A Distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction

MacParthaláin, Neil Seosamh; Shen, Qiang; Jensen, Richard

Published in:
IEEE Transactions on Knowledge and Data Engineering

DOI:
10.1109/TKDE.2009.119

Publication date:
2010

Citation for published version (APA):
MacParthaláin, N. S., Shen, Q., & Jensen, R. (2010). A Distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction. *IEEE Transactions on Knowledge and Data Engineering*, *22*(3), 306-317. https://doi.org/10.1109/TKDE.2009.119

A Distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction

Neil Mac Parthaláin, Qiang Shen, and Richard Jensen

Abstract—Feature Selection (FS) or Attribute Reduction techniques are employed for dimensionality reduction and aim to select a subset of the original features of a data set which are rich in the most useful information. The benefits of employing FS techniques include improved data visualization and transparency, a reduction in training and utilization times and potentially, improved prediction performance. Many approaches based on rough set theory up to now, have employed the dependency function, which is based on lower approximations as an evaluation step in the FS process. However, by examining only that information which is considered to be certain and ignoring the boundary region, or region of uncertainty, much useful information is lost. This paper examines a rough set FS technique which uses the information gathered from both the lower approximation dependency value and a distance metric which considers the number of objects in the boundary region and the distance of those objects from the lower approximation. The use of this measure in rough set feature selection can result in smaller subset sizes than those obtained using the dependency function alone. This demonstrates that there is much valuable information to be extracted from the boundary region. Experimental results are presented for both crisp and real-valued data and compared with two other FS techniques in terms of subset size, runtimes, and classification accuracy.

Index Terms—Rough sets, fuzzy sets, attribute reduction, boundary region, classification.

1 Introduction

It is often desirable to present a large number of features for a domain such that every possible aspect of that domain is represented. However, this can often result in many redundant or irrelevant features that may lead to poor results when using data mining tools for knowledge discovery. Feature selection is a process which chooses a subset of the original features present in a given data set which provides the most useful information. Following selection, the most important information of the data set should still remain. In fact, efficient FS techniques should be able to detect and ignore noisy and misleading features. As a result, the data set quality may even *increase* through feature selection.

Classification accuracy may be increased as a result of feature selection through the removal of noisy, irrelevant, or redundant features. Also in domains where features correspond to measurements (the water treatment plant in [28] demonstrates this well), fewer features offer advantages such as minimizing the expense and time consumed in recording such measurements. For data sets which are smaller in size, the runtimes of learning algorithms can be improved significantly. This is equally applicable to both training and application (e.g., classification) phases. Where there are fewer dimensions, identification of trends and correlations within the data becomes easier [32]. This is evident where a small number of features have an influence on data outcomes.

Methods which extract knowledge from data (e.g., rule induction) may also benefit from the use of FS and show improvement in the readability of the discovered knowledge. When induction algorithms are applied to reduced data, the resulting rules are more compact. A good feature selection algorithm will remove unnecessary attributes which may affect both rule comprehension and rule prediction performance.

The work on rough set theory (RST) [24] offers a formal methodology that can be employed to reduce the dimensionality of data sets, as a preprocessing step to assist any chosen modeling method for learning from data. It assists in identifying and selecting the most information-rich features in a data set. This is achieved without transforming the data, while simultaneously attempting to minimize information loss during the selection process. In terms of computational effort, this approach is highly efficient, and is based on simple set operations, which makes it suitable as a preprocessor for techniques that are much more complex. In contrast to statistical correlation-reduction approaches [7], RST requires no human input or domain knowledge other than the given data sets. Perhaps most importantly though, it retains the underlying semantics of the data, which results in data models that are more transparent to human scrutiny.

Most existing rough-set-based FS approaches [8], [9], [13], [14], [18], [20], [33], [37] rely on the information gathered from the lower approximation of a set to minimize data. These approaches have been adopted as the certainty that is embodied in the lower approximation is associated with greater importance in scientific analysis. Although successful, these lower approximation-based approaches

• *The authors are with the Department of Computer Science, Aberystwyth University, Penglais Campus, Aberystwyth, Ceredigion, Wales SY23 3DB, UK. Email: {ncm, qqs, rkj}@aber.ac.uk.*

Manuscript received 8 Nov. 2007; revised 17 Sept. 2008; accepted 5 Mar. 2009; published online 29 Apr. 2009.
Recommended for acceptance by F.Y. Wang.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2007-11-0550.
Digital Object Identifier no. 10.1109/TKDE.2009.119.

ignore the information that is contained in the boundary region, or region of uncertainty. While there are also some existing RST approaches which consider the boundary region information [4], [11], they adopt an approach which examines the upper approximation as a whole rather than examining the lower approximation and boundary region as conceptually separate entities. This paper presents a method which based on the initial work in [16], examines both the information in the lower approximation and the information contained in the boundary region for the selection of feature subsets. This can result in the selection of subsets which are smaller than those selected using the information gathered from the lower approximation alone.

The remainder of this paper is structured as follows: Section 2 summarizes the theoretical background and ideas of RSAR—as this forms the basis for the new approach—along with a look at the rough set QUICKREDUCT algorithm. Section 3 demonstrates the main contribution of the new approach with a description of the distance-metric-assisted approach to RSAR (DMRSAR). The corresponding algorithm is also presented, as well as an examination of the computational complexity of the approach. Section 4 documents the results of applying the RSAR, fuzzy-rough feature selection FRFS [12], Principal Component Analysis (PCA) [7], tolerance rough set feature selection [29], and DMRSAR approaches to a number of both crisp and real-valued data sets. These results compare the new approach with the other previously mentioned methods in terms of classification accuracies (using three different classifiers), and dimensionality reduction. Section 5 concludes the paper along with some suggestions for further development, and a discussion of future work.

2 BACKGROUND

2.1 Rough Set Attribute Reduction

The principal focus of this paper lies in distance-metric-assisted rough set attribute reduction (DMRSAR); however, an in-depth view of the current RSAR methodology is necessary to appreciate the DMRSAR approach fully.

At the heart of the RSAR approach is the concept of indiscernibility [24]. Let $I = (\mathbb{U}, \mathbb{S})$ be an information system, where $\mathbb{U}$ is a nonempty set of finite objects (the universe of discourse) and $\mathbb{S}$ is a nonempty finite set of attributes so that $a : \mathbb{U} \rightarrow V_a$ for every $a \in \mathbb{S}$. V_a is the set of values that a can take. For any $P \subseteq \mathbb{S}$, there exists an associated equivalence relation $IND(P)$:

$$IND(P) = \{(x, y) \in \mathbb{U}^2 | \forall a \in P, a(x) = a(y)\}. \quad (1)$$

The partition generated by $IND(P)$ is denoted $\mathbb{U}/IND(P)$ and is calculated as follows:

$$\mathbb{U}/IND(P) = \otimes\{\mathbb{U}/IND(\{a\}) : a \in P\}, \quad (2)$$

where,

$$S \otimes T = \{X \cap Y : \forall X \in S, \forall Y \in T, X \cap Y \neq \emptyset\}. \quad (3)$$

If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]_p$. Let $X \subseteq \mathbb{U}$. X can be approximated using only the information contained in P by constructing the P-*lower* and P-*upper* approximations of X:

QUICKREDUCT($\mathbb{C}$,$\mathbb{D}$).
$\mathbb{C}$, the set of all conditional features;
$\mathbb{D}$, the set of decision features.

(1) $R \leftarrow \{\}$
(2) **do**
(3) $T \leftarrow R$
(4) $\forall x \in (\mathbb{C} - R)$
(5) **if** $\gamma_{R \cup \{x\}}(\mathbb{D}) > \gamma_T(\mathbb{D})$
(6) $T \leftarrow R \cup \{x\}$
(7) $R \leftarrow T$
(8) **until** $\gamma_R(\mathbb{D}) == \gamma_{\mathbb{C}}(\mathbb{D})$
(9) **return** R

Fig. 1. The QUICKREDUCT algorithm.

$$\underline{P}X = \{x | [x]_p \subseteq X\}, \quad (4)$$

$$\overline{P}X = \{x | [x]_p \cap X \neq \emptyset\}. \quad (5)$$

Let P and Q be equivalence relations over $\mathbb{U}$, then the concepts of the positive, negative, and boundary regions can be defined:

$$POS_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \underline{P}X, \quad (6)$$

$$NEG_P(Q) = \mathbb{U} - \bigcup_{X \in \mathbb{U}/Q} \overline{P}X, \quad (7)$$

$$BND_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \overline{P}X - \bigcup_{X \in \mathbb{U}/Q} \underline{P}X. \quad (8)$$

By employing this definition of the positive region it is possible to calculate the rough set degree of dependency of a set of attributes Q on a set of attributes P. This can be achieved as follows: For $P, Q \subseteq S$, it can be said that Q depends on P in a degree k $(0 \leq k \leq 1)$, thus the higher the value of k the more dependent Q is upon P. This is denoted $(P \Rightarrow_k Q)$ if:

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|\mathbb{U}|}. \quad (9)$$

The reduction of attributes can be achieved through the comparison of equivalence relations generated by sets of attributes. Attributes are removed such that the reduced set provides identical predictive capability of the decision feature or features as that of the original or unreduced set of features, assuming of course that the data set is consistent. A reduct of set $\mathbb{N}$ is a minimal set of attributes $B \subseteq A$ such that $IND_N(B) = IND_N(A)$. In other words, a reduct is a minimal set of attributes from A that preserves the partitioning of the universe and hence the ability to perform classifications as the whole attribute set A does.

