Tolerance-based and Fuzzy-Rough Feature Selection

Jensen, Richard; Shen, Qiang

Publication date:
2007

Citation for published version (APA):
Jensen, R., & Shen, Q. (2007). *Tolerance-based and Fuzzy-Rough Feature Selection*. 877-882.

Tolerance-based and Fuzzy-Rough Feature Selection

Richard Jensen and Qiang Shen

***Abstract*— One of the main obstacles facing the application of computational intelligence technologies in pattern recognition (and indeed in many other tasks) is that of dataset dimensionality. To enable pattern classifiers to be effective, a dimensionality minimization step is usually carried out beforehand. Rough set theory has been successfully applied for this as it requires only the supplied data and no other information; most other methods require supplementary knowledge. However, the main limitation of traditional rough set-based selection in the literature is the restrictive requirement that all data is discrete; it is not possible to consider real-valued or noisy data. This has been tackled previously via the use of discretization methods, but may result in information loss. This paper investigates two approaches based on rough set extensions, namely fuzzy-rough and tolerance rough sets, that address these problems and retain dataset semantics. The methods are compared experimentally and utilized for the task of forensic glass fragment identification.**

I. Introduction

Feature selection [5] addresses the problem of selecting those input features that are most predictive of a given outcome; a problem encountered in many areas of computational intelligence. Unlike other dimensionality reduction methods, feature selectors preserve the original meaning of the features after reduction. This has found application in tasks that involve datasets containing huge numbers of features (in the order of tens of thousands) which, for some learning algorithms, might be impossible to process further. Recent examples include text processing and web content classification [8].

There are often many features involved, and combinatorially large numbers of feature combinations, to select from. Note that the number of feature subset combinations with m features from a collection of N total features is $N!/[m!(N-m)!]$. It might be expected that the inclusion of an increasing number of features would increase the likelihood of including enough information to distinguish between classes. Unfortunately, this is not necessarily true if the size of the training dataset does not also increase rapidly with each additional feature included. A high-dimensional dataset increases the chances that a learning algorithm will find spurious patterns that are not valid in general. Most techniques employ some degree of reduction in order to cope with large amounts of data, so an efficient and effective reduction method is required.

A technique that can reduce dimensionality using information contained within the dataset and that preserves the meaning of the features (i.e. semantics-preserving) is clearly desirable. Rough set theory (RST) [10] can be used as such a tool to discover data dependencies and to reduce the number of features contained in a dataset using the data alone. However, traditional RST methods are generally incapable of handling real-valued data directly. Previously, discretization methods were applied beforehand in order to transform the data into discrete values, but this may result in information loss. As a result of this, several extensions to the original theory have been proposed. Two significant developments in this area have been fuzzy-rough sets [6] and similarity- or tolerance-based rough set theory [15]. It is, therefore, desirable to develop techniques to provide the means of data reduction for crisp and real-value attributed datasets which utilize this additional information.

This paper presents two methods for feature selection that employ rough set extensions for this purpose. By using tolerance relations or fuzzy equivalence classes, the strict requirement of complete equivalence can be relaxed, and a more flexible approach to subset selection can be developed. The rest of the paper is structured as follows. Section 2 introduces the main theoretical concepts behind crisp and fuzzy-rough set-based feature selection. Section 3 presents the tolerance rough set-based selection method with a worked example. The experimentation carried out is detailed in the fourth section, and the approaches compared with respect to reduct size, time taken and resulting classification accuracy. Their utility is also evaluated with respect to a challenging task - glass fragment identification. The paper is concluded in section 5.

II. Approaches

Rough set theory [10] is an extension of conventional set theory that supports approximations in decision making. The rough set itself is the approximation of a vague concept (set) by a pair of precise concepts, called lower and upper approximations, which are a classification of the domain of interest into disjoint categories. The lower approximation is a description of the domain objects which are known with certainty to belong to the subset of interest, whereas the upper approximation is a description of the objects which possibly belong to the subset.

