Aiding Fuzzy Rule Induction with Fuzzy Rough Attribute Reduction

Richard Jensen and Qiang Shen
{richjens,qiangs}@dai.ed.ac.uk

Centre for Intelligent Systems and their Applications
Division of Informatics
The University of Edinburgh

Abstract

Many rule induction algorithms are unable to cope with high dimensional descriptions of input features. To enable such techniques to be effective, a redundancy-removing step is usually carried out beforehand. Rough Set Theory (RST) has been used as such a dataset pre-processor with much success, however it is reliant upon a *crisp* dataset; important information may be lost as a result of quantization. By using fuzzy-rough sets this loss is avoided, allowing the reduction of noisy, real-valued attributes. This paper demonstrates the applicability of fuzzy-rough attribute reduction to the problem of learning classifiers, resulting in simpler rules with little loss in classification accuracy.

1 Introduction

The knowledge acquisition bottleneck is a problem in the building of intelligent reasoning systems for complex applications, especially when experts are not readily available. High dimensionality of the domain attributes presents a further obstacle for a number of rule induction algorithms which would otherwise have the potential for automating knowledge acquisition. It is therefore not surprising that much research has been carried out on dimensionality reduction. However, existing work tends to destroy the underlying semantics of the features after reduction (e.g. transformation-based approaches [12]) or require additional information about the given data set for thresholding (e.g. entropy-based approaches [10]). A technique that can reduce dimensionality using information contained within the data set and preserving the meaning of the features is clearly desirable.

Fuzzy-rough sets [3] can be used as a tool to discover data dependencies and reduce the number of attributes contained in a dataset by purely structural methods. The Fuzzy-Rough Attribute Reduction (FRAR) technique [7] can be applied to datasets where conditional and decision attribute values are crisp or fuzzy. Of particular interest here is the case where both the conditional and decision values are fuzzy. To show the utility of the approach, it is applied as a pre-processor to an existing fuzzy rule induction method [2].

The rest of this paper is structured as follows. Section 2 introduces the fuzzy rule induction algorithm used, illustrating its operation with an example. Section 3 discusses the fundamentals of rough set theory, in particular focusing on dimensionality reduction. The fourth section builds on these definitions to outline a procedure for fuzzy-rough attribute reduction. The modular design of the system is described in section 5, with experimental results presented in the sixth section. Section 7 concludes the paper, and proposes further work in this area.

2 Fuzzy Rule Induction

To show the potential utility of fuzzy-rough attribute reduction, the method is applied as a pre-processor to an existing fuzzy rule induction algorithm. The algorithm used is a recent one as described in [2]. The original dataset used to outline this induction procedure can be seen in table 1. There are three attributes each with corresponding linguistic terms, e.g. A has terms $A1$, $A2$ and $A3$. The decision attribute $Plan$ is also fuzzy, separated into three linguistic decisions X, Y and Z.

The algorithm begins by organising the dataset objects into subgroups according to their highest decision value. Within each subgroup, the fuzzy subsethood [8, 15] is calculated between the decisions of the subgroup and each attribute term. Fuzzy subsethood is defined as follows:

Case	A			B			C		Plan		
	A1	A2	A3	B1	B2	B3	C1	C2	X	Y	Z
1	0.3	0.7	0.0	0.2	0.7	0.1	0.3	0.7	0.1	0.9	0.0
2	1.0	0.0	0.0	1.0	0.0	0.0	0.7	0.3	0.8	0.2	0.0
3	0.0	0.3	0.7	0.0	0.7	0.3	0.6	0.4	0.0	0.2	0.8
4	0.8	0.2	0.0	0.0	0.7	0.3	0.2	0.8	0.6	0.3	0.1
5	0.5	0.5	0.0	1.0	0.0	0.0	0.0	1.0	0.6	0.8	0.0
6	0.0	0.2	0.8	0.0	1.0	0.0	0.0	1.0	0.0	0.7	0.3
7	1.0	0.0	0.0	0.7	0.3	0.0	0.2	0.8	0.7	0.4	0.0
8	0.1	0.8	0.1	0.0	0.9	0.1	0.7	0.3	0.0	0.0	1.0
9	0.3	0.7	0.0	0.9	0.1	0.0	1.0	0.0	0.0	0.0	1.0

