</td>
<td>0.999</td>
</tr>
<tr>
<td>7</td>
<td>49</td>
<td>0.67</td>
<td>82.74%</td>
<td>52</td>
<td>0.999</td>
<td>84.57%</td>
<td>30</td>
<td>17</td>
<td>0.995</td>
<td>89.31%</td>
<td>36</td>
<td>0.999</td>
</tr>
<tr>
<td>8</td>
<td>51</td>
<td>0.77</td>
<td>83.37%</td>
<td>26</td>
<td>0.998</td>
<td>84.87%</td>
<td>31</td>
<td>34</td>
<td>0.995</td>
<td>89.35%</td>
<td>47</td>
<td>0.999</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>0.82</td>
<td>84.94%</td>
<td>24</td>
<td>0.9991</td>
<td>85.81%</td>
<td>32</td>
<td>47</td>
<td>0.996</td>
<td>89.39%</td>
<td>25</td>
<td>0.999</td>
</tr>
<tr>
<td>10</td>
<td>44</td>
<td>0.87</td>
<td>85.37%</td>
<td>4</td>
<td>0.9993</td>
<td>86.16%</td>
<td>33</td>
<td>25</td>
<td>0.996</td>
<td>89.51%</td>
<td>17</td>
<td>0.999</td>
</tr>
<tr>
<td>11</td>
<td>45</td>
<td>0.9</td>
<td>86.33%</td>
<td>15</td>
<td>0.9995</td>
<td>87.39%</td>
<td>34</td>
<td>22</td>
<td>0.997</td>
<td>90.13%</td>
<td>34</td>
<td>0.999</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>0.91</td>
<td>85.70%</td>
<td>6</td>
<td>0.9996</td>
<td>88.79%</td>
<td>35</td>
<td>5</td>
<td>0.997</td>
<td>89.76%</td>
<td>53</td>
<td>0.999</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0.93</td>
<td>85.11%</td>
<td>9</td>
<td>0.9997</td>
<td>88.72%</td>
<td>36</td>
<td>21</td>
<td>0.998</td>
<td>89.89%</td>
<td>5</td>
<td>0.999</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>0.94</td>
<td>85.70%</td>
<td>44</td>
<td>0.9997</td>
<td>89.05%</td>
<td>37</td>
<td>53</td>
<td>0.998</td>
<td>89.92%</td>
<td>12</td>
<td>0.999</td>
</tr>
<tr>
<td>15</td>
<td>32</td>
<td>0.95</td>
<td>85.87%</td>
<td>7</td>
<td>0.9998</td>
<td>89.55%</td>
<td>38</td>
<td>8</td>
<td>0.998</td>
<td>89.76%</td>
<td>30</td>
<td>0.999</td>
</tr>
<tr>
<td>16</td>
<td>52</td>
<td>0.95</td>
<td>86.76%</td>
<td>0</td>
<td>0.9998</td>
<td>89.39%</td>
<td>39</td>
<td>38</td>
<td>0.998</td>
<td>89.78%</td>
<td>40</td>
<td>0.999</td>
</tr>
<tr>
<td>17</td>
<td>9</td>
<td>0.96</td>
<td>87.26%</td>
<td>55</td>
<td>0.9998</td>
<td>89.35%</td>
<td>40</td>
<td>19</td>
<td>0.999</td>
<td>90.07%</td>
<td>22</td>
<td>0.999</td>
</tr>
<tr>
<td>18</td>
<td>41</td>
<td>0.97</td>
<td>87.68%</td>
<td>1</td>
<td>0.9999</td>
<td>89.59%</td>
<td>41</td>
<td>40</td>
<td>0.999</td>
<td>90.13%</td>
<td>38</td>
<td>0.999</td>
</tr>
<tr>
<td>19</td>
<td>55</td>
<td>0.97</td>
<td>87.83%</td>
<td>45</td>
<td>0.9999</td>
<td>89.46%</td>
<td>42</td>
<td>12</td>
<td>0.999</td>
<td>89.98%</td>
<td>21</td>
<td>0.999</td>
</tr>
<tr>
<td>20</td>
<td>43</td>
<td>0.97</td>
<td>87.63%</td>
<td>19</td>
<td>0.9999</td>
<td>89.76%</td>
<td>43</td>
<td>30</td>
<td>0.999</td>
<td>90.07%</td>
<td>35</td>
<td>0.999</td>
</tr>
<tr>
<td>21</td>
<td>50</td>
<td>0.98</td>
<td>87.50%</td>
<td>23</td>
<td>0.9999</td>
<td>90.07%</td>
<td>44</td>
<td>35</td>
<td>0.999</td>
<td>90.22%</td>
<td>8</td>
<td>0.999</td>
</tr>
<tr>
<td>22</td>
<td>36</td>
<td>0.98</td>
<td>87.46%</td>
<td>32</td>
<td>0.999</td>
<td>90.26%</td>
<td>45</td>
<td>27</td>
<td>1</td>
<td>90.42%</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>13</td>
<td>0.98</td>
<td>87.70%</td>
<td>41</td>
<td>0.999</td>
<td>90.52%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The most important observation about these results, however, is the obvious benefit fuzzy decision reducts have over crisp ones. For instance, a crisp $\gamma$-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 Figure 2b), if a 1% accuracy drop is permissible, fuzzy $\gamma$-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 $\gamma$ 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 Figure 3, show a largely similar picture. In this case, the advantage of fuzzy decision reducts is even more evident: $\gamma$ 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 $\gamma$ and $g$ results is smaller than in the SPAMBASE experiment, with a small advantage of $\gamma$ over $g$.

![Figure 3: CPU results: a) RMSE versus subset size b) RMSE increase versus subset size reduction](image-url)
Table 5: CPU results

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

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., $\gamma$ versus $\gamma'$), 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: \( \delta \) and \( f \) focus on its worst single violation within the data, \( \gamma \) 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 \( \delta/f \) 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), (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 \( \gamma \), then at some point shift to \( \delta \) to fine-tune the result).

Acknowledgment

Chris Cornelis would like to thank the Research Foundation—Flanders for funding his research.

References


[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.