A. Rough Set Attribute Reduction

Central to Rough Set Attribute Reduction (RSAR) [3], [7] is the concept of indiscernibility. Let $I = (\mathbb{U}, \mathbb{A})$ be an information system, where $\mathbb{U}$ is a non-empty set of finite objects (the universe) and $\mathbb{A}$ is a non-empty 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. With any $P \subseteq \mathbb{A}$ there is an associated equivalence relation $IND(P)$:

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

Richard Jensen and Qiang Shen are with the Department of Computer Science, The University of Wales, Aberystwyth, UK. (phone: +44 1970 621787; email: {rkj,qqs}@aber.ac.uk).

The partition of $\mathbb{U}$, generated by *IND(P)* is denoted $\mathbb{U}$*/IND(P)* (or $\mathbb{U}$*/P*) and can be calculated as follows:

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

where

$$A \otimes B = \{X \cap Y \mid \forall X \in A, \forall Y \in B, X \cap Y \neq \emptyset\} \quad (3)$$

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 within P by constructing the P-*lower* and P-*upper* approximations of X:

$$\underline{P}X = \{x \mid [x]_P \subseteq X\} \quad (4)$$

$$\overline{P}X = \{x \mid [x]_P \cap X \neq \emptyset\} \quad (5)$$

Let *P* and *Q* be equivalence relations over $\mathbb{U}$, then the positive region is defined as:

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

The positive region contains all objects of $\mathbb{U}$ that can be classified to classes of $\mathbb{U}$/*Q* using the information in attributes *P*. Using this definition of the positive region, the rough set degree of dependency of a set of attributes *Q* on a set of attributes *P* is defined in the following way:

For *P*, *Q* $\subseteq \mathbb{A}$, it is said that *Q* depends on *P* in a degree *k* ($0 \leq k \leq 1$), denoted $P \Rightarrow_k Q$, if

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|\mathbb{U}|} \quad (7)$$

The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same predictive capability of the set of decision features as the original. A *reduct*, *R*, is defined as a subset of minimal cardinality of the conditional attribute set C such that $\gamma_R(D) = \gamma_C(D)$. The QUICKREDUCT algorithm given in [3], attempts to calculate a reduct without exhaustively generating all possible subsets. It starts off with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough set dependency metric, until this produces its maximum possible value for the dataset.

B. Fuzzy-Rough Feature Selection

The reliance on discrete data for the successful operation of rough set-based feature selection methods such as [3] and expanded in [11], [12] can be seen as a significant drawback of the approach. Indeed, this requirement implies an objectivity in the data that is simply not present in many real problem domains. For example, in a medical dataset, the feature *Blood Pressure* is a real-valued measurement but for the purposes of rough set theory must be discretized into a small set of labels such as *Normal*, *High*, etc. Subjective judgments are required for establishing boundaries for objective measurements.

A better way of handling this problem is the use of fuzzy-rough sets, as envisaged in [7]. Subjective judgments are not entirely removed as fuzzy set membership functions still need to be defined. However, the method offers a high degree of flexibility when dealing with real-valued data, enabling the vagueness and imprecision present to be modelled effectively. By employing fuzzy-rough sets, it is possible to use this information to better guide feature selection.

1) Fuzzy Equivalence Classes: In the same way that crisp equivalence classes are central to rough sets, *fuzzy* equivalence classes are central to the fuzzy-rough set approach [6]. For pattern classification applications, this means that the class values and the feature values may all be fuzzy. In particular, the family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes.