The QUICKREDUCT algorithm shown in Fig. 1 searches for a minimal subset without exhaustively generating all possible subsets. The search begins with an empty subset; attributes which result in the greatest increase in the rough-set dependency value are added iteratively. This process continues until the search produces its maximum possible

dependency value for that data set ($\gamma_c(\mathbb{D})$). Note that this type of search does not guarantee a minimal subset and may only discover a local minimum.

2.2 Rough Set Extensions

There are a number of extensions to the rough set model. However, two approaches of note are variable precision rough sets (VPRS) [38] and the tolerance rough set model (TRSM) [29]. These particular extensions are considered important because they extend the rough set model and utilize the data of the boundary region—albeit indirectly.

2.2.1 Variable Precision Rough Sets

VPRS [38] attempts to introduce an element of "*fuzziness*" to the rough set model and hence (although indirectly) utilize the boundary region information. The principal idea behind VPRS is to allow objects to be classified with an error smaller than a certain (manually) predefined level. However, the introduction of this threshold is contrary to the rough set ideology of operating only on the information contained within the data itself.

Let $X, Y \subseteq U$, the relative classification error is then defined by

$$c(X,Y) = 1 - \frac{|X \cap Y|}{|X|}. \tag{10}$$

It is important to note that $c(X,Y) = 0$, if and only if $X \subseteq Y$.

The degree of inclusion is obtained by allowing a predefined level of error, β, in classification such that

$$X \subseteq_\beta Y \iff c(X,Y) \leq \beta, \quad 0 \leq \beta < 0.5. \tag{11}$$

By employing $\subseteq_\beta$ in place of $\subseteq$, the p-lower and p-upper approximations of set X can now be redefined as

$$\underline{P}X_\beta = \bigcup \{[x]_P \in U/P \mid [x]_P \subseteq_\beta X\}, \tag{12}$$

$$\overline{P}X_\beta = \bigcup \{[x]_P \in U/P \mid c([x]_P, X) < 1 - \beta\}. \tag{13}$$

Note also that $\underline{P}_\beta X = \underline{P}X$ for $\beta = 0$, therefore, degenerating to the traditional rough sets case.

This also allows the positive, negative, and boundary regions to be extended thus:

$$POS_{P\beta}(X) = \underline{P}_\beta X, \tag{14}$$

$$NEG_{P\beta}(X) = \mathbb{U} - \overline{P}_\beta X, \tag{15}$$

$$BND_{P\beta}(X) = \overline{P}_\beta X - \underline{P}_\beta X. \tag{16}$$

It can now be seen that VPRS can be applied in the same way as the traditional rough sets approach described previously.

2.2.2 Tolerance Rough Set Model

A similar approach in some respects to VPRS is the TRSM [29] and like VPRS it can be useful for application to real-valued data. TRSM employs a similarity relation to minimize data as opposed to the indiscernibility relation used in classical rough sets. This allows a relaxation in the way equivalence classes are considered, and the resulting tolerance classes can be generated according to the tolerance threshold which has been specified. Again, like VPRS this threshold is human defined.

In this approach, suitable similarity relations must be defined for each feature, although the same definition can be used for all features if applicable. A standard measure for this purpose, given in [29], is

$$SIM_a(x,y) = 1 - \frac{|a(x) - a(y)|}{|a_{max} - a_{min}|}, \tag{17}$$

where a is a considered feature, and a_{max} and a_{min} denote the maximum and minimum values of a, respectively. When considering the case where there is more than one feature, the defined similarities must be combined to provide an overall measure of similarity of objects. For a subset of features, P, this can be achieved in many ways; some common approaches are

$$(x,y) \in SIM_{P,\tau} \iff \prod_{a \in P} SIM_a(x,y) \geq \tau, \tag{18}$$

$$(x,y) \in SIM_{P,\tau} \iff \frac{\sum_{a \in P} SIM_a(x,y)}{|P|} \geq \tau, \tag{19}$$

where τ is a global similarity threshold and determines the required level of similarity for inclusion within tolerance classes. This framework allows for the specific case of traditional rough sets by defining a suitable similarity measure (e.g., equality of feature values and (15)) and threshold ($\tau = 1$). Further similarity relations are summarized in [21], but are not included here. From this, the so-called tolerance classes that are generated by a given similarity relation for an object x are defined as

$$SIM_{P,\tau}(x) = \{y \in U | (x,y) \in SIM_{P,\tau}\}. \tag{20}$$

Lower and upper approximations can now be defined in a similar way to that of traditional rough set theory:

$$\underline{P_\tau}X = \{x | SIM_{P,\tau}(x) \subseteq X\}, \tag{21}$$

$$\overline{P_\tau}X = \{x | SIM_{P,\tau}(x) \cap X \neq \emptyset\}. \tag{22}$$

The tuple $\langle \underline{P_\tau}X, \overline{P_\tau}X \rangle$ is known as a tolerance rough set [29]. Using this, the positive region and dependency functions can be defined as follows:

$$POS_{P,\tau}(Q) = \bigcup_{X \in U/Q} \underline{P_\tau}X, \tag{23}$$

$$\gamma_{P,\tau}(Q) = \frac{|POS_{P,\tau}(Q)|}{|U|}. \tag{24}$$

From these definitions, an attribute reduction method can be formulated that uses the tolerance-based degree of dependency, $\gamma_{P,t}(Q)$, to measure the significance of feature subsets (in a similar way to the rough set QUICKREDUCT algorithm described in the previous section).

2.2.3 Fuzzy-Rough Approaches

Other hybrid approaches such as rough-fuzzy [23] and fuzzy-rough sets [12], [15], have been proposed in order to improve the ability to deal with uncertainty and vagueness present in data. A fuzzy-rough set [5], [36] is defined by two fuzzy sets, fuzzy lower and upper approximations, obtained by extending the corresponding crisp rough set notions. In the crisp case, elements that belong to the lower approximation (i.e., have a membership of 1) are said to belong to the approximated set with absolute certainty. In the fuzzy-rough case, elements may have a membership in the range [0, 1], allowing greater flexibility in handling uncertainty.

Let $I = (\mathbb{U}, \mathbb{A})$ be an information system, where $\mathbb{U}$ is a nonempty set of finite objects (the universe) and $\mathbb{A}$ is a nonempty finite set of attributes such that $a : \mathbb{U} \rightarrow V_a$ for every $a \in \mathbb{A}$. V_a is the set of values that attribute a may take. For decision systems, $\mathbb{A} = \mathbb{C} \cup \mathbb{D}$ where $\mathbb{C}$ is the set of input or conditional features and $\mathbb{D}$ is the set of decision features.

Fuzzy equivalence classes [5], [6] are central to the fuzzy-rough set approach in the same way that crisp equivalence classes are central to classical rough sets. For typical applications, this means that the decision values and the conditional values may all be fuzzy. The concept of crisp equivalence classes can be extended by the inclusion of a fuzzy similarity relation S on the universe, which determines the extent to which two elements are similar in S. The usual properties of reflexivity ($\mu_S(x, x) = 1$), symmetry ($\mu_S(x, y) = \mu_S(y, x)$) and transitivity ($\mu_S(x, z) \geq \mu_S(x, y) \wedge \mu_S(y, z)$) hold.

Using the fuzzy similarity relation, the fuzzy equivalence class $[x]_S$ for objects close to x can be defined:

$$\mu_{[x]_S}(y) = \mu_S(x, y). \tag{25}$$

The following axioms should hold for a fuzzy equivalence class F:

- $\exists x, \mu_F(x) = 1$,
- $\mu_F(x) \wedge \mu_S(x, y) \leq \mu_F(y)$, and
- $\mu_F(x) \wedge \mu_F(y) \leq \mu_S(x, y)$.

The first axiom corresponds to the requirement that an equivalence class is nonempty. The second axiom states that elements in y's neighborhood are in the equivalence class of y. The final axiom states that any two elements in F are related via the fuzzy similarity relation S. Obviously, this definition degenerates to the normal definition of equivalence classes when S is nonfuzzy. The family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes [5].

