School of Informatics, University of Edinburgh

Centre for Intelligent Systems and their Applications

Fuzzy rough attribute reduction with application to web categorization

by

Richard Jensen, Qiang Shen

Informatics Research Report EDI-INF-RR-0197

School of Informatics **January 2004**
http://www.informatics.ed.ac.uk/

Fuzzy rough attribute reduction with application to web categorization

Richard Jensen, Qiang Shen

Informatics Research Report EDI-INF-RR-0197

SCHOOL *of* INFORMATICS
Centre for Intelligent Systems and their Applications

January 2004

appears in Fuzzy Sets and Systems

Abstract :

Due to the explosive growth of electronically stored information, automatic methods must be developed to aid users in maintaining and using this abundance of information e+ectively. In particular, the sheer volume of redundancy present must be dealt with, leaving only the information-rich data to be processed. This paper presents a novel approach, based on an integrated use of fuzzy and rough set theories, to greatly reduce this data redundancy. Formal concepts of fuzzy rough attribute reduction are introduced and illustrated with a simple example. The work is applied to the problem of web categorization, considerably reducing dimensionality with minimal loss of information. Experimental results show that fuzzy rough reduction is more powerful than the conventional rough set-based approach. Classi2ers that use a lower dimensional set of attributes which are retained by fuzzy rough reduction outperform those that employ more attributes returned by the existing crisp rough reduction method.

Keywords : Attribute reduction, Web categorization, Data redundancy

ARTICLE IN PRESS

Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT

ELSEVIER

Fuzzy Sets and Systems III (IIII) III–III

FUZZY
sets and systems

www.elsevier.com/locate/fss

Fuzzy–rough attribute reduction with application to web categorization

Richard Jensen, Qiang Shen*

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

Received 25 April 2002; received in revised form 25 September 2002; accepted 9 January 2003

Abstract

Due to the explosive growth of electronically stored information, automatic methods must be developed to aid users in maintaining and using this abundance of information effectively. In particular, the sheer volume of redundancy present must be dealt with, leaving only the information-rich data to be processed. This paper presents a novel approach, based on an integrated use of fuzzy and rough set theories, to greatly reduce this data redundancy. Formal concepts of fuzzy–rough attribute reduction are introduced and illustrated with a simple example. The work is applied to the problem of web categorization, considerably reducing dimensionality with minimal loss of information. Experimental results show that fuzzy–rough reduction is more powerful than the conventional rough set-based approach. Classifiers that use a lower dimensional set of attributes which are retained by fuzzy–rough reduction outperform those that employ more attributes returned by the existing crisp rough reduction method.

Keywords: Attribute reduction; Web categorization; Data redundancy

1. Introduction

It is well known that the amount of electronically stored information increases exponentially with time. Sorting through even a fraction of the available data by hand can be very difficult. Information filtering and retrieval systems are therefore acquiring increasing prominence as automated aids in quickly sifting through information.

The World Wide Web (WWW) is an information resource, whose full potential may not be realized unless its content is adequately organized and described. This not only applies to the vast network of web pages, but also to users' personal repositories of web page bookmarks. However,

* Corresponding author.
E-mail addresses: richjens@dai.ed.ac.uk (R. Jensen), qiangs@dai.ed.ac.uk (Q. Shen).

0165-0114/03/$ - see front matter
doi:10.1016/S0165-0114(03)00021-6

due to the immense size and dynamicity of the web, manual categorization is not a practical solution to this problem. There is a clear need for automated classification of web content.

Many classification problems involve high-dimensional descriptions of input features. It is therefore not surprising that much research has been done on dimensionality reduction [4,11,12]. However, existing work tends to destroy the underlying semantics of the features after reduction (e.g. transformation-based approaches [5]) or require additional information about the given data set for thresholding (e.g. entropy-based approaches [15]). A technique that can reduce dimensionality using information contained within the data set and preserving the meaning of the features is clearly desirable. Rough set theory (RST) can be used as such a tool to discover data dependencies and reduce the number of attributes contained in a data set by purely structural methods [17].

Over the past 10 years, RST has indeed become a topic of great interest to researchers and has been applied to many domains. This success is due in part to the following aspects of the theory:

- Only the facts hidden in data are analysed.
- No additional information about the data is required such as thresholds or expert knowledge.
- A minimal knowledge representation can be attained.

Given a data set with discretized attribute values, it is possible to find a subset (termed a *reduct*) of the original attributes using RST that are the most informative; all other attributes can be removed from the data set with minimal information loss.