Table 1: Original dataset

$$S(A,B) = \frac{M(A \cap B)}{M(A)} = \frac{\sum_{u \in \mathbb{U}} min(\mu_A(u), \mu_B(u))}{\sum_{u \in \mathbb{U}} \mu_A(u)} \quad (1)$$

For example, taking the subgroup of objects that belong to the decision X, the subsethood of the first term, $A1$, may be calculated as follows:

$$\begin{aligned}
\text{M}(X) &= 0.8 + 0.6 + 0.7 = 2.1 \\
\text{M}(X \cap A1) &= min(0.8, 1) + min(0.6, 0.8) \\
&\quad + min(0.7, 1) \\
&= 0.8 + 0.6 + 0.7 \\
&= 2.1 \\
\text{hence S}(X, A1) &= 2.1/2.1 = 1
\end{aligned}$$

From table 1, the following subsethood values can be obtained:

Subgroup1(X):
A:
$S(X, A1) = 1$, $S(X, A2) = 0.1$, $S(X, A3) = 0$
B:
$S(X, B1) = 0.71$, $S(X, B2) = 0.43$, $S(X, B3) = 0.14$
C:
$S(X, C1) = 0.52$, $S(X, C2) = 0.76$

Subgroup2(Y):
A:
$S(Y, A1) = 0.33$, $S(Y, A2) = 0.58$, $S(Y, A3) = 0.29$
B:
$S(Y, B1) = 0.42$, $S(Y, B2) = 0.58$, $S(Y, B3) = 0.04$
C:
$S(Y, C1) = 0.13$, $S(Y, C2) = 0.92$

Subgroup3(Z):
A:
$S(Z, A1) = 0.14$, $S(Z, A2) = 0.64$, $S(Z, A3) = 0.29$
B:
$S(Z, B1) = 0.32$, $S(Z, B2) = 0.61$, $S(Z, B3) = 0.14$
C:
$S(Z, C1) = 0.82$, $S(Z, C2) = 0.25$

These values are an indication of the relatedness of the individual terms to the decisions. A suitable level threshold, $\alpha \in [0,1]$, must be chosen beforehand in order to determine whether terms are close enough or not. At most, one term is selected per attribute. For example, setting $\alpha = 0.9$ means that the term with the highest fuzzy subsethood value (or its negation) above this threshold will be chosen. Applying this process to the first two decision values X and Y generates the rules:

Rule 1: IF A is $A1$ THEN $Plan$ is X
Rule 2: IF B is NOT $B3$ AND C is $C2$ THEN $Plan$ is Y

A problem is encountered here when there are no suitably representative terms for a decision (as is the case for decision Z). In this situation, a rule is produced that classifies cases to the decision value if the other rules do not produce reasonable classifications. This introduces another threshold value, $\beta \in [0,1]$, which determines whether a classification is reasonable or not. For decision Z, the following rule is produced:

Rule 3: IF $MF(Rule1) < \beta$ AND $MF(Rule2) < \beta$ THEN $Plan$ is Z

where MF(Rule i) = MF(condition part of Rule i) and MF means the membership function value. The classification results when using these rules on the example dataset can be found in the results section.

This technique has been shown to produce highly competitive results [2] in terms of both classification accuracy and number of rules generated. However, for most rule induction algorithms, the resultant rules may be unnecessarily complex due to the presence of redundant or misleading attributes. Fuzzy-Rough Attribute Reduction may be used here to significantly reduce dataset dimensionality, removing redundant attributes that would otherwise increase rule complexity and reducing the time for the induction process itself. This technique is based on Rough Set Attribute Reduction, which is outlined in the next section.

3 Rough Set Attribute Reduction

A rough set [11] is an approximation of a vague concept by a pair of precise concepts, called lower and upper approximations (which are informally a classification of the domain of interest into disjoint categories). Objects belonging to the same category characterised by the same attributes (or features) are not distinguishable.