2) Fuzzy Lower and Upper Approximations: The fuzzy lower and upper approximations are fuzzy extensions of their crisp counterparts, and can be redefined as:

$$\mu_{\underline{P}X}(x) = \sup_{F \in \mathbb{U}/P} \min(\mu_F(x), \inf_{y \in \mathbb{U}} \max\{1 - \mu_F(y), \mu_X(y)\}) \quad (8)$$

$$\mu_{\overline{P}X}(x) = \sup_{F \in \mathbb{U}/P} \min(\mu_F(x), \sup_{y \in \mathbb{U}} \min\{\mu_F(y), \mu_X(y)\}) \quad (9)$$

The tuple $< \underline{P}X, \overline{P}X >$ 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 subsets of features, the following is used:

$$\mathbb{U}/P = \otimes\{a \in P \mid \mathbb{U}/IND(\{a\})\} \quad (10)$$

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, $i = 1, 2, ..., n$:

$$\mu_{F_1 \cap ... \cap F_n}(x) = \min(\mu_{F_1}(x), \mu_{F_2}(x), ..., \mu_{F_n}(x)) \quad (11)$$

3) Fuzzy-Rough Reduction Process: Fuzzy-Rough Feature Selection (FRFS) builds on the notion of the fuzzy lower approximation to enable reduction of datasets containing real-valued features. The process becomes identical to the crisp approach when dealing with nominal well-defined features.

The crisp positive region in the standard RST is defined as the union of the lower approximations. By the extension principle, the membership of an object $x \in \mathbb{U}$, belonging to the fuzzy positive region can be defined by

$$\mu_{POS_P(Q)}(x) = \sup_{X \in \mathbb{U}/Q} \mu_{\underline{P}X}(x) \quad (12)$$

Using the definition of the fuzzy positive region, a new dependency function between a set of features Q and another set P can be defined as follows:

$$\gamma'_P(Q) = \frac{|\mu_{POS_P(Q)}(x)|}{|\mathbb{U}|} = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_P(Q)}(x)}{|\mathbb{U}|} \quad (13)$$

As with crisp rough sets, the dependency of Q on P is the proportion of objects that are discernible out of the entire dataset. In the present approach, this corresponds to determining the fuzzy cardinality of $\mu_{POS_P(Q)}(x)$ divided by the total number of objects in the universe.

FRQUICKREDUCT(C,D).
C, the set of all conditional features;
D, the set of decision features.

(1) $R \leftarrow \{\}$, $\gamma'_{best} \leftarrow 0$, $\gamma'_{prev} \leftarrow 0$
(2) **do**
(3) $T \leftarrow R$
(4) $\gamma'_{prev} \leftarrow \gamma'_{best}$
(5) $\forall x \in (C - R)$
(6) **if** $\gamma'_{R\cup\{x\}}(D) > \gamma'_T(D)$
(7) $T \leftarrow R \cup \{x\}$
(8) $\gamma'_{best} \leftarrow \gamma'_T(D)$
(9) $R \leftarrow T$
(10) **until** $\gamma'_{best} = \gamma'_{prev}$
(11) **return** R

Fig. 1. Fuzzy-rough QUICKREDUCT

Computationally, the fuzzy-rough QUICKREDUCT algorithm of Figure 1 employs the dependency function γ' to choose which features to add to the current reduct candidate. The algorithm terminates when the addition of any remaining feature does not increase the dependency. The complexity of the approach is O($(n^2+n)/2$).

III. TOLERANCE-BASED FEATURE SELECTION

Another way of attempting to handle the problem of real-valued data is to introduce a measure of similarity of feature values and define the lower and upper approximations based on these similarity measures.

A. Similarity Measures

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 [16], is:

$$SIM_a(x,y) = 1 - \frac{|a(x) - a(y)|}{|a_{\max} - a_{\min}|} \tag{14}$$

where a is the feature under consideration, and $a_{\max}$ and $a_{\min}$ denote the maximum and minimum values respectively for this feature. When considering more than one feature, the defined similarities must be combined to provide a measure of the overall similarity of objects. For a subset of features, P, this can be achieved in many ways; two commonly adopted approaches are:

$$(x,y) \in SIM_{P,\tau} \text{ iff } \prod_{a\in P} SIM_a(x,y) \geq \tau \tag{15}$$

$$(x,y) \in SIM_{P,\tau} \text{ iff } \frac{\sum_{a\in P} SIM_a(x,y)}{|P|} \geq \tau \tag{16}$$

where τ is a global similarity threshold; it 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 equation (15)) and threshold ($\tau = 1$). Further similarity relations are investigated in [9], but are omitted here.