The fuzzy lower and upper approximations are fuzzy extensions of their crisp counterparts. Informally, in crisp rough set theory, the lower approximation of a set contains those objects that belong to it with certainty. The upper approximation of a set contains the objects that possibly belong. From the literature, the fuzzy P-lower and P-upper approximations are defined as [5]

$$\mu_{\underline{P}X}(F_i) = \inf_x \max\{1 - \mu_{F_i}(x), \mu_X(x)\}, \quad \forall i, \tag{26}$$

$$\mu_{\overline{P}X}(F_i) = \sup_x \min\{\mu_{F_i}(x), \mu_X(x)\}, \quad \forall i, \tag{27}$$

where $\mathbb{U}/P$ stands for the partition of the universe of discourse, $\mathbb{U}$, with respect to a given subset P of features, and F_i denotes a fuzzy equivalence class belonging to $\mathbb{U}/P$. Note that although the universe of discourse in feature reduction is finite, this is not the case in general, hence the use of sup and inf above. These definitions diverge a little from the crisp upper and lower approximations, as the memberships of individual objects to the approximations are not explicitly available (further discussion can be found in [10]). As a result of this, the fuzzy lower and upper approximations are redefined as [12]

$$\mu_{\underline{P}X}(x) = \sup_{F \in \mathbb{U}/P} \min\Big(\mu_F(x), \inf_{y \in \mathbb{U}} \max\{1 - \mu_F(y), \mu_X(y)\}\Big), \tag{28}$$

$$\mu_{\overline{P}X}(x) = \sup_{F \in \mathbb{U}/P} \min\Big(\mu_F(x), \sup_{y \in \mathbb{U}} \min\{\mu_F(y), \mu_X(y)\}\Big). \tag{29}$$

The tuple $\langle \underline{P}X, \overline{P}X \rangle$ is called a fuzzy-rough set.

For an individual feature, a, the partition of the universe by $\{a\}$ (denoted $\mathbb{U}/IND(\{a\})$) is considered to be the set of those fuzzy equivalence classes for that feature. For example, if two fuzzy sets N_a and Z_a are generated for feature a during fuzzification, the partition $\mathbb{U}/IND(\{a\}) = \{N_a, Z_a\}$. If the fuzzy-rough feature selection process is to be useful, it must be able to deal with multiple features, finding the dependency between various subsets of the original feature set. For instance, it may be necessary to be able to determine the degree of dependency of the decision feature(s) with respect to feature set $P = \{a, b\}$. In the crisp case, $\mathbb{U}/P$ contains sets of objects grouped together that are indiscernible according to both features a and b. In the fuzzy case, objects may belong to many equivalence classes, so the cartesian product of $\mathbb{U}/IND(\{a\})$ and $\mathbb{U}/IND(\{b\})$ must be considered in determining $\mathbb{U}/P$. In general,

$$\mathbb{U}/P = \otimes\{a \in P : \mathbb{U}/IND(\{a\})\}. \tag{30}$$

For example, if $P = \{a, b\}$, $\mathbb{U}/IND(\{a\}) = \{N_a, Z_a\}$, and $\mathbb{U}/IND(\{b\}) = \{N_b, Z_b\}$, then

$$\mathbb{U}/P = \{N_a \cap N_b, N_a \cap Z_b, Z_a \cap N_b, Z_a \cap Z_b\}.$$

Clearly, each set in $\mathbb{U}/P$ denotes an equivalence class. The extent to which an object belongs to such an equivalence class is, therefore, calculated by using the conjunction of constituent fuzzy equivalence classes, say F_i, $i = 1, 2, \ldots, n$:

$$\mu_{F_1 \cap \ldots \cap F_n}(x) = min(\mu_{F_1}(x), \mu_{F_2}(x), \ldots, \mu_{F_n}(x)). \tag{31}$$

3 Distance Measure Assisted Rough Set Attribute Reduction

As discussed previously, almost all techniques for rough set attribute reduction adopt an approach to minimization that examines only the information contained within the lower approximation of a set. Currently, there are no mechanisms in rough-set-based methods to deal with the uncertainty of the boundary region. Any useful information that may be contained in the boundary region is, therefore, lost when only the lower approximation is employed for minimization.

The approach described in this section uses both the information contained in the lower approximation and the information contained in the boundary region to search for reducts. The DMRSAR [16] method uses a distance measure to determine the proximity of objects in the boundary region to those in the lower approximation and assigns a significance value to these distances.

3.1 Distance Metric and Mean Positive Region

The distance metric attempts to qualify the objects in the boundary region with regard to their proximity to the lower approximation. From an intuitive point-of-view, the closer the proximity of an object in the boundary region to objects of the lower approximation, the greater the likelihood that it actually belongs to the set of interest. For the method detailed here, all of the distances of objects in the boundary region are calculated. From this, the significance value for a subset can be obtained.

Since calculating the margin of the lower approximation for an n-dimensional space would involve considerable computational effort, a more pragmatic solution is employed—the mean of all object attribute values in the positive region (POS_P) or union of lower approximations is calculated. This can be defined as follows:

$$POS_{P_{MEAN}} = \left\{ \frac{\sum_{o \in \underline{P}X} a(o)}{|POS_P X|} : \forall a \in P \right\}. \tag{32}$$

Using this definition of the mean of the P positive region, the distance function for the proximity of objects in the boundary region from the P positive region mean can be defined by

$$\delta_P\big(POS_{P_{MEAN}}, y\big), \quad y \in BND_P(Q). \tag{33}$$

Clearly this definition only holds true if either $POS_{P_{MEAN}}$ or $BND_P(Q)$ is nonempty.

The exact distance function is not defined here as a number of strategies may be employed for the calculation of the distance of objects in the boundary. In Section 3.4, a euclidean type distance metric is employed.

In order to measure the quality of the boundary region, a significance value ω for subset P is calculated by obtaining the sum of all object distances and inverting it such that

$$\omega_P(Q) = \left(\sum_{y \in BND_P(Q)} \delta_P\big(POS_{P_{MEAN}}, y\big) \right)^{-1}. \tag{34}$$

It is important to note that if $POS_P(Q) = \emptyset$ there are no certain objects from which to generate a $POS_{P_{MEAN}}$, in which case no distance function can be defined and hence the significance degree $\omega_P(Q)$ is set to 0. Also, when $BND_P(Q) = \emptyset$ there is no uncertainty about the concept being approximated and so there are no uncertain objects to measure using the distance function, in which case the significance degree $\omega_P(Q)$ is set to its maximum value of 1.

This significance measure is used in conjunction with the rough-set dependency value to gauge the utility of attribute subsets in a similar way to that of the rough-set dependency measure. As one measure only operates on the objects in the lower approximation and the other only on the objects in the boundary, both entities are considered separately and then combined to create a new evaluation measure M:

$$M_P(Q) = \frac{\omega_P(Q) + \gamma_P(Q)}{2}. \tag{35}$$

Obviously, if $\gamma_P(Q) = 1$, then the concept being approximated has no uncertainty with respect to P and by (33) $\omega_P(Q) = 1$ also. A mean of both values is obtained as both operate in the range [0, 1] and this allows the approach to be data driven. Initial investigative work has shown that manual manipulation of the participation can affect subset selection and even improve results (see conclusion and future work section). A new feature selection mechanism can be constructed that uses both the significance value and the rough dependency value to guide the search for the best feature subset.

An alternative to the mean positive region and distance metric is another approach which uses the *Hausdorff* metric to calculate the distance between nonempty sets. It measures the extent to which each point in a set is located relative to those of another set. The *Hausdorff* metric has been applied to facial recognition [26], image processing [27], and FS [22] with much success. It can be defined as

$$h(A, B) = \max_{a \in A} \Big\{ \min_{b \in B} \{ d(a, b) \} \Big\}, \tag{36}$$

where a and b are points (objects) of sets A and B, respectively, and $d(a, b)$ is any metric between these points. A basic implementation of this has been incorporated into the above framework using euclidean distance as a metric. Experimentation using this approach can be found in Section 5. The primary disadvantage to this approach, however, is the computational overhead involved in calculating the distance of all objects in the boundary region from each other. For n boundary region objects, this means that $O(n^2)$ distance calculations must be made, unlike the mean positive region which results in $O(n)$ distance calculations.

DMQUICKREDUCT($\mathbb{C}$,$\mathbb{D}$).
$\mathbb{C}$, the set of all conditional features;
$\mathbb{D}$, the set of decision features.