Rough sets have been employed to remove redundant conditional attributes from discrete-valued datasets, while retaining their information content. A successful example of this is the Rough Set Attribute Reduction (RSAR) method [13]. Central to RSAR is the concept of indiscernibility. Let $I = (\mathbb{U}, A)$ be an information system, where $\mathbb{U}$ is a non-empty set of finite objects (the universe of discourse); A is a non-empty finite set of attributes such that $a : \mathbb{U} \rightarrow V_a \ \forall a \in A$, V_a being the value set of attribute a. In a decision system, $A = \{C \cup D\}$ where C is the set of conditional attributes and D is the set of decision attributes. With any $P \subseteq A$ there is an associated equivalence relation $IND(P)$:

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

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

$$\mathbb{U}/P = \otimes\{a \in P : \mathbb{U}/IND(\{a\})\}, \tag{3}$$

where

$$A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \emptyset\} \tag{4}$$

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 $X \subseteq \mathbb{U}$, the P-*lower* and P-*upper* approximations of a set can now be defined as:

$$\underline{P}X = \{x \mid [x]_P \subseteq X\} \tag{5}$$

$$\overline{P}X = \{x \mid [x]_P \cap X \neq \emptyset\} \tag{6}$$

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

$$POS_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \tag{7}$$

In terms of classification, the positive region contains all objects of $\mathbb{U}$ that can be classified to classes of $\mathbb{U}/Q$ using the knowledge in attributes P.

An important issue in data analysis is discovering dependencies between attributes. Intuitively, a set of attributes Q depends totally on a set of attributes P, denoted $P \Rightarrow Q$, if all attribute values from Q are uniquely determined by values of attributes from P. Dependency can be measured in the following way:

For $P, Q \subseteq A$, Q depends on P in a degree k $(0 \leq k \leq 1)$, denoted $P \Rightarrow_k Q$, if

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|\mathbb{U}|} \tag{8}$$

where $|S|$ stands for the cardinality of set S. If $k = 1$ Q depends totally on P, if $0 < k < 1$ Q depends partially (in a degree k) on P, and if $k = 0$ Q does not depend on P.

By calculating the change in dependency when an attribute is removed from the set of considered conditional attributes, an estimate of the significance of the attribute can be obtained. The higher the change in dependency, the more significant the attribute is. If the significance is 0, then the attribute is dispensible. More formally, given P, Q and an attribute $x \in P$, the significance of attribute x upon Q is defined by

$$\sigma_P(Q,x) = \gamma_P(Q) - \gamma_{P-\{x\}}(Q) \tag{9}$$

3.1 Reducts

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 quality of classification as the original. In the context of decision systems, a *reduct* is formally defined as a subset R of the conditional attribute set C such that $\gamma_R(D) = \gamma_C(D)$. A given dataset may have many attribute reduct sets, and the collection of all reducts is denoted by

$$R = \{X : X \subseteq C, \gamma_X(D) = \gamma_C(D)\} \tag{10}$$

The intersection of all the sets in R is called the *core*, the elements of which are those attributes that cannot be eliminated without introducing more contradictions to the dataset. In

RSAR, a reduct with minimum cardinality is searched for; in other words an attempt is made to locate a single element of the minimal reduct set $R_{min} \subseteq R$:

$$R_{min} = \{X : X \in R, \forall Y \in R, |X| \leq |Y|\} \tag{11}$$

A basic way of achieving this is to calculate the dependencies of all possible subsets of C. Any subset X with $\gamma_X(D) = 1$ is a reduct; the smallest subset with this property is a minimal reduct. However, for large datasets this method is impractical and an alternative strategy is required.

```
R ← {}
do
    T ← R
    ∀x ∈ (C − R)
        if γ_{R∪{x}}(D) > γ_T(D)
            T ← R ∪ {x}
    R ← T
until γ_R(D) = γ_C(D)
return R
```

Figure 1: The QuickReduct Algorithm

The QuickReduct algorithm given in figure 1, borrowed from [6, 13], attempts to calculate a minimal 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 $\gamma_P(Q)$, until this produces its maximum possible value for the dataset (usually 1). However, it has been proved that this method does not always generate a *minimal* reduct, as $\gamma_P(Q)$ is not a perfect heuristic. It does result in a close-to-minimal reduct, though, which is still useful in greatly reducing dataset dimensionality. An intuitive understanding of QuickReduct implies that, for a dimensionality of n, $n!$ evaluations of the dependency function may be performed for the worst-case dataset. From experimentation, the average complexity has been determined to be approximately O(n) [13].

