Computers in Industry 97 (2018) 111–116

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

A novel feature selection method using fuzzy rough sets

Sheeja T.K.[a,*], A. Sunny Kuriakose[b]

[a] Department of Mathematics, T. M. Jacob Memorial Govt. College, Manimalakunnu, Kerala, India
[b] Federal Institute of Science and Technology, Angamaly, Kerala, India

ARTICLE INFO

Article history:
Received 22 May 2017
Received in revised form 6 November 2017
Accepted 31 January 2018

Keywords:
Information systems
Approximations
Rough set
Divergence measure
Fuzzy rough set
Feature selection

ABSTRACT

The fuzzy set theory and the rough set theory are two distinct but complementary theories that deal with uncertainty in data. The salient features of both the theories are encompassed in the domain of the fuzzy rough set theory so as to cope with the problems of vagueness and indiscernibility in real world data. This hybrid theory has been found to be a potential tool for data mining, particularly useful for feature selection. Most of the existing approaches to fuzzy rough sets are based on fuzzy relations. In this paper, a new definition for fuzzy rough sets in an information system based on the divergence measure of fuzzy sets is introduced. The properties of the fuzzy rough approximations are explored. Moreover, an algorithm for feature selection using the proposed approximations is presented and experimented using real data sets.

1. Introduction

One of the greatest challenges faced by the present information age is the extraction of useful knowledge from a vast amount of raw data. The data mining tools and techniques are of great help in finding and describing the structural patterns in data. The conventional data analysis techniques such as regression analysis, time series analysis, cluster analysis, stochastic models, etc. deal with extracting quantitative data characteristics [1]. They cannot handle qualitative or imperfect data or produce a qualitative description of the dependencies in the data. Fuzzy set theory [2] and rough set theory [3] are two formal mathematical tools which can address the problem of vagueness, imperfection or incompleteness in data. They handle two different aspects of uncertainty namely vagueness and indiscernibility. The successful applications of fuzzy sets and rough sets in data mining have lead to a hybrid theory namely fuzzy rough set theory which can manage both types of uncertainty.

Fuzzy rough set theory is found to be a very effective tool for feature selection [4]. The task of feature selection is to remove the irrelevant and redundant features from a data set and choose only those features that facilitate extraction of useful knowledge. Using fuzzy rough feature selection it is possible to reduce discrete or real valued noisy data without providing any additional information. There are many different approaches to fuzzy rough sets available in the literature. Most of the definitions are based on fuzzy relations [5–9]. In this paper the fuzzy rough lower and upper approximations of a fuzzy set in a fuzzy information system are defined using the divergence measure of fuzzy sets. Divergence measures are fuzzy measures that express the extent to which two fuzzy sets differ from each other. The properties of the proposed approximations are explored. Further, an algorithm for feature selection using the fuzzy positive region is presented and experimented with real data sets.

The rest of the paper is organized as follows: Section 2 gives a brief review of the existing fuzzy rough set models and applications to feature selection. The concept of divergence based fuzzy rough sets in a fuzzy information system are introduced in Section 3 and their properties are studied. Section 4 provides a feature selection technique using the fuzzy positive region in the proposed approach and the process is illustrated with an example. The results obtained on experimentation of the proposed method with real data sets are discussed in Section 5. The conclusion and future work are given in Section 6.

2. Related work

In this section, some preliminary definitions and a brief review of the different approaches to fuzzy rough set theory existing in the

[*] Corresponding author.
E-mail addresses: sheejatk@fisat.ac.in (S. T.K.), asunnyk@fisat.ac.in (A.S. Kuriakose).

https://doi.org/10.1016/j.compind.2018.01.014
0166-3615/© 2018 Elsevier B.V. All rights reserved.

literature are presented. The basic notions of fuzzy set theory and rough set theory as described in [10,12] respectively, are followed throughout this paper.