(1) $T \leftarrow \{\}, R \leftarrow \{\}$
(2) **do**
(3) $\quad \forall x \in (\mathbb{C} - R)$
(4) $\quad\quad$ **if** $M(R \cup \{x\}) > M(T)$
(5) $\quad\quad\quad T \leftarrow R \cup \{x\}$
(6) $\quad R \leftarrow T$
(7) **until** $\gamma_R(\mathbb{D}) == \gamma_{\mathbb{C}}(\mathbb{D})$
(8) **return** R

Fig. 2. The rough-set distance-metric-based QUICKREDUCT algorithm.

3.2 Distance Measure-Based Selection Algorithm

Fig. 2 shows a rough-set-based DMQUICKREDUCT algorithm based on the previously described rough-set-based algorithm in Fig. 1.

DMQUICKREDUCT is similar to the RSAR algorithm but uses a combined distance and rough-set dependency value of a subset to guide the feature selection process. If the

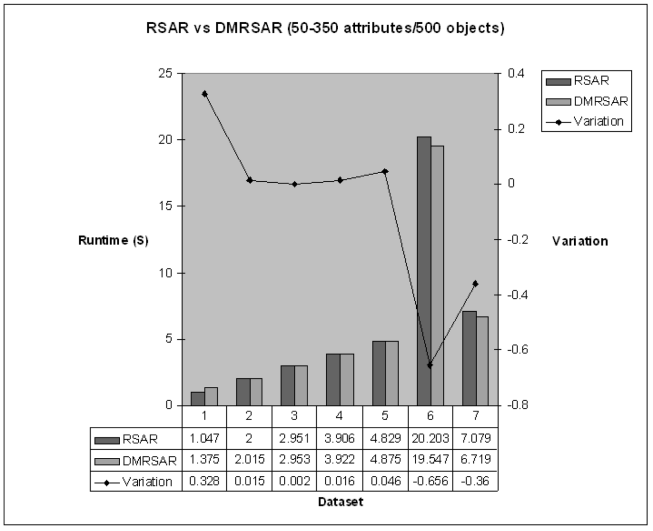

Fig. 3. RSAR and DMRSAR runtimes for 50-350 attributes.

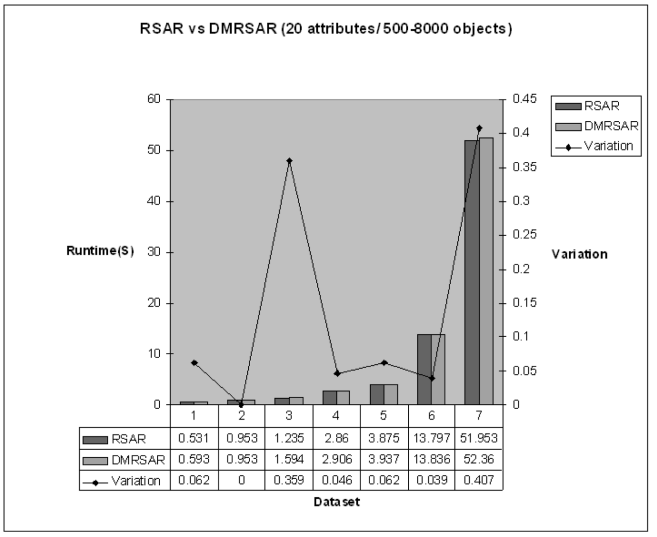

Fig. 4. RSAR and DMRSAR runtimes for 500-8,000 objects.

combined value M of the current reduct candidate is greater than that of the previous, then this subset is retained and used in the next iteration of the loop. It is important to point out that the subset is evaluated by examining the value of M, termination only occurs when the addition of any remaining features results in the dependency function value (γ_T) reaching that of the unreduced data set. The value of M is, therefore, not used as a termination criterion.

The algorithm begins with an empty subset R. The do-until loop works by examining the combined dependency and significance value of a subset and incrementally adding a single conditional feature at a time. For each iteration, a conditional feature that has not already been evaluated will be temporarily added to the subset R. The combined measure of the subset currently being examined (line 6) is then evaluated and compared with that of T (the previous subset). If the combined measure of the current subset is greater, then the attribute added in (line 5) is retained as part of the new subset T (line 6).

The loop continues to evaluate in the above manner by adding conditional features, until the dependency value of the current reduct candidate ($\gamma_R(\mathbb{D})$) equals the consistency of the data set (1 if the data set is consistent).

3.3 Computational Complexity

As the DMRSAR algorithm is based on a greedy hill-climbing type of search. The computational complexity will be similar to that of other approaches which use this method. However, in addition to the factors which govern the computational complexity of the rough set QUICKREDUCT algorithm, other factors must also be taken into account. In the DMRSAR approach objects in the boundary region are also considered and this inevitably adds to the computational overhead. Furthermore, all of those objects in the lower approximation are also considered when calculating a positive region object for each concept—where the objects of the positive region are "collapsed" to form a single representative object. At this lower level, the additional factors that must be considered (also those that are not employed in the rough set case) include the calculation of the collapsed lower approximation mean, the calculation of the upper approximation, and the calculation of the distances of objects in the boundary from the collapsed lower approximation mean.

From a high level point-of-view the DMQUICKREDUCT has an intuitive complexity of $(n^2 + n)/2$ for a dimensionality of n. This is the number of evaluations of the dependency function and distance measure performed in the "worst case." For instance, if the feature set consists of $\{a_1, a_2\}$, then the DMQUICKREDUCT algorithm will make three evaluations, one each for $\{a_1\}$ and $\{a_2\}$, and finally one for $\{a_1, a_2\}$ in the worst case.

In an attempt to compare the complexity of both the RSAR and DMRSAR approaches from an application viewpoint, a number of artificial data sets were generated. There are 14 data sets in total ranging from 20 to 350 attributes, and 500 to 8,000 objects. Both FS approaches were applied to these data sets and the time taken to find a reduct was recorded in each case. The results show that there is only a marginal increase in runtime for the DMRSAR approach. There is even a decrease in some cases, but this relates to the fact that DMRSAR found smaller subsets than RSAR in these particular cases. However, Figs. 3 and 4 demonstrate that for increased dimensionality and numbers of objects there is little overall difference in runtime between the approaches.

3.4 A Worked Example

To illustrate the operation of the new distance measure-based algorithm, a small example data set is considered, containing discrete-valued conditional and decision attributes. The data used in the experimentation section are real-valued; however, crisp data are used in this example to aid explanation of the approach. Note also for brevity, that only the selection of two subsets is shown here.

Table 1 contains seven objects. It has four crisp-valued conditional attributes and a single crisp-valued decision attribute.

If attribute d is considered for selection, for example, the lower and upper approximations must first be calculated:

$$\underline{\{d\}} = \{\{\emptyset\}, \{2\}, \{\emptyset\}\}.$$

TABLE 1
Example Data Set: Crisp Attributes

Object	a	b	c	d	e
0	1	0	2	2	0
1	0	1	0	0	2
2	1	0	0	1	1
3	1	0	0	2	2
4	1	2	0	0	1
5	1	2	0	2	0
6	0	1	2	0	1

Similarly, for the upper approximation:

$$\overline{\{d\}} = \{\{0,3,5\},\{1,2,4,6\},\{0,1,3,4,6\}\}.$$

Having calculated the upper and lower approximations for $\{d\}$, the positive and boundary regions can be shown to be

$$\begin{aligned} POS_{\{d\}}(\{e\}) &= \bigcup\{\emptyset,\{2\}\} = \{2\},\\ BND_{\{d\}}(\{e\}) &= \bigcup\{\{0,3,5\},\{2\}\\ &\quad \{1,4,6\},\{1,4,6\}\} - \{2\}\\ &= \{0,1,3,4,5,6\}. \end{aligned}$$

The rough-set dependency, the positive region mean, and object distances can now all be calculated. As mentioned in the previous section, there are many distance metrics which can be applied to measure the distance of the objects in the boundary from the lower approximation mean. For simplicity, a variation of euclidean distance is used in the approach documented here, and this is defined as

$$\delta_P(POS_{P_{MEAN}}, y) = \sqrt{\sum_{a\in P} f_a\big(POS_{P_{MEAN}}, y\big)^2}, \qquad (37)$$

where

$$\begin{aligned} f_a\big(POS_{P_{MEAN}}, y\big) = 1 &\Longleftrightarrow a\big(POS_{P_{MEAN}}\big) \neq a(y)\\ &= 0\ otherwise. \end{aligned}$$

From this, the distances of all of the objects in the boundary region in relation to the lower approximation mean can now be calculated.