4 Fuzzy-Rough Attribute Reduction

The RSAR process described previously can only operate effectively with datasets containing discrete values. As most datasets contain real-valued attributes, it is necessary to perform a discretization step beforehand. This is typically implemented by standard fuzzification techniques [13]. However, membership degrees of attribute values to fuzzy sets are not exploited in the process of RSAR. By using *fuzzy-rough* sets [3], it is possible to use this information to better guide attribute selection.

4.1 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 [3]. For typical RSAR 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 is defined by:

$$\mu_{[x]_S}(y) = \mu_S(x, y) \tag{12}$$

The following axioms hold for a fuzzy equivalence class F [5]:

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

The first axiom corresponds to the requirement that an equivalence class is non-empty. The second axiom states that elements in y's neighbourhood are in the equivalence class of y. The final axiom states that any two elements in F are related via S. Obviously, this definition degenerates to the normal definition of equivalence classes when S is non-fuzzy.

The family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes [3]. Consider the crisp partitioning $\mathbb{U}/Q = \{\{1,3,6\},\{2,4,5\}\}$. This contains two equivalence classes ({1,3,6} and {2,4,5}) that can be thought of as degenerated fuzzy sets, with those elements belonging to the class possessing a membership of one, zero otherwise. For the first class, for instance, the objects 2, 4 and 5 have a membership of zero. Extending this to the case of fuzzy equivalence classes is straightforward: objects can be allowed to assume membership values, with respect to any given class, in the interval [0,1]. $\mathbb{U}/Q$ is not restricted to crisp partitions only; fuzzy partitions are equally acceptable.

4.2 Fuzzy Lower and Upper Approximations

From the literature, the fuzzy P-lower and P-upper approximations are defined as [3]:

$$\mu_{\underline{P}X}(F_i) = inf_x max\{1 - \mu_{F_i}(x), \mu_X(x)\} \quad \forall i \tag{13}$$

$$\mu_{\overline{P}X}(F_i) = sup_x min\{\mu_{F_i}(x), \mu_X(x)\} \quad \forall i \tag{14}$$

where F_i denotes a fuzzy equivalence class belonging to $\mathbb{U}/P$. Note that although the universe of discourse in attribute reduction is finite, this is not the case in general, hence the use of sup and inf. 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. As a result of this, the fuzzy lower and upper approximations are herein redefined as:

$$\mu_{\underline{P}X}(x) = \sup_{F \in \mathbb{U}/P} min(\mu_F(x), \inf_{y \in \mathbb{U}} max\{1 - \mu_F(y), \mu_X(y)\}) \tag{15}$$

$$\mu_{\overline{P}X}(x) = \sup_{F \in \mathbb{U}/P} min(\mu_F(x), \sup_{y \in \mathbb{U}} min\{\mu_F(y), \mu_X(y)\}) \tag{16}$$

In implementation, not all $y \in \mathbb{U}$ are needed to be considered - only those where $\mu_F(y)$ is non-zero, i.e. where object y is a fuzzy member of (fuzzy) equivalence class F. The tuple $< \underline{P}X, \overline{P}X >$ is called a fuzzy-rough set. It can be seen that these definitions degenerate to traditional rough sets when all equivalence classes are crisp.

4.3 Fuzzy-Rough Reduction Process

Fuzzy RSAR builds on the notion of the fuzzy lower approximation to enable reduction of datasets containing real-valued attributes. As will be shown, the process becomes identical to the traditional approach when dealing with nominal well-defined attributes.

The crisp positive region in traditional rough set theory 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

$$\mu_{POS_P(Q)}(x) = \sup_{X \in \mathbb{U}/Q} \mu_{\underline{P}X}(x) \tag{17}$$

Object x will not belong to the positive region only if the equivalence class it belongs to is not a constituent of the positive region. This is equivalent to the crisp version where objects belong to the positive region only if their underlying equivalence class does so.