However, it is most often the case that the values of attributes may be both crisp and *real-valued*, and this is where traditional rough set theory encounters a problem. It is not possible in the theory to say whether two attribute values are similar and to what extent they are the same; for example, two close values may only differ as a result of noise, but in RST they are considered to be as different as two values of a different order of magnitude.

One answer to this problem has been to discretize the data set beforehand, producing a new data set with crisp values. This is often still inadequate, however, as the degrees of membership of values to discretized values are not considered at all. For example, two values may both be mapped to the same class "Negative", but one may be much more negative than the other. This is a source of information loss, which is against the rough set ideology of retaining information content.

It is, therefore, desirable to develop these techniques to provide the means of data reduction for crisp and real-value attributed data sets which utilizes the extent to which values are similar. This could be achieved through the use of *fuzzy–rough* sets. Fuzzy–rough sets encapsulate the related but distinct concepts of vagueness (for fuzzy sets [27]) and indiscernibility (for rough sets), both of which occur as a result of uncertainty in knowledge [6]. This paper, based on initial work reported in [9,10], presents such a method which employs fuzzy–rough sets to improve the handling of this uncertainty. Newly introduced concepts for fuzzy–rough attribute reduction are formally defined and illustrated with a simple example. The work is applied to the domain of website classification with very promising results. In particular, the present approach retains less attributes than the crisp RST-based reduction method, whilst entailing higher classification accuracy.

The rest of this paper is structured as follows. Section 2 discusses the fundamentals of rough set theory, in particular focusing on dimensionality reduction. The third section introduces the hybrid of rough and fuzzy sets and builds on these definitions to outline a procedure for fuzzy–rough attribute reduction. To help in the understanding of this procedure, a demonstrative case is given

to show the key stages involving the use of the introduced concepts. The work is then applied to website categorization and compared with traditional crisp RSAR, demonstrating the power of the new approach. Finally, the paper concludes with a discussion of the results and outlines future work to be carried out.

2. Background

The theory of rough sets provides rigorous mathematical techniques for creating approximate descriptions of objects for data analysis, optimization and recognition. A rough set itself is an approximation of a vague concept by a pair of precise concepts, called lower and upper approximations [7,17]. 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.

2.1. Rough set attribute reduction

Rough sets have been employed to remove redundant conditional attributes from discrete-valued data sets, while retaining their information content. A successful example of this is the rough set attribute reduction (RSAR) method [21]. 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 $IND(P)$:

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

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

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

where

$$A\otimes B=\{X\cap Y:\ \forall X\in A,\forall Y\in B,X\cap Y\neq\emptyset\}. \tag{3}$$

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

$$\underline{P}X=\{x\,|\,[x]_P\subseteq X\}. \tag{4}$$

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

$$POS_P(Q)=\bigcup_{X\in\mathbb{U}/Q}\underline{P}X. \tag{5}$$

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 defined in the following way:

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

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|\mathbb{U}|}, \tag{6}$$

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, a measure 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

$$\sigma_P(Q, x) = \gamma_P(Q) - \gamma_{P-\{x\}}(Q). \tag{7}$$

2.2. 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 data set may have many attribute reduct sets, and the collection of all reducts is denoted by

$$R = \{X: \ X \subseteq C, \gamma_X(D) = \gamma_C(D)\} \tag{8}$$

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 data set. 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 $R_{\min} \subseteq R$:

$$R_{\min} = \{X: \ X \in R, \ \forall Y \in R, |X| \leqslant |Y|\}. \tag{9}$$

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 data sets this method is impractical and an alternative strategy is required.

The QuickReduct algorithm given in Fig. 1, borrowed from [9,21], 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 data set (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 data set dimensionality.

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

Fig. 1. The QUICKREDUCT algorithm.

An intuitive understanding of QUICKREDUCT implies that, for a dimensionality of n, $(n^2 + n)/2$ evaluations of the dependency function may be performed for the worst-case data set. From experimentation, the average complexity has been determined to be approximately O(n) [21].

3. Fuzzy–rough attribute reduction

The RSAR process described previously can only operate effectively with data sets containing discrete values. As most data sets contain real-valued attributes, it is necessary to perform a discretization step beforehand. This is typically implemented by standard fuzzification techniques [21]. However, membership degrees of attribute values to fuzzy sets are not exploited in the process of dimensionality reduction. By using *fuzzy–rough* sets [6,22,16], it is possible to use this information to better guide feature selection. This forms the central contribution of this paper.