As there is only a single object in the lower approximation, the mean of the lower approximation does not need to be calculated in this case. The individual distances of objects in the boundary of $\{d\}$ can be shown to be

$$\begin{aligned} obj\ 0\ &\sqrt{f_d\big(POS_{P_{MEAN}}, 0\big)^2} = 1,\\ obj\ 1\ &\sqrt{f_d\big(POS_{P_{MEAN}}, 1\big)^2} = 1,\\ obj\ 3\ &\sqrt{f_d\big(POS_{P_{MEAN}}, 3\big)^2} = 1,\\ obj\ 4\ &\sqrt{f_d\big(POS_{P_{MEAN}}, 4\big)^2} = 1,\\ obj\ 5\ &\sqrt{f_d\big(POS_{P_{MEAN}}, 5\big)^2} = 1,\\ obj\ 6\ &\sqrt{f_d\big(POS_{P_{MEAN}}, 6\big)^2} = 1. \end{aligned}$$

Where there is more than one object in the potential reduct lower approximation, calculating the $POS_{P_{MEAN}}$ object can be achieved in the manner described in the previous section. Examine all of those attribute values for each of the objects that appear in the lower approximation of the considered subset. For example, considering the subset $\{a,d\}$, the lower approximation and boundary regions are

$$\begin{aligned} POS_{\{a,d\}}(\{e\}) &= \bigcup\{\emptyset,\{2\},\{4\}\},\\ BND_{\{a,d\}}(\{e\}) &= \bigcup\{\{0,3,5\},\{0,3,5\}\{1,6\},\{1,6\}\}\\ &= \{0,1,3,5,6\}. \end{aligned}$$

The attribute values for $\{a,d\}$ for objects $\{2,4\}$ can be obtained by referring to Table 1:

$$\begin{aligned} for\ \{a\} : object\ 2 &= '1',\\ object\ 4 &= '1',\\ for\ \{d\} : object\ 2 &= '1',\\ object\ 4 &= '0'. \end{aligned}$$

This results in $POS_{P_{MEAN}} = \{1, 0.5\}$ for $\{a,d\}$.

These real-valued numbers, however, are not meaningful when dealing with crisp-valued data (1 is considered as different from 1.1 as it is from 100). The strategy employed to address this problem was to examine all of the attribute values for the attribute in question and assign it a value which appears in that range of values to which it is closest in terms of magnitude. So, as the $POS_{P_{MEAN}}$ value for the attribute a is an existing value, this does not need to be considered; the $POS_{P_{MEAN}}$ value assigned to d, however, is not in the range of values taken by the attribute d. Values of 0.5 or less are considered to be closer to 0, thus approximated to '0', and becomes $POS_{P_{MEAN}} = \{1, 0\}$.

Again by utilization of euclidean distance and the new $POS_{P_{MEAN}}$, the distances of objects in the boundary region can be calculated:

$$\begin{aligned} ob\ 0\ &\sqrt{\big(f_a(POS_{P_{MEAN}}, 0\big)^2 + f_d\big(POS_{P_{MEAN}}, 0\big)^2\big)} = 1,\\ ob\ 1\ &\sqrt{\big(f_a(POS_{P_{MEAN}}, 1\big)^2 + f_d\big(POS_{P_{MEAN}}, 1\big)^2\big)} = 1,\\ ob\ 3\ &\sqrt{\big(f_a(POS_{P_{MEAN}}, 3\big)^2 + f_d\big(POS_{P_{MEAN}}, 3\big)^2\big)} = 1,\\ ob\ 5\ &\sqrt{\big(f_a(POS_{P_{MEAN}}, 5\big)^2 + f_d\big(POS_{P_{MEAN}}, 5\big)^2\big)} = 1,\\ ob\ 6\ &\sqrt{\big(f_a(POS_{P_{MEAN}}, 6\big)^2 + f_d\big(POS_{P_{MEAN}}, 6\big)^2\big)} = 1. \end{aligned}$$

It is perhaps worth noting at this point, that although a form of euclidean distance is used to calculate the distance of the objects from the $POS_{P_{MEAN}}$, in calculating that distance, the difference between two values is always considered in Boolean terms for crisp data. The reason for this is that the values are states rather than real-valued. This means that if the value for a particular attribute in the $POS_{P_{MEAN}}$ happened to be 1 and that of the corresponding attribute value of an object in the boundary region was 1,563, the difference between these two states would be $(1 - 1{,}563) = 1$. For real-valued data, however, this would not be the case as the values of attributes are real numerical values.

Although the individual distances may be useful in identifying objects that are similar to those in the lower approximation, they are not individually indicative of the

subset goodness. A method of achieving this measure is to calculate the sum of all of the distances and invert it thus giving a significance value to each subset considered for selection. The significance value is real-valued and has membership in the range [0, 1] for the purpose of dealing with crisp data.

Thus, for $\{a, d\}$:

$$\omega_{\{a,d\}}(\{e\}) = (1 + 1 + 1 + 1 + 1)^{-1} = 0.2.$$

Although the significance measure alone can be used to search for subsets, the results from some initial investigation indicated that these were not of equal quality as those returned by RSAR. So, the significance value was combined with the rough-set dependency value. This results in a combined metric in which both dependency and significance have equal participation.

By calculating the change in combined significance and dependency value (M) when an attribute is removed from the set of considered conditional attributes, a measure of the goodness of that attribute can be obtained. The greater the change in M the greater the measure of goodness that attribute has attached to it.

Using the previous examples of the DMRSAR method, the values for the combined metric can be calculated for all considered subsets of $\mathbb{C}$ using DMRSAR:

$$\begin{aligned} M_{\{b\}}(\{e\}) &= 0.0 \quad M_{\{b,d\}}(\{e\}) = 0.3910, \\ M_{\{c\}}(\{e\}) &= 0.0 \quad M_{\{c,d\}}(\{e\}) = 0.3026, \\ M_{\{d\}}(\{e\}) &= 0.342 \quad M_{\{a,b,d\}}(\{e\}) = 0.3026, \\ M_{\{a,d\}}(\{e\}) &= 0.2425 \quad M_{\{b,c,d\}}(\{e\}) = 1.0. \end{aligned}$$

It is obvious from the above example that the search finds a subset in the manner $\{d\} \rightarrow \{b, d\} \rightarrow \{b, c, d\}$. As $\{a, d\}$ and $\{c, d\}$ and also $\{a, b, d\}$ do not result in the same increase in combined metric these subsets are ignored.

4 Experimentation

This section presents the results of experimental studies using both crisp-valued and real-valued data sets. The DMRSAR method is initially compared with a rough-set-based feature selection method (RSAR) [2], and PCA [7]. Additionally, DMRSAR is also compared with fuzzy-rough-set-based FS (FRFS) [12] and a tolerance rough-set-based feature selection method [29] for real-valued data. It is important to note that DMRSAR operates on discretized versions of the real-valued data sets listed. All of the data sets presented are of the same format as that used in the example of the previous section. All data have been obtained from [1] and [19]. A comparison of the RSAR, FRFS, and distance-based dimensionality reduction techniques is made based on subset size, classification accuracy, and time taken to discover subsets.

4.1 Classifiers

Three classifier learners were employed for the classification of the data, JRip, J48, and PART [34].

J48 [25] creates decision trees by choosing the most informative features and recursively partitioning the data into subtables based on their values. Each node in the tree represents a feature with branches from a node representing the alternative values this feature can take according to the current subtable. Partitioning stops when all data items in the subtable have the same classification. A leaf node is then created, and this classification assigned.

JRip [3] learns propositional rules by repeatedly growing rules and pruning them. During the growth phase, antecedents are added greedily until a termination condition is satisfied. Antecedents are then pruned in the next phase subject to a pruning metric. Once the rule set is generated, a further optimization is performed where rules are evaluated and deleted based on their performance on randomized data.

PART [35] generates rules by means of repeatedly creating partial decision trees from data. The algorithm adopts a divide-and-conquer strategy such that it removes instances covered by the current rule set during processing. Essentially, a rule is created by building a pruned tree for the current set of instances; the leaf with the highest coverage is promoted to a rule.

4.2 Comparison of DMRSAR and RSAR

In this section, DMRSAR is compared with RSAR [2]. Results are presented both in terms of subset size and classification accuracy. The data sets employed range in size from 47 to 2,000 objects and between 7 and 57 attributes. Conditional attributes and decision attributes are crisp and discrete-valued.

4.2.1 Classification Accuracy

The data presented in Table 2 show the average classification accuracy as a percentage using each of the previously described classifiers. The classification was initially performed on the unreduced data set, followed by the reduced data sets which were obtained by using the RSAR and DMRSAR dimensionality reduction techniques, respectively.

Noting the classification results, the DMRSAR approach performs very well and shows increases in classification accuracies for at least one classifier where there has been a corresponding decrease in dimensionality (e.g., *credit, exactly*, etc.) Notably for the *exactly* data set DMRSAR shows an increase of up to 30 percent while simultaneously demonstrating a reduction in dimensionality. Even where there has been a decrease in the case of some classifiers and data sets which are of similar size to those of RSAR, this decrease is insignificant. Indeed, DMRSAR may sometimes discover subsets of similar size (but contain different features) to RSAR yet demonstrate an increase in classification accuracy (e.g., *derm, ionosphere, heart*).