From this, the so-called tolerance classes that are generated by a given similarity relation for an object x are defined as:

$$SIM_{P,\tau}(x) = \{y \in \mathbb{U} | (x,y) \in SIM_{P,\tau}\} \tag{17}$$

B. Approximations and Dependency

Lower and upper approximations are then defined in a similar way to traditional rough set theory:

$$\underline{P_\tau} X = \{x | SIM_{P,\tau}(x) \subseteq X\} \tag{18}$$

$$\overline{P_\tau} X = \{x | SIM_{P,\tau}(x) \cap X \neq \emptyset\} \tag{19}$$

The tuple $< \underline{P_\tau} X, \overline{P_\tau} X >$ is called a tolerance rough set [15]. Positive region and dependency functions then become:

$$POS_{P,\tau}(Q) = \bigcup_{X\in\mathbb{U}/Q} \underline{P_\tau} X \tag{20}$$

$$\gamma_{P,\tau}(Q) = \frac{|POS_{P,\tau}(Q)|}{|\mathbb{U}|} \tag{21}$$

From these definitions, feature reduction methods can be constructed that use the tolerance-based degree of dependency, $\gamma_{P,\tau}(Q)$, to gauge the significance of feature subsets (in a similar way as fuzzy-rough QUICKREDUCT). The resulting algorithm can be found in Figure 2.

TOLERANCEREDUCT(C,D,τ).
C, the set of all conditional features;
D, the set of decision features;
τ, the similarity threshold.

(1) $R \leftarrow \{\}$; $\gamma^\tau_{best} = 0$;
(2) **do**
(3) $T \leftarrow R$
(4) $\gamma^\tau_{prev} = \gamma^\tau_{best}$
(5) $\forall x \in (C - R)$
(6) **if** $\gamma_{R\cup\{x\},\tau}(D) > \gamma_{T,\tau}(D)$
(7) $T \leftarrow R \cup \{x\}$
(8) $\gamma^\tau_{best} = \gamma_{T,\tau}(D)$
(9) $R \leftarrow T$
(10) **until** $\gamma^\tau_{best} == \gamma^\tau_{prev}$
(11) **return** R

Fig. 2. Tolerance QUICKREDUCT

Note that for a dimensionality of n, $(n^2+n)/2$ evaluations of the tolerance-based dependency function may be performed for the worst-case dataset. However, as with fuzzy-rough QUICKREDUCT, the algorithm is used for dimensionality reduction prior to any involvement of the system which will employ those features belonging to the resultant reduct. Thus, this operation has no negative impact upon the runtime efficiency of the system.

TABLE I

EXAMPLE DATASET

Object	a	b	c	q
1	-0.4	-0.3	-0.5	no
2	-0.4	0.2	-0.1	yes
3	-0.3	-0.4	-0.3	no
4	0.3	-0.3	0	yes
5	0.2	-0.3	0	yes
6	0.2	0	0	no

C. Worked Example

To illustrate the operation of the tolerance QUICKREDUCT algorithm, it is applied to the example data given in table I, which contains three real-valued conditional attributes and a crisp-valued decision attribute. For this example, the similarity measure is the same as that given in equation (14) and equation (15) for all conditional attributes, with $\tau = 0.7$. This choice of threshold permits attribute values to differ to a limited extent, allowing close values to be considered as identical. For the decision feature, τ is set to 1 (i.e. objects must have identical values to appear in the same tolerance class) as the decision value is nominal. Setting $A = \{a\}$, $B = \{b\}$, $C = \{c\}$ and $Q = \{q\}$, the following tolerance classes are obtained:

$$\mathbb{U}/SIM_{A,\tau} = \{\{1,2,3\},\{4,5,6\}\}$$
$$\mathbb{U}/SIM_{B,\tau} = \{\{1,3,4,5\},\{2\},\{6\}\}$$
$$\mathbb{U}/SIM_{C,\tau} = \{\{1\},\{2,4,5,6\},\{3\}\}$$
$$\mathbb{U}/SIM_{Q,\tau} = \{\{1,3,6\},\{2,4,5\}\}$$
$$\mathbb{U}/SIM_{\{a,b\},\tau} = \{\{1,3\},\{2\},\{4,5\},\{4,5,6\},\{5,6\}\}$$
$$\mathbb{U}/SIM_{\{b,c\},\tau} = \{\{1,3\},\{2,6\},\{4,5,6\},\{2,4,5,6\}\}$$
$$\mathbb{U}/SIM_{\{a,b,c\},\tau} = \{\{1,3\},\{2\},\{4,5,6\}\}$$

It can be seen that some objects belong to more than one tolerance class. This is due to the additional flexibility of employing similarity measures rather than strict equivalence.

Based on these partitions, the degree of dependency can be calculated for attribute subsets, providing an evaluation of their significance. The tolerance QUICKREDUCT algorithm considers the addition of attributes to the currently stored best subset (initially the empty set) and selects the feature that results in the highest increase of the dependency degree. Considering attribute b, the lower approximations of the decision classes are calculated as follows:

$$\underline{B_\tau}\{1,3,6\} = \{x | SIM_{B,\tau}(x) \subseteq \{1,3,6\}\} = \{6\}$$
$$\underline{B_\tau}\{2,4,5\} = \{x | SIM_{B,\tau}(x) \subseteq \{1,3,6\}\} = \{2\}$$

Hence, the positive region can be constructed:

$$\begin{aligned} POS_{B,\tau}(Q) &= \bigcup_{X \in U/Q} \underline{B_\tau} X \\ &= \underline{B_\tau}\{1,3,6\} \cup \underline{B_\tau}\{2,4,5\} \\ &= \{2,6\} \end{aligned}$$

The resulting degree of dependency is:

$$\gamma_{B,\tau}(Q) = \frac{|POS_{B,\tau}(Q)|}{|U|} = \frac{|\{2,6\}|}{|\{1,2,3,4,5,6\}|} = \frac{2}{6}$$

For the other conditional features in the dataset, the corresponding dependency degrees are:

$$\gamma_{A,\tau}(Q) = \frac{|\{\emptyset\}|}{|\{1,2,3,4,5,6\}|} = \frac{0}{6}$$
$$\gamma_{C,\tau}(Q) = \frac{|\{1,3\}|}{|\{1,2,3,4,5,6\}|} = \frac{2}{6}$$

Following the tolerance QUICKREDUCT algorithm, attribute b is added to the reduct candidate ($R = \{b\}$) and the search continues. The algorithm makes an arbitrary choice here between attributes b and c as they produce equally high degrees of dependency (although they generate different positive regions). As attribute b was considered before attribute c, it is selected. The algorithm continues by evaluating subsets containing this attribute in combination with the remaining individual attributes from the dataset.

$$\gamma_{\{a,b\},\tau}(Q) = \frac{|\{1,2,3,4,5\}|}{|\{1,2,3,4,5,6\}|} = \frac{5}{6}$$
$$\gamma_{\{b,c\},\tau}(Q) = \frac{|\{1,3\}|}{|\{1,2,3,4,5,6\}|} = \frac{2}{6}$$

The subset $\{a,b\}$ is chosen as this results in a higher dependency degree than $\{b\}$. The algorithm then evaluates the combination of this subset with the remaining attributes (in this example only one attribute, c, remains):

$$\gamma_{\{a,b,c\},\tau}(Q) = \frac{|\{1,2,3\}|}{|\{1,2,3,4,5,6\}|} = \frac{3}{6}$$

As this value is less than that for subset $\{a,b\}$, the algorithm terminates and outputs the reduct $\{a,b\}$. Note that this is the same subset as that found by the fuzzy-rough method [7] for this dataset, but applying RSAR leads to a non-minimal reduct $\{a,b,c\}$.