3.1. 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 [6,23,26]. 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 ($\mu_S(x,z) \geqslant \mu_S(x,y) \wedge \mu_S(y,z)$) hold.

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

$$\mu_{[x]_S}(y) = \mu_S(x,y). \tag{10}$$

The following axioms should hold for a fuzzy equivalence class F [8]:

- $\exists x,\ \mu_F(x) = 1$,

- $\mu_F(x) \wedge \mu_S(x, y) \leqslant \mu_F(y)$,
- $\mu_F(x) \wedge \mu_F(y) \leqslant \mu_S(x, y)$.

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 [6]. 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.

3.2. Fuzzy lower and upper approximations

From the literature, the fuzzy P-lower and P-upper approximations are defined as [6]

$$\mu_{\underline{P}X}(F_i) = \inf_x \max\{1 - \mu_{F_i}(x), \mu_X(x)\} \quad \forall i, \tag{11}$$

$$\mu_{\overline{P}X}(F_i) = \sup_x \min\{\mu_{F_i}(x), \mu_X(x)\} \quad \forall i, \tag{12}$$

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

$$\mu_{\underline{P}X}(x) = \sup_{F \in \mathbb{U}/P} \min\left(\mu_F(x), \inf_{y \in \mathbb{U}} \max\{1 - \mu_F(y), \mu_X(y)\}\right), \tag{13}$$

$$\mu_{\overline{P}X}(x) = \sup_{F \in \mathbb{U}/P} \min\left(\mu_F(x), \sup_{y \in \mathbb{U}} \min\{\mu_F(y), \mu_X(y)\}\right). \tag{14}$$

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 $\langle \underline{P}X, \overline{P}X \rangle$ is called a fuzzy–rough set. It can be seen that these definitions degenerate to traditional rough sets when all equivalence classes are crisp. It is useful to think of the crisp lower approximation as characterized by the following membership function:

$$\mu_{\underline{P}X}(x) = \begin{cases} 1, & x \in F,\ F \subseteq X, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

This states that an object x belongs to the P-lower approximation of X if it belongs to an equivalence class that is a subset of X. The behaviour of the fuzzy lower approximation must be exactly that of the crisp definition for crisp situations. This is indeed the case as the fuzzy lower approximation may be rewritten as

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

where $\to$ denotes the fuzzy implication operator. In the crisp case, $\mu_F(x)$ and $\mu_X(x)$ will take values from $\{0,1\}$. Hence, it is clear that the only time $\mu_{\underline{P}X}(x)$ will be zero is when at least one object in its equivalence class F fully belongs to F but not to X. This is exactly the same as the definition for the crisp lower approximation. Similarly, the definition for the P-upper approximation can be established to make sense.

Other generalizations are possible [18]. In their work, Beaubouef et al. [2] relate the concepts of information theoretic measures to rough sets, comparing these to established rough set models of uncertainty. They apply the work to the rough and fuzzy–rough relational database models. However, they choose an alternative definition of fuzzy–rough sets which originates from the rough membership function [17].

Rough sets may be expressed by a fuzzy membership function to represent the negative, boundary and positive regions [24]. All objects in the positive region have a membership of one and those belonging to the boundary region have a membership of 0.5. Those that are contained in the negative region (and therefore do not belong to the rough set) have zero membership. In so doing, a rough set can be expressed as a fuzzy set, with suitable modifications to the rough union and intersection operators.

The reason for integrating fuzziness into the rough set model is to quantify the levels of roughness in the boundary region by using fuzzy membership values. It is necessary, therefore, to allow elements in the boundary region to have membership values in the range of 0–1, not just the value 0.5. Hence, a fuzzy rough set Y is defined (by this approach) as a membership function $\mu_Y(x)$ that associates a grade of membership from the interval [0,1] with every element of $\mathbb{U}$. For a rough set X and a crisp equivalence relation R:

$$\mu_Y(\underline{R}X) = 1,$$
$$\mu_Y(\mathbb{U} - \overline{R}X) = 0,$$
$$0 < \mu_Y(\overline{R}X - \underline{R}X) < 1.$$

However, this is not a true hybridization of the two approaches, it merely assigns a degree of membership to the elements depending on the crisp positive, boundary or negative region they belong to. Fuzzy equivalence classes are not used and so this does not offer a particularly useful approach for fuzzy–rough attribute reduction.