2.1. Divergence measure

Let $\mathcal{F}(U)$ be the family of all fuzzy sets on U. Then a function $\delta : \mathcal{F}(U) \times \mathcal{F}(U) \to R$ is a *divergence measure* [11] if and only if $\forall A, B \in \mathcal{F}(U)$

1. $\delta(A, B)=\delta(B, A)$
2. $\delta(A, A)=0$
3. $\max\{\delta(A\cup C, B\cup C), \delta(A\cap C, B\cap C)\}\leq\delta(A, B)$

2.2. Fuzzy rough sets

Fuzzy rough sets encapsulate the related but distinct concepts of vagueness and indiscernibility. A fuzzy rough set consists of a pair of fuzzy membership functions which correspond to the fuzzy lower and upper approximations of a fuzzy set in a fuzzy approximation space. A fuzzy approximation space is a pair (U, R), where U is a non-empty set of objects and R is a fuzzy equivalence relation. The first attempt to define fuzzy rough sets in a fuzzy approximation space was done by Nakamura [5]. He defined the lower and upper approximations of a fuzzy set A on U as the fuzzy sets on U given by

$$\mu_{{}_{R}(A)}(x) = \inf_{R(x,y) \geq \alpha} \mu_A(y) \tag{1}$$

$$\mu_{\bar{R}(A)}(x) = \sup_{R(x,y) \geq \alpha} \mu_A(y) \tag{2}$$

respectively. Dubois and Prade [6] defined fuzzy rough approximations as

$$\mu_{{}_{R}(A)}(x) = \inf_{y \in U} \max(1 - R(x,y), \mu_A(y)) \tag{3}$$

$$\mu_{\bar{R}(A)}(x) = \sup_{y \in U} \min(R(x,y), \mu_A(y)) \tag{4}$$

Since then, intensive studies have been conducted on fuzzy rough sets both in theoretical and application point of view and several extensions and generalizations are proposed [7–9]. A detailed study of the different approaches to fuzzy rough set was done by D'eer et al [14].

2.3. Feature selection using fuzzy rough sets

Since its inception, several researchers have attempted to apply fuzzy rough set theory to feature selection. Many of them used the dependency degree, which is a measure of how well an attribute set can discern between elements as a criteria for feature selection. Whichever definition of fuzzy rough set is used, the dependency degree is defined using the fuzzy cardinality of the fuzzy positive region. A number of papers were authored by Jensen and Shen [15–19] in which the development of a fuzzy rough quick reduct algorithm was outlined. The attribute set having the maximum dependency degree is selected. Another approach to fuzzy-rough feature selection is to use fuzzy entropy as a criteria for feature selection [20]. Algorithms based on discernibility matrix to compute the attribute reducts are also proposed by many authors [19,21]. Fuzzy boundary region based feature selection methods are also available in the literature [19,22].

3. Divergence based fuzzy rough sets

Let (U, C, D) be a fuzzy information system, where U is a non-empty set of objects, C is the set of conditional attributes each of which is a fuzzy set on U and D is the set of decision attributes. Each $x\in U$ can be associated with a fuzzy set on $P\subseteq C$, with membership function

$$\mu_x^P(a) = \begin{cases} a(a), & \text{if } a \in P, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

Unless there is no ambiguity, the above fuzzy set may be denoted by μ_x.

Let $\delta(A, B)$ be a normalized divergence measure of fuzzy sets. Then, $\delta(\mu_x, \mu_y)$ measure the extent to which the fuzzy set μ_x on P differ from the fuzzy set μ_y on P. In other words, $\delta(\mu_x, \mu_y)$ expresses the dissimilarity between the objects x and y, with respect to the given set of fuzzy conditional attributes, whereas $R(x, y)$ expresses the indiscernibility between x and y. Thus a natural procedure to define divergence based fuzzy rough approximations is to replace $R(x, y)$ by $1-\delta(\mu_x, \mu_y)$.