4.2.2 Subset Size and Runtimes

Table 3 shows a comparison of subset size, and runtimes for both the RSAR and DMRSAR approaches. Although it still at least equals RSAR in terms of performance. However, despite this, it does show that there are gains to be made with crisp-valued data (*credit, exactly, exactly2, wq*), demonstrating that there is much information contained in the boundary region.

There is little relative increase in runtimes when comparing RSAR with DMRSAR, indeed DMRSAR sometimes demonstrates a reduction in dimensionality along

TABLE 2
Average Classification Accuracy—Crisp Data

	J48			JRip			PART		
Dataset	Unreduced	RSAR	DMRSAR	Unreduced	RSAR	DMRSAR	Unreduced	RSAR	DMRSAR
credit	72.60	70.10	69.59	71.90	72.0	70.30	68.9	68.3	69.4
derm	95.90	80.32	82.51	92.07	77.32	76.77	95.62	78.76	81.96
derm2	95.25	94.41	94.41	93.85	92.73	93.85	92.73	93.85	93.85
ionosphere	85.21	88.26	87.39	86.60	84.34	86.93	86.08	87.39	89.56
exactly	85.5	69.4	98.1	69.30	68.00	91.30	92.10	67.32	99.20
exactly 2	74.9	74.9	73.1	75.0	75.0	74.8	74.2	74.20	78.20
heart	77.89	80.95	81.29	79.59	76.10	77.55	77.21	78.57	81.63
led	100	100	100	100	100	100	100	100	100
lung	84.38	84.38	78.12	68.75	84.38	68.75	71.88	84.38	78.12
m-of-n	100	100	100	97.3	98.60	98.60	100	100	100
monk3	100	100	100	99.76	99.07	99.07	100	100	100
soybean	91.35	89.84	87.59	88.72	88.72	80.89	92.10	87.96	84.21
tic-tac-toe	92.38	88.10	87.89	98.32	91.10	91.44	95.30	87.68	87.68
vote	93.67	93.67	93.67	95.00	93.67	93.67	91.67	93.67	93.67
wq	71.07	64.87	67.37	70.44	68.71	67.51	67.17	65.25	66.02

TABLE 3
Comparison of Reduct Size and Runtimes—Crisp Data

	Original number	Reduct size		Time taken to locate reduct	
Dataset	of features	RSAR	DMRSAR	RSAR	DMRSAR
credit	21	9	8	0.937	1.656
derm	35	7	7	0.625	0.625
derm2	35	10	10	0.578	0.640
ionosphere	35	8	8	0.313	0.313
exactly	14	9	8	0.203	0.172
exactly2	14	13	10	0.328	0.235
heart	14	7	7	0.188	0.188
led	25	12	12	2.168	2.375
lung	57	4	4	0.125	0.132
m-of-n	14	8	7	0.171	0.142
monk3	7	3	3	0.063	0.063
soybean	36	12	12	0.797	0.828
tic-tac-toe	10	8	8	0.188	0.203
vote	17	9	9	0.157	0.172
wq	39	15	14	3.250	2.766

with a reduction in runtime. Considering also that no runtime optimization has been performed for DMRSAR, these results are very encouraging, but also show that there is some improvement required in terms of the mean positive region calculation which would result in more accurate measurement of distances.

4.3 Comparison of DMRSAR and PCA

PCA [7] is a versatile transformation-based DR technique which projects the data onto a new coordinate system of reduced dimensions. This process of linear transformation, however, also transforms the underlying semantics or meaning of the data. This results in data that are difficult for humans to interpret, but which may still provide useful automatic classification of new data. In order to ensure that the comparison of DMRSAR and PCA is balanced, the same subset sizes discovered for each data set are also employed in the analysis of PCA. Each of the best number of transformed features are also utilized for PCA.

The results in Table 4 show that of the 15 data sets only *credit*, *derm*, and *tic-tac-toe* demonstrate a small decrease in classification accuracy performance when compared with DMRSAR. These decreases are small in magnitude and DMRSAR outperforms PCA in all other cases, sometimes significantly.

TABLE 4
Subset Size and Classification Accuracy Results for PCA

Dataset	(predefined) subset size	J48	JRIP	PART
credit	8	71.10	71.00	72.00
derm	7	90.40	94.08	93.98
derm2	10	93.29	91.34	93.52
ionosphere	8	81.30	76.95	79.12
exactly	8	67.80	66.70	68.60
exactly2	10	75.90	74.30	75.80
heart	7	77.89	79.58	77.21
led	12	99.38	98.55	99.38
lung	4	71.85	68.75	65.62
m-of-n	7	76.2	73.30	75.30
monk3	3	77.77	76.62	77.31
soybean	12	77.81	72.18	75.18
tic-tac-toe	8	96.18	94.57	95.92
vote	9	89.00	89.00	87.67
wq	14	67.32	67.37	66.41

TABLE 5
Classification Accuracy of Unreduced, DMRSAR Reduced, and FRFS Reduced, Data Using JRIP, PART, and J48 Classifiers

	Unreduced data			DMRSAR reduced data			FRFS reduced data		
Dataset	JRIP	PART	J48	JRIP	PART	J48	JRIP	PART	J48
water 2	83.84	83.33	85.64	85.89	84.36	86.67	84.36	82.56	80.26
water 3	81.28	77.43	79.48	81.79	83.33	79.74	82.05	78.97	79.74
cleveland	52.18	51.85	50.16	53.53	51.51	54.20	55.55	52.18	53.87
glass	67.75	67.28	67.75	69.62	72.89	69.15	69.62	69.62	68.22
heart	77.40	76.66	73.30	82.22	81.82	77.78	80.00	78.51	75.55
ionosphere	86.52	87.82	86.26	84.78	86.10	86.10	87.82	91.30	91.30
olitos	70.83	67.50	57.50	67.33	67.50	68.33	70.83	67.50	62.50
wine	92.69	94.33	93.82	95.86	94.94	93.25	88.76	92.13	93.82

It should be emphasized, however, that while PCA might marginally outperform DMRSAR in three instances in terms of classification accuracy, the semantics of the data is irreversibly transformed following dimensionality reduction. This can have consequences where human interpretability of the data is important, which is one of the key reasons for performing feature selection tasks to begin with. As DMRSAR is a *feature selection* approach as opposed to a *feature ranking* method, a predefined threshold is not required; selection is complete as soon as the termination criterion (rough-set dependency) is fulfilled. The rough-set dependency value is integral to the selection process and as such, in contrast to PCA does not need to be predefined.

Finally, it is worth noting that PCA is selected for comparison here due to recognition of the fact that it is an established approach for dimensionality reduction.

4.4 Comparison of DMRSAR and FRFS

The real-valued data used in this section comprise of data sets which are small-to-medium in size, with between 120 and 390 objects per data set and feature sets ranging from 5 to 39. The unreduced data classification is illustrated in Table 5. The data have been discretized for use with DMRSAR as it is unable to handle real-valued data. The DMRSAR selected subsets are, however, employed when reducing and classifying the original real-valued data.

4.4.1 Classification Accuracy

It is interesting to note that where a decrease in classification accuracy is recorded for FRFS, with respect to the unreduced data the same is also true for DMRSAR. This decrease in classification accuracy is small when comparing both FRFS and DMRSAR approaches to the unreduced data. Also, when comparing classification results, where the DMRSAR approach shows a fall in classification accuracy, the corresponding reduction in dimensionality (shown in Table 5) is significantly better than that of FRFS.

4.4.2 Subset Size and Runtimes

It is clear also from the runtime figures that DMRSAR runs considerably faster than FRFS. This primarily, can be attributed to the computational complexity of FRFS which is related to the time taken in calculating fuzzy equivalence classes. Clearly, DMRSAR has a considerable advantage in this respect as the figures in Table 4 demonstrate.

The advantages of the DMRSAR method in terms of subset size are more pronounced when compared with FRFS than those for RSAR as demonstrated in Table 6. This is a strong indicator that the approach is perhaps more efficient when applied to domains where the data are real-valued; this is borne out by the marked contrast between the subset-size results obtained for both approaches. There are, however, two data sets where DMRSAR fails to outperform FRFS in terms of subset size—*water 2* and *water 3* (see conclusion and future work for further discussion of this). However, it should be noted that FRFS is considerably more mature and refined in terms of both research effort and development.