3.3. Fuzzy–rough reduction process

Fuzzy RSAR builds on the notion of the fuzzy lower approximation to enable reduction of data sets 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 data set. 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.

The definition of dependency degree covers the crisp case as its specific instance. This can be easily shown by recalling the definition of the crisp dependency degree given in (6). If a function $\mu_{POS_P(Q)}(x)$ is defined which returns 1 if the object x belongs to the positive region, 0 otherwise, then the above definition may be rewritten as

$$\gamma_P(Q) = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_P(Q)}(x)}{|\mathbb{U}|}, \tag{19}$$

which is identical to (18).

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,

$$\mathbb{U}/P = \otimes\{a \in P:\ \mathbb{U}/IND(\{a\})\}. \tag{20}$$

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, $i = 1, 2, \ldots, n$:

$$\mu_{F_1 \cap \cdots \cap F_n}(x) = \min(\mu_{F_1}(x), \mu_{F_2}(x), \ldots, \mu_{F_n}(x)). \tag{21}$$

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

Fig. 2. The fuzzy–rough QUICKREDUCT algorithm.

3.4. Reduct computation

A problem may arise when this approach is compared to the crisp approach. In conventional 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 data set 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.

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 (for normalization rather than $|\mathbb{U}|$), allowing γ' to reach 1. With these issues in mind, a new QUICKREDUCT algorithm has been developed as given in Fig. 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). As with the original QUICKREDUCT algorithm, for a dimensionality of n, the worst case data set will result in $(n^2 + n)/2$ 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.

Note that it is also possible to reverse the search process; that is, start with the full set of attributes and incrementally remove the least informative attributes. This process continues until no more attributes can be removed without reducing the total number of discernible objects in the data set.

3.5. Fuzzy RSAR example

To illustrate the operation of fuzzy RSAR, an example data set is given here. In crisp RSAR, the data set would be discretized using the non-fuzzy sets. However, in the new approach membership

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

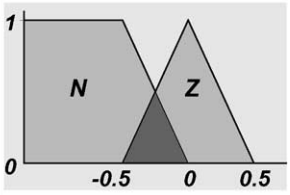

Fig. 3. Data set and corresponding fuzzy sets.

degrees are used in calculating the fuzzy lower approximations and fuzzy positive regions. To begin with, the fuzzy–rough QuickReduct algorithm initializes the potential reduct (i.e. the current best set of attributes) to the empty set.

Using the fuzzy sets defined in Fig. 3 (for all conditional attributes), and setting $A=\{a\}$, $B=\{b\}$, $C=\{c\}$ and $Q=\{q\}$, the following equivalence classes are obtained:

$$\mathbb{U}/A = \{N_a, Z_a\},$$

$$\mathbb{U}/B = \{N_b, Z_b\},$$

$$\mathbb{U}/C = \{N_c, Z_c\},$$

$$\mathbb{U}/Q = \{\{1,3,6\},\{2,4,5\}\}.$$

The first step is to calculate the lower approximations of the sets A, B and C. For simplicity, only A will be considered here; that is, using A to approximate Q. For the first decision equivalence class $X=\{1,3,6\}$, $\mu_{\underline{A}\{1,3,6\}}(x)$ needs to be calculated:

$$\mu_{\underline{A}\{1,3,6\}}(x) = \sup_{F\in\mathbb{U}/A} \min\left(\mu_F(x), \inf_{y\in\mathbb{U}} \max\{1-\mu_F(y), \mu_{\{1,3,6\}}(y)\}\right).$$

Considering the first fuzzy equivalence class of A, N_a:

$$\min\left(\mu_{N_a}(x), \inf_{y\in\mathbb{U}} \max\{1-\mu_{N_a}(y), \mu_{\{1,3,6\}}(y)\}\right).$$

For object 2 this can be calculated as follows:

$$\min(0.8, \inf\{1, 0.2, 1, 1, 1, 1\}) = 0.2.$$

Similarly for Z_a

$$\min(0.2, \inf\{1, 0.8, 1, 0.6, 0.4, 1\}) = 0.2.$$

Thus,

$$\mu_{\underline{A}\{1,3,6\}}(2) = 0.2.$$

Calculating the A-lower approximation of $X = \{1, 3, 6\}$ for every object gives

$$\mu_{\underline{A}\{1,3,6\}}(1) = 0.2, \quad \mu_{\underline{A}\{1,3,6\}}(2) = 0.2,$$
$$\mu_{\underline{A}\{1,3,6\}}(3) = 0.4, \quad \mu_{\underline{A}\{1,3,6\}}(4) = 0.4,$$
$$\mu_{\underline{A}\{1,3,6\}}(5) = 0.4, \quad \mu_{\underline{A}\{1,3,6\}}(6) = 0.4.$$