Using the definition of the fuzzy positive region, the new dependency function can be defined as follows

$$\gamma'_P(Q) = \frac{|\mu_{POS_P(Q)}(x)|}{|\mathbb{U}|} = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_P(Q)}(x)}{|\mathbb{U}|} \tag{18}$$

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 $|\mu_{POS_P(Q)}(x)|$, the fuzzy cardinality of $\mu_{POS_P(Q)}(x)$, divided by the total number of objects in the universe. The definition of dependency degree covers the crisp case as its specific instance.

If the fuzzy-rough reduction process is to be useful, it must be able to deal with multiple attributes, finding the dependency between various subsets of the original attribute set. For example, it may be necessary to be able to determine the degree of dependency of the decision attribute(s) with respect to $P = \{a, b\}$. In the crisp case, $\mathbb{U}/P$ contains sets of objects grouped together that are indiscernible according to both attributes 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,

$$\mathbb{U}/P = \otimes\{a \in P : \mathbb{U}/IND(\{a\})\} \tag{19}$$

Each set in $\mathbb{U}/P$ denotes an equivalence class. For example, if $P = \{a, b\}$, $\mathbb{U}/IND(\{a\}) = \{N_a, Z_a\}$ and $\mathbb{U}/IND(\{b\}) = \{N_b, Z_b\}$, then

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

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

$$\mu_{F_1 \cap \ldots \cap F_n}(x) = min(\mu_{F_1}(x), \mu_{F_2}(x), ..., \mu_{F_n}(x)) \quad (20)$$

4.4 Reduct Computation

A problem may arise when this approach is compared to the original crisp one. In RSAR, a reduct is defined as a subset R of the attributes which have the same information content as the full attribute set A. In terms of the dependency function this means that the values $\gamma(R)$ and $\gamma(A)$ are identical and equal to 1 if the dataset is consistent. However, in the fuzzy-rough approach this is not necessarily the case as the uncertainty encountered when objects belong to many fuzzy equivalence classes results in a reduced total dependency.

$R \leftarrow \{\}$, $\gamma'_{best} \leftarrow 0$, $\gamma'_{prev} \leftarrow 0$
do
  $T \leftarrow R$
  $\gamma'_{prev} \leftarrow \gamma'_{best}$
  $\forall x \in (C - R)$
    if $\gamma'_{R \cup \{x\}}(D) > \gamma'_T(D)$
      $T \leftarrow R \cup \{x\}$
      $\gamma'_{best} \leftarrow \gamma'_T(D)$
  $R \leftarrow T$
until $\gamma'_{best} = \gamma'_{prev}$
return R

Figure 2: The fuzzy-rough QuickReduct algorithm

A possible way of combatting this would be to determine the degree of dependency of the full attribute set and use this as the denominator, allowing γ' to reach 1. With these issues in mind, a new QuickReduct algorithm has been developed as given in figure 2. It employs the new dependency function γ' to choose which attributes to add to the current reduct candidate in the same way as the original QuickReduct process. The algorithm terminates when the addition of any remaining attribute does not increase the dependency (such a criterion could be used with the original QuickReduct algorithm). For a dimensionality of n, the worst case dataset will still result in $n!$ evaluations of the dependency function. However, as fuzzy RSAR is used for dimensionality reduction prior to any involvement of the system which will employ those attributes belonging to the resultant reduct, this potentially costly operation has no negative impact upon the run-time efficiency of the system.

Using the fuzzy-rough QuickReduct algorithm, table 1 can be reduced in size. Firstly, the initial reduct candidate is the empty set and the dependency of each individual conditional attribute is calculated:

$\gamma'_{\{A\}}(\{Plan\}) = 3.7/9$
$\gamma'_{\{B\}}(\{Plan\}) = 1.3/9$
$\gamma'_{\{C\}}(\{Plan\}) = 2.4/9$

As attribute A causes the greatest increase in dependency degree, it is added to the reduct candidate and the search progresses:

$\gamma'_{\{A,B\}}(\{Plan\}) = 3.7/9,$
$\gamma'_{\{A,C\}}(\{Plan\}) = 4.7/9$

Here, C is added to the reduct candidate as the dependency is increased. There is only one attribute addition to be checked at the next stage, namely

$\gamma'_{\{A,B,C\}}(\{Plan\}) = 4.7/9$

This causes no dependency increase, resulting in the algorithm terminating and outputting the reduct $\{A, C\}$. Hence, the original dataset can be reduced to these attributes with minimal information loss (according to the algorithm). The resulting fuzzy rules using the smaller dataset can be found in the section 6.