4.5 Comparison of TRSM and DMRSAR

In this section, an extension of the rough set model—the TRSM [29] is compared with DMRSAR. TRSM employs a similarity relation to minimize data as opposed to the indiscernibility relation used in classical rough sets. This allows a relaxation in the way equivalence classes are considered. This flexibility allows a blurring of the boundaries of the former rough or crisp equivalence classes and objects may now belong to more than one tolerance

TABLE 6
Comparison of Subset Size, Dependency Value, and Runtimes—FRFS

	Original number of		Subset size		Time taken to locate subset	
Dataset	features	objects	FRFS	DMRSAR	FRFS	DMRSAR
water 2	39	390	11	12	96.58	0.860
water 3	39	390	12	18	158.73	1.266
cleveland	14	297	11	9	24.11	0.219
glass	10	214	9	6	1.61	0.156
heart	14	270	11	10	11.84	0.158
ionosphere	35	230	5	4	0.488	0.512
olitos	26	120	10	8	11.20	0.156
wine	14	178	10	8	1.42	0.125

TABLE 7
Comparison of Subset Size for Each Tolerance Threshold Value

Dataset	Original number of features	TRSM $\tau = 0.90$	TRSM $\tau = 0.95$
water 2	39	8	12
water 3	39	9	12
cleveland	14	11	8
glass	10	3	8
heart	14	12	8
ionosphere	34	6	8
olitos	25	9	6
wine	13	5	5

class, thus allowing the consideration of real-valued data. Thus, as for FRFS real-valued data are also employed for the evaluation of this approach.

The ideal tolerance threshold value can be obtained by repeated experimentation for a given data set. This is where the TRSM diverges from the approaches to which DMRSAR has been compared up until now, which have all been data driven. Further work which examines a non-data-driven feature selection approach and which utilizes the boundary region of the TRSM can be found in [17]. For the comparison of DMRSAR and TRSM, results are presented in the following sections for two different values of tolerance threshold (τ)—0.90, and 0.95.

4.5.1 Subset Size

The subset sizes for both values of tolerance threshold are outlined in Table 7. The results demonstrate that the TRSM method can sometimes outperform both FRFS and DMRSAR in terms of subset size. However, it should be borne in mind that the TRSM is not completely data driven and much experimentation may be required before optimal results are achieved for each individual data set. Additionally, the results also demonstrate that the TRSM method does not perform consistently and in some cases returns a larger subset while simultaneously displaying a decrease in classification accuracy.

4.5.2 Classification Accuracy

The results presented in Tables 8 and Tables 9 show that DMRSAR when compared with TRSM performs favorably.

TABLE 8
Classification Accuracy Using JRIP, PART, and J48 Classifiers ($\tau = 0.90$)

Dataset	TRSM JRIP	TRSM PART	TRSM J48
water 2	85.38	82.30	87.43
water 3	80.00	81.53	76.67
cleveland	54.20	53.87	52.52
glass	65.88	69.15	68.69
heart	79.25	75.19	78.88
ionosphere	85.65	86.52	85.21
olitos	70.00	65.83	61.66
wine	96.06	94.94	96.62

TABLE 9
Classification Accuracy Using JRIP, PART, and J48 Classifiers ($\tau = 0.95$)

Dataset	TRSM JRIP	TRSM PART	TRSM J48
water 2	82.82	83.07	82.05
water 3	81.02	80.77	81.02
cleveland	50.54	50.84	54.54
glass	69.62	68.22	69.62
heart	80.38	78.57	81.48
ionosphere	86.08	87.39	87.39
olitos	64.16	66.67	64.16
wine	93.25	95.50	96.02

The results obtained for both tolerance values, show that for four of the eight data sets for $\tau = 0.9$, the TRSM performs poorly and for the remaining four data sets the results are comparable. When $\tau = 0.95$, DMRSAR outperforms TRSM in six cases. The TRSM, however, defeats DMRSAR marginally for the *ionosphere* data set but the corresponding subset is twice the size. The remaining data set, *wine*, shows a classification result that is comparable to DMRSAR.

4.6 Hausdorff Metric Implementation

The Hausdorff metric approach to distance measurement has been described previously as an alternative to the mean positive region and euclidean distance-based method which was used to generate the empirical results shown above.

The DMRSAR approach was augmented with the Hausdorff metric to measure the distance between the lower approximation and the boundary region was implemented in order to investigate the performance of this method in terms of subset size and runtimes. The results of this investigation are included here in Table 10.

It is apparent that this particular implementation of the *Hausdorff* metric fails to capture the useful information of the boundary region in the same way that the mean positive region method does. Examining the results for subset size, it can be seen that the existing DMRSAR approach returns superior results in all cases. Perhaps

TABLE 10
DMRSAR-Hausdorff Metric Implementation Subset Size and Runtimes

Dataset	DMRSAR Subset Size	Hausdorff Metric Subset Size	Hausdorff Metric Runtime
credit	8	10	41.64
derm	7	32	19.343
derm2	10	32	18.437
ionosphere	8	28	7.000
exactly	8	13	17.422
exactly2	10	13	19.250
Heart	7	10	1.734
LED	12	13	566.05
lung	4	5	0.484
m-of-n	7	9	22.03
monk3	3	6	0.422
soybean	12	19	23.518
tic-tac-toe	8	8	5.859
vote	9	9	3.205
wq	14	27	57.031

even more apparent is the results for the runtimes with the LED data set taking 566 s to run. This was to be expected as there are a large number of distance calculations performed even for small data sets (exponential $O(n^2)$ for n upper approximation objects).

5 Conclusion and Future Work

Comparison of DMRSAR with FRFS has shown that the DMRSAR method is a good starting point for further work based on the distance metric for exploring the boundary region of rough sets. The subset size results show that there is still some additional optimization required in order to equal FRFS. Classification accuracy results have been shown to be very similar to those of FRFS, and in some cases the DMRSAR method has even shown an increase while simultaneously demonstrating a reduction in dimensionality. Where a decrease has been observed in relation to FRFS, it has been small and, as discussed previously, the actual decrease is not significant. It is perhaps worth noting that FRFS is a state-of-the-art approach when considering real-valued data and much research effort has been invested in refining its performance.

Additional comparison with a TRSM-based feature selection method has demonstrated that while this method may sometimes marginally outperform DMRSAR, it requires an additional thresholding value. In order to determine the optimal value, however, repeated experimentation is required for each data set. DMRSAR requires no such thresholding value and relies only on the information in the data.

The experimental evaluation emphasizes how much useful information is contained in the boundary region of a rough set. However, it is clear from the results obtained in the previous section that an increase in the efficiency of the DMRSAR algorithm is highly desirable and will lead to further increases in performance. The experimental work detailed in this paper did not take advantage of any optimizations that are expected to reflect this.

Future work would include a reevaluation of how the mean positive region is calculated. Implementation of a more accurate calculation of the lower approximation boundary would mean that distances of objects in the boundary region could be more accurately measured. Other aspects of the DMRSAR approach which require investigation include the current distance metric. For the worked example described in this paper, a euclidean distance metric is employed. Metrics such as Mahalanobis distance, ellipsoid distance, and others could also be considered.

It is also expected that a more thorough investigation of the *Hausdorff* metric and methods of implementation would improve both performance and runtimes. Indeed a *fuzzy-Hausdorff* version of the distance metric may be useful in dealing with real-valued data as this would reflect the continuous nature of the data.

The significance measure which is employed for DMRSAR is quite basic, and considers the boundary region as a single significance value which is expressed as membership value of a unary fuzzy set. By redefining this as a number of fuzzy sets, the boundary region could be quantified more accurately by expressing membership in terms of weights of objects in the boundary in relation to distance from the mean positive region. Apart from the use of extra fuzzy sets, the way in which objects in the boundary are correlated is another area which is worthy of investigation. By examining the correlation of objects and their individual distances, it may be possible to qualify the individual objects and their information value.

In the DMRSAR approach described in Section 3.4, both the dependency measure and the significance value (distance measure) are allowed equal weighting as part of the combined evaluation metric. However, by assigning different weights to each of these metrics, more or less significance can be placed on either measure. Some initial work has been undertaken in this area which indicates that further improvements in performance in terms of subset size and classification accuracy can be obtained through the adoption of such a strategy. A more thorough investigation would explore the level of participation of each evaluation metric and ascertain the effect this has on performance.

Given that the extension of RSAR for the consideration of objects in the boundary region has shown to be successful, additional work in the form of an investigation of the boundary region of tolerance rough sets [17] has also been carried out.

Other areas of application which are currently being investigated include classification and the use of the boundary region of fuzzy-rough sets.

Acknowledgments

The authors wish to thank both the referees and the editor for their invaluable advice in revising this paper.

References

[1] C. Armanino, R. Leardi, S. Lanteri, and G. Modi, "Chemometric Analysis of Tuscan Olive Oils," *Chemometrics and Intelligent Laboratory Systems*, vol. 5, no. 4, pp. 343-354, Apr. 1989.