The corresponding values for $X = \{2, 4, 5\}$ can also be determined this way. Using these values, the fuzzy positive region for each object can be calculated via using

$$\mu_{POS_A(Q)}(x) = \sup_{X \in \mathbb{U}/Q} \mu_{\underline{A}X}(x).$$

This results in

$$\mu_{POS_A(Q)}(1) = 0.2, \quad \mu_{POS_A(Q)}(2) = 0.2,$$
$$\mu_{POS_A(Q)}(3) = 0.4, \quad \mu_{POS_A(Q)}(4) = 0.4,$$
$$\mu_{POS_A(Q)}(5) = 0.4, \quad \mu_{POS_A(Q)}(6) = 0.4.$$

It is a coincidence here that $\mu_{POS_A(Q)}(x) = \mu_{\underline{A}\{1,3,6\}}(x)$ for this example. The next step is to determine the degree of dependency of Q on A:

$$\gamma'_A(Q) = \frac{\sum_{x \in U} \mu_{POS_A(Q)}(x)}{|U|} = 2/6.$$

Calculating for B and C gives

$$\gamma'_B(Q) = \frac{2.4}{6}, \quad \gamma'_C(Q) = \frac{1.6}{6}.$$

From this it can be seen that attribute b will cause the greatest increase in dependency degree. This attribute is chosen and added to the potential reduct. The process iterates and the two dependency degrees calculated are

$$\gamma'_{\{a,b\}}(Q) = \frac{3.4}{6}, \quad \gamma'_{\{b,c\}}(Q) = \frac{3.2}{6}.$$

Adding attribute a to the reduct candidate causes the larger increase of dependency, so the new candidate becomes $\{a, b\}$. Lastly, attribute c is added to the potential reduct:

$$\gamma'_{\{a,b,c\}}(Q) = \frac{3.4}{6}.$$

As this causes no increase in dependency, the algorithm stops and outputs the reduct $\{a, b\}$. The steps taken by the fuzzy-rough QuickReduct algorithm in reaching this decision can be seen in Fig. 4. The data set can now be reduced to only those attributes appearing in the reduct. When crisp RSAR is performed on this data set (after using the same fuzzy sets to discretize the real-valued attributes), the reduct generated is $\{a, b, c\}$, i.e. the full conditional attribute set [10]. Unlike crisp RSAR, the true minimal reduct was found using the information on degrees of membership. It is clear from this example alone that the information lost by using crisp RSAR can be important when trying to discover the smallest reduct from a data set.

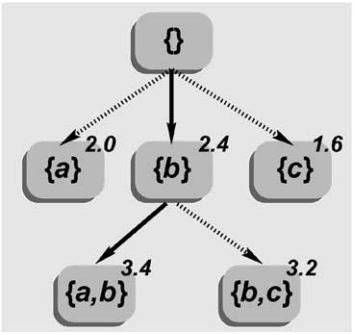

Fig. 4. Path taken by the fuzzy–rough QUICKREDUCT algorithm.

4. Application to the web

To demonstrate the applicability of the described methods, two relevant domains of interest were selected, namely bookmark classification and website classification. However, these domains are quite distinct, possessing features and problems that must be independently addressed. This section will initially present those features that are common to both before focusing on domain-dependent issues.

4.1. System overview

Both applications employ a similar general system architecture in order to reduce dimensionality and perform categorization. A key issue in the design of the system was that of modularity; it should be able to integrate with existing (or new) techniques. The current implementations allow this flexibility by dividing the overall process into several independent sub-modules (see Fig. 5):

- *Splitting of training and testing data sets*. Data sets were generated from large textual corpora and separated randomly into training and testing sets. Each data set is a collection of documents, either bookmarks or web pages depending on the application.
- *Keyword acquisition*. Given the output from the previous module, keywords/terms are extracted and weighted according to their perceived importance in the document, resulting in a new data set of weight-term pairs. Note that in this work, no sophisticated keyword acquisition techniques methods are used as the current focus of attention is on the evaluation of attribute reduction. However, the use of more effective keyword acquisition techniques recently built in the area of information retrieval would help improve the system's overall classification performance further.
- *Keyword selection*. As the newly generated data sets are too large, mainly due to keyword redundancy, to perform classification at this stage, a dimensionality reduction step is carried out using the techniques described previously.
- *Keyword filtering*. Used only in testing, this simple module filters the keywords obtained during acquisition, using the reduct generated in the keyword selection module.
- *Classification*. This final module uses the reduced data set to perform the actual categorization of the test data. Classifiers used are the vector space model (VSM) [20] and the Boolean inexact model (BIM) [19]. Again, more efficient and effective classifiers can be employed for this, but