5 Overview of the Implementation

The system implemented to demonstrate the ideas is herein comprised of several modules (figure 3), allowing various sub-components to be replaced with alternative techniques. Firstly, the dataset from which rules are to be generated is input into the system. Depending on the attribute format of the dataset, an additional fuzzification step may need to be carried out.

The Fuzzy Rough Attribute Reduction (FRAR) module takes as input any dataset containing real-valued or crisp attributes, or attributes defined entirely by their fuzzy membership degrees. Using this information only (no thresholds are required), the dataset is reduced if it involves redundant attributes. When no attribute reduction is desired, the FRAR module is bypassed and the unprocessed dataset is sent directly to the Rule Induction Algorithm (RIA).

Given the dataset (either reduced or unreduced), the RIA extracts fuzzy rules for use in

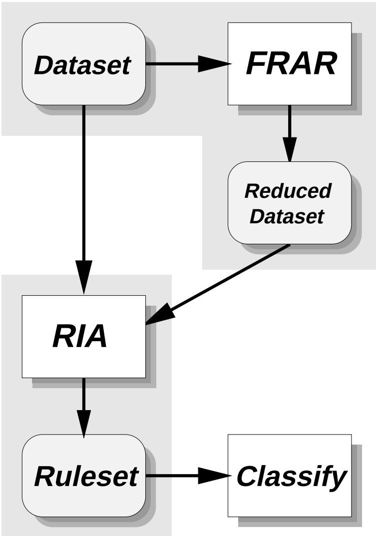

Figure 3: Block diagram of the implemented system

the classification phase. Due to the modularity of the system, different RIAs other than the one outlined previously may be used here. Using these fuzzy rules, data may then be classified into the categories present in the training dataset.

6 Results

The original rules produced using table 1 and the fuzzy rule induction procedure outlined in section 2 can be seen in figure 4. These rules are generated using the entire dataset, with $\alpha = 0.9$ and $\beta = 0.6$. Using these rules, the dataset may be classified and the generated classifications compared with the original dataset's plan.

Rule 1: IF A is $A1$ THEN $Plan$ is X
Rule 2: IF B is NOT $B3$ AND C is $C2$ THEN $Plan$ is Y
Rule 3: IF $MF(Rule1) < \beta$ AND $MF(Rule2) < \beta$ THEN $Plan$ is Z

Figure 4: Original rules

As has been shown in section 4, the dataset may be reduced by the removal of attribute B with little reduction in classification accuracy (according to FRAR). Using this reduced dataset, the rule induction algorithm generates the rules given in figure 5. From this, it can be seen that rule 2 has been simplified due to the redundancy of attribute B. Although the extent of simplification is small in this case, with larger datasets the effect can be expected to be greater.

Rule 1: IF A is $A1$ THEN $Plan$ is X
Rule 2: IF C is $C2$ THEN $Plan$ is Y
Rule 3: IF $MF(Rule1) < \beta$ AND $MF(Rule2) < \beta$ THEN $Plan$ is Z

Figure 5: Generated rules using the reduced dataset

Table 2 shows the membership degrees of the cases to each classification for the calculated plan and the actual plan present in the dataset. It can be seen that the resulting classifications are the same when the min operator is used.

Case	Calc.			Actual		
	X	Y	Z	X	Y	Z
1	0.3	0.7	0.0	0.1	0.9	0.0
2	1.0	0.3	0.0	0.8	0.2	0.0
3	0.0	0.4	1.0	0.0	0.2	0.8
4	0.8	0.7	0.0	0.6	0.3	0.1
5	0.5	1.0	0.0	0.6	0.8	0.0
6	0.0	1.0	0.0	0.0	0.7	0.3
7	1.0	0.8	0.0	0.7	0.4	0.0
8	0.1	0.3	1.0	0.0	0.0	1.0
9	0.3	0.0	1.0	0.0	0.0	1.0

Table 2: Resulting plan with all attributes

The results using the FRAR-reduced dataset are provided in table 3. The differences between the classifications of the reduced and unreduced approaches have been highlighted (cases 4 and 7). In case 4, only the membership degree for Y has changed. This value has increased from 0.7 to 0.8, resulting in an ambiguous classification. Again, for case 7, the membership degree for Y is the only value to have changed; this time it more closely resembles the classification present in the training dataset.