IV. EXPERIMENTATION

This section presents the experimental evaluation of the selection methods for the task of pattern classification, over nine benchmark datasets and one real-world problem with several different classifiers.

A. Experimental Setup

FRFS uses a pre-categorisation step which generates associated fuzzy sets for a dataset. For the tolerance-based method, the threshold is selected by initially setting τ to 1 and applying tolerance QUICKREDUCT to the data. τ is then decremented by 0.01 and this process is repeated for a set number of iterations. From this, the threshold value resulting in the largest reduct is chosen. After feature selection, the datasets are reduced according to the discovered reducts. These reduced datasets are then classified using the relevant classifier. (Note that the feature selection step is not employed for the unreduced dataset.)

Four classifiers were employed for the purpose of evaluating the resulting subsets from the feature selection phase: J48, JRip, PART (from [18]) and MODLEM [13]. J48 [14] creates decision trees by choosing the most informative features and recursively partitioning the data into subtables

TABLE II
REDUCT SIZE AND TIME TAKEN FOR FRFS AND TOLERANCE METHODS.

Dataset	Objects	Features	Reduct size		Time taken (s)	
			FRFS	Tol.	FRFS	Tol.
Cleveland	297	14	11	11	24.11	3.37
Glass	214	10	9	7	1.61	0.61
Heart	270	14	11	10	11.84	2.79
Ionosphere	230	35	11	10	61.80	10.04
Iris	150	5	5	4	0.031	0.206
Olitos	120	26	10	10	11.20	2.60
Water 2	390	39	11	8	96.58	35.02
Water 3	390	39	12	11	158.73	57.81
Wine	178	14	10	8	1.42	0.63

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 [4] learns propositional rules by repeatedly growing rules and pruning them. During the growth phase, features are added greedily until a termination condition is satisfied. Features are then pruned in the next phase subject to a pruning metric. Once the ruleset is generated, a further optimization is performed where classification rules are evaluated and deleted based on their performance on randomized data. PART [17] 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 ruleset during processing. Essentially, a classification 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. Additionally, experimentation is carried out for a leading rough set-based rule induction method, MODLEM, using extended minimal covering [13].

B. Benchmark Data

This section presents the results of experimental studies using nine datasets as given in Table II. These datasets are small-to-medium in size, with between 120 and 390 objects per dataset and feature sets ranging from 5 to 39. All datasets have been obtained from [1], [2] and [3].

Table II compares the reduct size and runtime data, using both FRFS and tolerance-based approaches. It can be seen that the tolerance-based method consistently locates reducts that are smaller or equal in size to those found by FRFS. In fact, in all but two of the test datasets (Cleveland and Olitos) the reducts are smaller. It is clear from the results that the runtimes of the tolerance-based technique are also considerably shorter in general than those of FRFS.

Table III shows the average classification accuracy as a percentage obtained using 10-fold cross validation. The classification was initially performed on the unreduced dataset, followed by the reduced datasets which were obtained by using both FRFS and tolerance-based feature selection techniques. In most cases the classification accuracy increases or remains at approximately the same level for both FRFS and tolerance-based methods. There are some notable exceptions however, where a decrease in classification accuracy is observed. When such decreases are compared to the reduction in dimensionality, it is apparent that they are not significant.

For the J48 classifier, FRFS maintains or improves upon the performance of the unreduced dataset in all but two cases. The tolerance approach does not perform as well as FRFS for this classifier, but still retains reasonably high classification accuracy with fewer number of features. When JRip is employed, the tolerance-based method performs better than FRFS. In comparison to the unreduced dataset results, accuracy is improved or maintained in six cases, whereas, for FRFS, this only occurs in three. For PART, the methods perform similarly. FRFS produces higher accuracies for six datasets, maintains accuracy in two, and results in worse accuracy in two. The tolerance method improves accuracy in seven datasets, and displays poorer performance in three. With MODLEM, both feature selection methods result in similar accuracies, and perform slightly worse in general than the unreduced method.