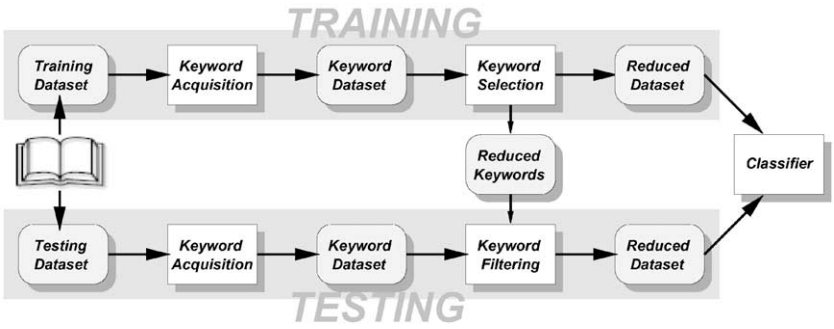

Fig. 5. Modular decomposition of the classification system.

for simplicity only these conventional classifiers are adopted here to show the power of attribute reduction. Better classifiers are expected to produce more accurate results, though not necessarily enhance the comparisons between classifiers that use reduced or unreduced data sets.

4.2. Bookmark classification

As the use of the World Wide Web becomes more prevalent and the size of personal repositories grows, adequately organizing and managing bookmarks becomes crucial. Several years ago, in recognition of this problem, web browsers included support for tree-like folder structures for organizing bookmarks. These enable the user to browse through their repository to find the necessary information. However manual URL classification and organization can be difficult and tedious when there are more than a few dozen bookmarks to classify—something that goes against the whole grain of the bookmarking concept.

Many usability studies [13] indicate that a deep hierarchy results in less efficient information retrieval as many traversal steps are required, so users are more likely to make mistakes. Users also do not have the time and patience to arrange their collection into a well-ordered hierarchy. The present work can help extract information from this relatively information-poor domain in order to classify bookmarks automatically.

In order to retain as much information as possible, all fields residing within the bookmark database are considered. For every bookmark, the Uniform Resource Locater (URL) is divided into its slash-separated parts, with each part regarded as a keyword. In a similar fashion, the bookmark title is split into terms with stop words removed.

For this domain, the fuzzy RSAR approach produces exactly the same results as the crisp approach as all equivalence classes are crisp. In particular, as shown in Table 1, it can be seen that using the present work, the amount of attributes was reduced to around 35% of the original. This demonstrates the large amount of redundancy present in the original data sets. In light of the fact that bookmarks contain very little useful information, the results are a little better than anticipated.

It is also interesting to compare how the use of reduced data sets may affect the classification accuracy as compared to that of the unreduced data set. The results of this comparison are listed in Table 1 as well. Importantly, the performance of the reduced data set is almost as good as the original. Although a very small amount of important information may have been lost in the

Table 1
Classification accuracies using unreduced and reduced data sets

	Attributes		Average accuracy	
Data set	URL	Title	BIM	VSM
Unreduced	1397	1283	98.5%	98.7%
RS-reduced	514	424	94.3%	98.1%

attribute reduction, this information loss is not significant enough to reduce classification accuracy significantly, while the reduction of dimensionality is substantial.

4.3. Website classification

There are an estimated 1 billion web pages available on the WWW with around 1.5 million web pages being added every day. The task to find a particular web page, which satisfies a user's requirements by traversing hyper-links, is very difficult. To aid this process, many web directories have been developed—some rely on manual categorization whilst others make decisions automatically. However, as web page content is vast and dynamic, manual categorization is becoming increasingly impractical. Automatic web site categorization is therefore required to deal with these problems.

There is usually much more information contained in a web document than a bookmark. Additionally, information can be structured within a web page that may indicate a relatively higher or lower importance of the contained text. For example, terms appearing within a ⟨TITLE⟩ tag would be expected to be more informative than those appearing within the document body at large. Because of this, keywords are weighted not only according to their statistical occurrence but also to their location within the document itself. These weights are almost always real-valued, hence the need for the fuzzy approach.