Case	Calc.			Actual		
	X	Y	Z	X	Y	Z
1	0.3	0.7	0.0	0.1	0.9	0.0
2	1.0	0.3	0.0	0.8	0.2	0.0
3	0.0	0.4	1.0	0.0	0.2	0.8
4	**0.8**	**0.8**	**0.0**	0.6	0.3	0.1
5	0.5	1.0	0.0	0.6	0.8	0.0
6	0.0	1.0	0.0	0.0	0.7	0.3
7	**1.0**	**0.3**	**0.0**	0.7	0.4	0.0
8	0.1	0.3	1.0	0.0	0.0	1.0
9	0.3	0.0	1.0	0.0	0.0	1.0

Table 3: Resulting plan with reduced attributes

7 Conclusion

This paper has shown the potential utility of employing Fuzzy Rough Attribute Reduction as a pre-processor to fuzzy rule induction. Not only are the runtimes of the induction and classification processes improved by this step (which for some systems are important factors), but the resulting rules are less complex. The resulting loss in classification accuracy for the presented example is small. Indeed, for one case the accuracy has been *improved*. Further experimentation is required to investigate the impact of the attribute reduction on runtime and rule complexity for larger datasets.

Currently, work is being carried out on a fuzzified dependency function. Ordinarily, the dependency function returns values for sets of attributes in the range [0,1]; the fuzzy dependency function will return qualitative fuzzy labels for use in the new QuickReduct algorithm. Additionally, research is being carried out into the potential utility of *fuzzy reducts*, which would allow attributes to have a varying possibility of becoming a member of the resultant reduct.

Acknowledgements

This work is partly funded by the UK EPSRC grant 00317404. The authors are grateful to Dave Robertson and the rest of the Advanced Knowledge Technologies team at Edinburgh for their support.

References

[1] Advanced Knowledge Technologies homepage: http://www.aktors.org/

[2] S. Chen, S.L. Lee and C. Lee. A new method for generating fuzzy rules from numerical data for handling classification problems. Applied Artificial Intelligence, 15(7):645-664, 2001.

[3] D. Dubois and H. Prade. Putting rough sets and fuzzy sets together. In [14] pp203-232. 1992.

[4] A Brief Introduction to Rough Sets, Copyright 1993, EBRSC. Information available at: http://cs.uregina.ca/roughset/rs.intro.txt

[5] U. Höhle. Quotients with respect to similarity relations. Fuzzy Sets and Systems 27, pp. 31-44.

[6] R. Jensen and Q. Shen. A Rough Set-Aided System for Sorting WWW Bookmarks. In N. Zhong et al. (Eds.), Web Intelligence: Research and Development. pp. 95-105, 2001.

[7] R. Jensen and Q. Shen. Fuzzy-Rough Sets for Descriptive Dimensionality Reduction. Proceedings of the 11th International Conference on Fuzzy Systems, pp. 29-34, 2002.

[8] B. Kosko. Fuzzy entropy and conditioning. Information Sciences 40(2): 165-174. 1986.

[9] O. Maudal. Preprocessing data for Neural Network based Classifiers: Rough Sets vs Principal Component Analysis. Project Report, Department of Artificial Intelligence, The University of Edinburgh, 1996.

[10] T. Mitchell. Machine Learning. McGraw-Hill, 1997

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

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

[13] Q. Shen and A. Chouchoulas. A modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems. Engineering Applications of Artificial Intelligence, 13(3):263-278, 2000.

[14] R. Slowinski, editor. Intelligent Decision Support. Kluwer Academic Publishers, Dordrecht, 1992.

[15] Y. Yuan and M. J. Shaw. Induction of fuzzy decision trees. Fuzzy Sets and Systems 69(2): 125-139. 1995.