Overall, the results show that both feature selection methods greatly reduce dimensionality while generally maintaining or improving performance. Although there are some instances where the classification accuracy may decrease, it is small in comparison to the overall reduction of dimensionality. The tolerance-based approach often chooses smaller subsets than FRFS, whilst exhibiting a comparable level of classification performance.

C. Application to Forensics

One of the less obvious, but frequent, sources of forensic evidence are traces of glass. This is regularly encountered at crime scenes, particularly those involving motor vehicle accidents, car theft and burglaries. The forensic scientist's role in analysing glass is to clearly and unambiguously determine the origin of the sample. Variation in the manufacture of glass allows considerable discrimination even with very small fragments. To demonstrate the domain-independence and utility of the work in this paper, the feature selection methods were applied as pre-processors to this challenging task of glass fragment classification.

The data was obtained from the Institute of Forensic Research, Krakow, Poland. 800 glass fragments were extracted from six glass sources. From these, the chemical concentrations of 8 elements (oxygen, sodium, potassium, aluminium, iron, magnesium, calcium and silicon) were measured via a scanning electron microscope with an energy dispersive X-ray spectrometer. These constitute the features of the dataset. The six glass sources are bulb glass, car window glass, headlamp glass, optic glass, glass containers and building window glass. These constitute the classes of the dataset. The task is to derive classifiers that correctly identify the source of glass based solely on chemical concentrations.

The results from the experimentation can be seen in Table IV. The classifiers used previously were employed for the purpose of predicting glass source. Again, these were evaluated using 10-fold cross-validation.

TABLE III

RESULTING CLASSIFICATION ACCURACIES (%) FOR UNREDUCED, FRFS AND TOLERANCE METHODS.

Dataset	J48			JRip			PART			MODLEM		
	Unred.	FRFS	Tol.	Unred.	FRFS	Tol.	Unred.	FRFS	Tol.	Unred.	FRFS	Tol.
Cleveland	51.85	55.22	51.17	53.87	53.87	53.87	50.17	52.19	57.23	55.80	51.45	53.56
Glass	67.29	69.63	69.16	69.16	67.76	67.76	67.76	68.22	69.62	58.42	57.53	52.01
Heart	76.67	78.89	80.37	79.63	81.85	82.59	73.33	78.52	80.37	77.41	77.41	72.59
Ionosphere	87.83	91.30	87.39	86.96	86.52	86.96	88.26	91.30	86.52	86.52	86.52	86.52
Iris	96.00	96.00	96.00	95.33	95.33	94.67	94.00	94.00	95.33	91.33	91.33	93.33
Olitos	67.50	67.50	61.67	70.00	66.67	61.67	57.50	62.50	72.50	71.67	63.33	64.17
Water 2	83.33	80.26	81.79	81.03	80.51	82.31	85.64	82.56	81.28	81.79	83.33	81.54
Water 3	77.44	79.74	75.64	83.85	80.76	78.46	79.49	78.97	72.05	80.00	81.03	75.90
Wine	94.38	92.14	94.94	91.57	90.45	94.38	93.82	93.82	94.38	94.93	92.65	93.82

TABLE IV

RESULTS FOR FORENSIC DATA.

Method	Features	J48	JRip	PART	MODLEM
Unreduced	8	83.13	79.38	80.50	78.00
FRFS	8	83.13	79.38	80.50	78.00
Tolerance	6	82.00	78.25	78.00	74.13

Applying FRFS to the data resulted in all features being chosen for the classification task, and hence the results are equal to the unreduced approach. The tolerance method selected a smaller number of features (6) for the task. It can be seen that the FRFS method was correct in retaining the entire feature set as the reduction in dimensionality produced by the tolerance method resulted in a slight decrease in performance for all classifiers. However, it should be noted that there is a cost (both monetary and time) associated with measuring each chemical within fragments. The results show that (via tolerance rough sets) two of these can be eliminated with only a small drop in classification accuracy.