The training and testing data sets were generated using Yahoo [25]. Five classification categories were used, namely Art & Humanity, Entertainment, Computers & Internet, Health, Business & Economy. A total of 280 web sites were collected from Yahoo categories and classified into these categories. From this collection of data, the keywords, weights and corresponding classifications were collated into a single data set.

For this set of experiments, crisp RSAR is compared with the new fuzzy RSAR approach. As the unreduced data set exhibits high dimensionality (2557 attributes), it is too large to evaluate (hence the need for keyword selection). Using crisp RSAR the original attribute set was reduced to 29 (1.13% of the full set of attributes). However, using fuzzy RSAR the number of selected attributes was only 23 (0.90% of the full attribute set). It is interesting to note that the crisp RSAR reduct and fuzzy RSAR reduct share four attributes in common. With such a large reduction in attributes, it must be shown that classification accuracy does not suffer in a fuzzy RSAR-reduced system.

To see the effect of dimensionality reduction on classification accuracy, the system was tested on the original training data first and the results are summarized in Table 2. The results are averaged over all the classification categories. Clearly, the fuzzy method exhibits better precision and error rates. This performance was achieved using 20.7% *fewer* attributes than the crisp approach.

Table 2
Performance: training data (using VSM)

Method	Attributes	Average precision	Average error
Original	2557	—	—
Crisp	29	73.7%	23.8%
Fuzzy	23	78.1%	16.7%

Table 3
Performance: unseen data

Method	% Original attributes	Classifier	Average precision	Average error
Crisp	1.13	BIM	23.3%	88.3%
		VSM	45.2%	49.7%
Fuzzy	0.90	BIM	16.7%	80.0%
		VSM	74.9%	35.9%

Table 3 contains the results for experimentation on 140 previously unseen web sites. For the crisp case, the average precision is rather low and average error high. With fuzzy RSAR there is a significant improvement in both the precision and classification error. Again, this more accurate performance is achieved while using 20.7% fewer attributes.

It must be pointed out here that although the testing accuracy is rather low, this is largely to do with the poor performance of the simple classifiers used. The fact that VSM-based results are much better than those using BIM-based classifiers shows that when a more accurate classification system is employed, the accuracy can be considerably improved with the involvement of the same attributes. Nevertheless, the purpose of the present experimental studies is to compare the performance of the two attribute reduction techniques, based on the common use of any given classifier. Thus, only the relative accuracies are important. Also, the fuzzy RSAR approach requires a reasonable fuzzification of the input data, whilst the fuzzy sets are herein generated by simple statistical analysis of the data set with no attempt made at optimizing these sets. A fine-tuned fuzzification will certainly improve the performance of fuzzy RSAR-based systems [14]. Finally, it is worth noting that the classifications were checked automatically. Many websites can be classified to more than one category, however, only the designated category is considered to be correct here.

5. Conclusion

This paper has presented a fuzzy–rough method for attribute reduction which alleviates important problems encountered by traditional RSAR such as dealing with noise and real-valued attributes. This novel approach has been applied to aid classification of web content, with very promising results. In particular, whilst retaining less attributes than the conventional crisp rough set-based technique, the work entails the classifiers that employ the retained attributes to have a higher classification rate. In all experimental studies there has been no attempt to optimize the fuzzifications or the classifiers

employed. It can be expected that the results obtained with optimization would be even better than those already observed.

There are several ways of improving the original QuickReduct algorithm. Many modifications have since been added and are reported in [3]. Similar optimizations may be included in the extended fuzzy–rough approach, yielding significant performance improvements. Even though the method proposed here is partially restricted (e.g. non-minimal reducts, difficulty in handling nested partitions of objects), it is still highly useful in locating close-to-minimal reducts. Other developments include ReverseReduct where the strategy is backward elimination of attributes as opposed to the current forward selection process. Initially, all attributes appear in the reduct candidate; the least informative ones are incrementally removed until no further attribute can be eliminated without introducing inconsistencies. As both forward and backward methods perform well, it is thought that a combination of these within one algorithm would be effective. Any version of the QuickReduct algorithm can be readily extended to dealing with fuzzy–rough cases.

Although the present work is focussed on web categorization, the generality of this approach should enable it to be applied to other domains. Relevant in-house experiments (e.g. for physical systems monitoring) have empirically demonstrated this with the results to be reported elsewhere. In addition, 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 very grateful to David Robertson and the other members of the Advanced Knowledge Technologies [1] team at Edinburgh, and to Alexandros Valarakos and Alexios Chouchoulas for their support.