[2] A. Chouchoulas and Q. Shen, "Rough Set-Aided Keyword Reduction for Text Categorisation," *Applied Artificial Intelligence*, vol. 15, no. 9, pp. 843-873, 2001.

[3] W.W. Cohen, "Fast Effective Rule Induction," *Proc. 12th Int'l Conf. Machine Learning*, pp. 115-123, 1995.

[4] J.S. Deogun, V.V. Raghavan, and H. Sever, "Exploiting Upper Approximation in the Rough Set Methodology," *Proc. First Int'l Conf. Knowledge Discovery and Data Mining*, pp. 1-10, 1995.

[5] D. Dubois and H. Prade, "Putting Rough Sets and Fuzzy Sets Together," *Intelligent Decision Support*, Kluwer Academic Publishers, pp. 203-232, 1992.

[6] *Rough-Fuzzy Hybridization: A New Trend in Decision Making*, S.K. Pal and A. Skowron, eds. Springer Verlag, 1999.

[7] P. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach.* Prentice Hall, 1982.

[8] A. Hassanien, "Rough Set Approach for Attribute Reduction and Rule Generation: A Case of Patients with Suspected Breast Cancer," *J. Am. Soc. Information Science and Technology*, vol. 55, no. 11, pp. 954-962, 2004.

[9] A. Hedar, J. Wang, and M. Fukushima, "Tabu Search for Attribute Reduction in Rough Set Theory," Technical Report 2006-008, Dept. of Applied Mathematics and Physics, Kyoto Univ., 2006.

[10] M. Inuiguchi and T. Tanino, *New Fuzzy-Rough Sets Based on Certainty Qualification, Rough-Neural Computing: Techniques for Computing with Words*, S.K. Pal, L. Polkowski, and A. Skowron, eds. Springer-Verlag, 2003.

[11] M. Inuiguchi and M. Tsurumi, "Measures Based on Upper Approximations of Rough Sets for Analysis of Attribute Importance and Interaction," *Int'l J. Innovative Computing, Information and Control*, vol. 2, no. 1, pp. 1-12, 2006.

[12] R. Jensen and Q. Shen, "Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approaches," *IEEE Trans. Knowledge and Data Eng.*, vol. 16, no. 12, pp. 1457-1471, Dec. 2004.

[13] H.R. Li and W.X. Zhang, "Applying Indiscernibility Attribute Sets to Knowledge Reduction," *Lecture Notes in Artificial Intelligence,* pp. 816-821, Springer, 2005.

[14] K. Li, Y. Liu, "Rough Set Based Attribute Reduction Approach in Data Mining," *Proc. 2002 Int'l Conf. Machine Learning and Cybernetics,* vol. 1, pp. 60-63, 2002.

[15] N. Mac Parthaláin, R. Jensen, and Q. Shen, "Fuzzy Entropy-Assisted Fuzzy-Rough Feature Selection," *Proc. 15th Int'l Conf. Fuzzy Systems (FUZZ-IEEE '06)* 2006.

[16] N. Mac Parthaláin, R. Jensen, and Q. Shen, "Distance Measure Assisted Rough Set Feature Selection," *Proc. 16th Int'l Conf. Fuzzy Systems (FUZZ-IEEE '07),* pp. 1084-1089, 2007.

[17] N. Mac Parthaláin and Q. Shen, "Exploring the Boundary Region of Tolerance Rough Sets for Feature Selection," *Pattern Recognition,* vol. 42, pp. 655-667, http://www.sciencedirect.com/science/article/B6V14-4TDC09M-1/2/9735ab90392246f032a2632eda77ae0e, May 2009.

[18] M. Modrzejewski, "Feature Selection Using Rough Sets Theory," *Proc. European Conf. Machine Learning,* P.B. Brazdil, ed., pp. 213-226, 1993.

[19] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "UCI Repository of Machine Learning Databases," Dept. of Information and Computer Science, Univ. of California, http://www.ics.uci.edu/mlearn/MLRepository.html, 1998.

[20] R. Nie and J. Yue, "An Attribute Reduction Method Based on Rough Set and SVM and with Application in Oil-Gas Prediction," *Proc. Sixth IEEE/ACIS Int'l Conf. Computer and Information Science (ICIS '07),* pp. 502-506, 2007.

[21] S.H. Nguyen and A. Skowron, "Searching for Relational Patterns in Data," *Proc. First European Symp. Principles of Data Mining and Knowledge Discovery,* pp. 265-276, 1997.

[22] S. Piramuthu, "The Hausdorff Distance Measure for Feature Selection in Learning Applications," *Proc. 32nd Ann. Hawaii Int'l Conf. System Sciences,* vol. 6, 1999.

[23] S.K. Pal and P. Mitra, "Case Generation Using Rough Sets with Fuzzy Representation," *IEEE Trans. Knowledge and Data Eng.,* vol. 16, no. 3, pp. 292-300, Mar. 2004.

[24] Z. Pawlak, "Rough Sets," *Int'l J. Computer and Information Science,* vol. 11, pp. 341-356, 1982.

[25] J.R. Quinlan, *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, 1993.

[26] W. Rucklidge, *Efficient Visual Recognition Using the Hausdorff Distance.* Springer, 1996.

[27] B. Sendov, "Hausdorff Distance and Image Processing," *Russian Math Surveys,* vol. 59, no. 2, pp. 319-328, 2004.

[28] Q. Shen and R. Jensen, "Selecting Informative Features with Fuzzy-Rough Sets and Its Application for Complex Systems Monitoring," *Pattern Recognition,* vol. 37, no. 7, pp. 1351-1363, 2004.

[29] A. Skowron and J. Stepaniuk, "Tolerance Approximation Spaces," *Fundamenta Informaticae,* vol. 27, pp. 245-253, 1996.

[30] D. Slezak, "Various Approaches to Reasoning with Frequency Based Decision Reducts: A Survey," *Rough Set Methods and Applications,* L. Polkowski, S. Tsumoto, T.Y. Lin, eds., pp 235-285, Physica-Verlag, 2000.

[31] *Intelligent Decision Support,* R. Slowinski, ed. Kluwer Academic Publishers, 1992.

[32] E.P.M. de Sousa, C. Traina, A.J.M. Traina, L. Wu, and C. Faloutsos, "A Fast and Effective Method to Find Correlations among Attributes in Databases," *Data Mining and Knowledge Discovery,* vol. 14, pp. 367-407, 2007.

[33] R.W. Swiniarski and A. Skowron, "Rough Set Methods in Feature Selection and Recognition," *Pattern Recognition Letters,* vol. 24, no. 6, pp. 833-849, 2003.

[34] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools with Java Implementations.* Morgan Kaufmann, 2000.

[35] I.H. Witten and E. Frank, "Generating Accurate Rule Sets without Global Optimization," *Proc. 15th Int'l Conf. Machine Learning,* 1998.

[36] Y. Yao, "A Comparative Study of Fuzzy Sets and Rough Sets," *Information Sciences,* vol. 109, pp. 21-47, 1998.

[37] N. Zhong, J. Dong, and S. Ohsuga, "Using Rough Sets with Heuristics for Feature Selection," *J. Intelligent Information Systems,* vol. 16, no. 3, pp. 199-214, 2001.

[38] W. Ziarko, "Variable Precision Rough Set Model," *J. Computer and System Sciences,* vol. 46, no. 1, pp. 39-59, 1993.

Neil Mac Parthaláin is a research associate with the Vision Graphics and Visualisation Group in the Department of Computer Science, Aberystwyth University, Wales, UK. His research interests include rough set theory, fuzzy set theory, feature selection, and rule induction. He has published around 10 peer-refereed conference papers and academic journal articles in these areas.

Qiang Shen is a professor with the Department of Computer Science at Aberystwyth University, Wales, UK, and an honorary fellow at the University of Edinburgh, UK. His research interests include fuzzy and imprecise modeling, model-based inference, pattern recognition, and knowledge refinement and reuse. He is an associate editor of the *IEEE Transactions on Fuzzy Systems* and of the *IEEE Transactions on Systems, Man, and Cybernetics (Part B)*, and an editorial board member of the *Fuzzy Sets and Systems Journal* among others. He has published more than 240 peer-refereed papers in academic journals and conferences on topics within artificial intelligence and related areas, including one winning the IEEE Outstanding Paper Award in the *Transactions on Fuzzy Systems*.

Richard Jensen is a lecturer in the Department of Computer Science at Aberystwyth University, Wales, UK. His research interests include rough and fuzzy set theory, pattern recognition, information retrieval, feature selection, and swarm intelligence. He has published more than 40 peer-refereed conference papers and academic journal articles in these areas.