V. CONCLUSION

This paper has presented two methods for handling real-valued data in feature selection, by the use of tolerance rough sets and fuzzy-rough sets. Through the benchmark data experimentation, it was found that the tolerance approach discovered smaller subsets whilst maintaining a comparably high level of performance. The methods were also applied to the real-world problem of glass identification - an important area in forensic glass analysis.

The experimental work presented here did not take advantage of any optimization for the fuzzifications, similarities or classifiers involved. It is expected that the results obtained through the use of such optimization would reflect a marked improvement. Future work would include the implementation of optimization methods for both algorithms - improving fuzzy set definitions for FRFS, and fine-tuning feature similarity measures for the tolerance-based approach.

ACKNOWLEDGMENTS

This work is partly funded by the UK EPSRC grant GR/S98603/01. The authors are very grateful to Professor Colin Aitken and Mr Burkhard Schafer of the University of Edinburgh for their support, and to Dr Grzegorz Zadora of the Forensic Research Institute, Krakow, Poland for the provision of the glass data.

REFERENCES

[1] C. Armanino, R. Leardi, S. Lanteri and G. Modi, "Chemometric analysis of Tuscan olive oils," *Chemom. Intell. Lab. Syst.*, vol. 5, pp. 343–354. 1989.

[2] C.L. Blake and C.J. Merz. UCI Repository of machine learning databases. Irvine, University of California, 1998. http://www.ics.uci.edu/~mlearn/.

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

[4] W.W. Cohen, "Fast effective rule induction," In *Machine Learning: Proceedings of the 12th International Conference*, pp. 115–123. 1995.

[5] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, vol. 1, no. 3, pp. 131-156, 1997.

[6] D. Dubois and H. Prade, "Putting Rough Sets and Fuzzy Sets Together," *Intelligent Decision Support*, pp. 203–232, 1992.

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

[8] R. Jensen and Q. Shen, "Fuzzy-Rough Sets Assisted Attribute Selection," Accepted for publication in *IEEE Transactions on Fuzzy Systems*.

[9] S.H. Nguyen and A. Skowron, "Searching for Relational Patterns in Data," In *Proc. First European Symposium on Principles of Data Mining and Knowledge Discovery*, pp. 265–276, 1997.

[10] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishing, 1991.

[11] Z. Pawlak and A. Skowron, "Rudiments of rough sets," *Information Sciences*, vol. 177, pp. 3-27, 2006.

[12] Z. Pawlak and A. Skowron, "Rough sets: Some extensions," *Information Sciences*, vol. 177, pp. 28-40, 2006.

[13] B. Predki and S. Wilk, "Rough Set Based Data Exploration Using ROSE System," In: Z.W. Ras, A. Skowron, eds. *Foundations of Intelligent Systems*, LNAI, vol. 1609, pp. 172–180. 1999.

[14] J.R. Quinlan, *C4.5: Programs for Machine Learning*, The Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[15] A. Skowron and J. Stepaniuk, "Tolerance Approximation Spaces," *Fundamenta Informaticae*, vol. 27, no. 2, pp. 245–253, 1996.

[16] J. Stepaniuk, "Optimizations of rough set model," *Fundamenta Informaticae*, vol. 36, no. 2-3, pp. 265–283, 1998.

[17] I.H. Witten and E. Frank, "Generating Accurate Rule Sets Without Global Optimization," In *Machine Learning: Proceedings of the 15th International Conference*, Morgan Kaufmann Publishers, San Francisco, 1998.

[18] I.H. Witten and E. Frank, *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann Publishers, San Francisco, 2000.