References

[1] Advanced Knowledge Technologies homepage: http://www.aktors.org/.
[2] T. Beaubouef, F.E. Petry, G. Arora, Information measures for rough and fuzzy sets and application to uncertainty in relational databases, in: S.K. Pal, A. Skowron (Eds.), Rough–Fuzzy Hybridization: A New Trend in Decision Making, Springer, Singapore, 1999.
[3] A. Chouchoulas, J. Halliwell, Q. Shen, On the implementation of rough set attribute reduction, Proc. 2002 UK Workshop on Computational Intelligence, 2002, pp. 18–23.
[4] M. Dash, H. Liu, Feature selection for classification, Intell. Data Anal. 1 (3) (1997) 131–156.
[5] P. Devijver, J. Kittler, Pattern Recognition: A Statistical Approach, Prentice-Hall, Englewood Cliffs, NJ, 1982.
[6] D. Dubois, H. Prade, Putting rough sets and fuzzy sets together, in: R. Slowinski (Ed.), Intelligent Decision Support, Kluwer Academic Publishers, Dordrecht, 1992, pp. 203–232.
[7] I. Düntsch, G. Gediga, H.S. Nguyen, Rough set data analysis in the KDD process, 1999.
[8] U. Höhle, Quotients with respect to similarity relations, Fuzzy Sets and Systems 27 (1988) 31–44.
[9] R. Jensen, Q. Shen, A rough set-aided system for sorting WWW bookmarks, in: N. Zhong et al. (Eds.), Web Intelligence: Research and Development, 2001, pp. 95–105.

[10] R. Jensen, Q. Shen, Fuzzy–rough sets for descriptive dimensionality reduction, Proc. 11th Internat. Conf. on Fuzzy Systems, 2002, pp. 29–34.
[11] K. Kira, L.A. Rendell, The feature selection problem: traditional methods and a new algorithm, in: Proc. Ninth National Conf. on Artificial Intelligence, Anaheim, CA, 1992, pp. 129–134.
[12] P. Langley, Selection of relevant features in machine learning, in: Proc. AAAI Fall Symp. on Relevance, New Orleans, LA, 1994, pp. 1–5.
[13] K. Larson, M. Czerwinski, Web page design: implications of memory, structure and scent for information retrieval, in: Proc. 1998 ACM SIGCHI Conf. on Human Factors in Computing Systems, Los Angeles, CA, April 1998, pp. 25–32.
[14] J.G. Marin-Blázquez, Q. Shen, From approximative to descriptive fuzzy classifiers, IEEE Trans. Fuzzy Systems 10 (4) (2002) 484–497.
[15] T. Mitchell, Machine Learning, McGraw-Hill, New York, 1997.
[16] S.K. Pal, A. Skowron (Eds.), Rough–Fuzzy Hybridization: A New Trend in Decision Making, Springer, Singapore, 1999.
[17] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishing, Dordrecht, 1991.
[18] W. Pedrycz, Shadowed sets: bridging fuzzy and rough sets, in: S.K. Pal, A. Skowron (Eds.), Rough–Fuzzy Hybridization: A New Trend in Decision Making, Springer, Berlin, 1999.
[19] G. Salton, E.A. Fox, H. Wu, Extended boolean information retrieval, Comm. ACM 26 (12) (1983) 1022–1036.
[20] G. Salton, A. Wong, C.S. Yang, A vector space model for automatic indexing, Comm. ACM 18 (11) (1975) 613–620.
[21] Q. Shen, A. Chouchoulas, A modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems, Eng. Appl. Artif. Intell. 13 (3) (2000) 263–278.
[22] R. Slowinski (Ed.), Intelligent Decision Support, Kluwer Academic Publishers, Dordrecht, 1992.
[23] H. Thiele, Fuzzy rough sets versus rough fuzzy sets—an interpretation and a comparative study using concepts of modal logics, Tech. Report no. CI-30/98, University of Dortmund, 1998.
[24] M. Wygralak, Rough sets and fuzzy sets—some remarks on interrelations, Fuzzy Sets and Systems 29 (1989) 241–243.
[25] Yahoo. www.yahoo.com.
[26] Y.Y. Yao, A comparative study of fuzzy sets and rough sets, Inform. Sci. 109 (1988) 21–47.
[27] L.A. Zadeh, Fuzzy sets, Inform. Control 8 (1965) 338–353.