Definition 3.1. The fuzzy rough lower and upper approximations of a fuzzy set A on U with respect to the divergence measure δ are defined $\forall x\in U$ as

$$\mu_{{}_{\delta}(A)}(x) = \inf_{y \in U} \max[\delta(\mu_x, \mu_y), \mu_A(y)] \tag{6}$$

$$\mu_{\bar{\delta}(A)}(x) = \sup_{y \in U} \min[1 - \delta(\mu_x, \mu_y), \mu_A(y)] \tag{7}$$

respectively.

The following proposition asserts that the above defined approximations are fuzzy sets on U.

Proposition 3.1. *The divergence based fuzzy rough lower and upper approximations of a fuzzy set on an information system are fuzzy sets on U.*

Proof. Both $\mu_A(y)$ and $\delta(\mu_x, \mu_y)\in[0, 1]$, $\forall x, y\in U$. Hence, $\max[\delta(\mu_x, \mu_y), \mu_A(y)]\in[0, 1]$. Using Eq. (6), $\mu_{{}_{\delta}(A)}(x) \in [0, 1]$, $\forall x \in U$. Similarly $\mu_{\bar{\delta}(A)}(x) \in [0, 1]$, $\forall x \in U$.

The properties of the proposed approximations are expressed in the next two theorems.

Theorem 3.1. *The general properties of the fuzzy rough lower and upper approximations with respect to δ are as follows:*

1. ${}_{\delta}(\phi) = \phi = \bar{\delta}(\phi)$
2. ${}_{\delta}(U) = U = \bar{\delta}(U)$
3. ${}_{\delta}(A) \subseteq A \subseteq \bar{\delta}(A)$, $\forall A \in \mathcal{F}(U)$
4. $A \subseteq B \Rightarrow {}_{\delta}(A) \subseteq {}_{\delta}(B)$ *and* $\bar{\delta}(A) \subseteq \bar{\delta}(B)$
5. ${}_{-}\delta(\hat{\alpha}) = \hat{\alpha} = \bar{\delta}(\hat{\alpha})$, $\forall\alpha\in[0, 1]$
6. $({}_{\delta}(A^C))^C = \bar{\delta}(A)$, $\forall A \in \mathcal{F}(U)$
7. $(\bar{\delta}(A^C))^C = {}_{\delta}(A)$, $\forall A \in \mathcal{F}(U)$

Proof.

1. $\mu_\phi(x) = 0$, $\delta(\mu_x, \mu_x) = 0$, $\forall x \in U \Rightarrow \max[\delta(\mu_x, \mu_x), \mu_\phi(x)] = 0$
   $\Rightarrow \mu_{\underline{\delta}(\phi)}(x) = \inf_{y \in U}\max[\delta(\mu_x, \mu_y), \mu_\phi(y)] = 0.$
   Also, $\min[1 - \delta(\mu_x, \mu_x), \mu_\phi(x)] = 0$, $\forall x \in U$.
   So, $\mu_{\bar{\delta}(\phi)}(x) = \sup_{\underline{y} \in U}\min[\delta(\mu_x, \mu_y), \mu_\phi(y)] = 0.$
   Thus, $\underline{\delta}(\phi) = \phi = \bar{\delta}(\phi)$
2. $\mu_U(x) = 1$, $\forall x \in U \Rightarrow \max[\delta(\mu_x, \mu_y), \mu_U(y)] = 1$, $\forall y \in U$.
   Therefore, $\mu_{\underline{\delta}(U)}(x) = \inf_{y \in U}\max[\delta(\mu_x, \mu_y), \mu_U(y)] = 1.$
   Also, $\min[1 - \delta(\mu_x, \mu_y), \mu_U(y)] = 1 - \delta(\mu_x, \mu_y)$.

Hence, $\mu_{\overline{\delta}(U)}(x) = \sup_{y\in U}(1-\delta(\mu_x, \mu_y))$. Since $\delta(\mu_x, \mu_x) = 0$, $\mu_{\overline{\delta}(U)}(x) = 1$. Thus, $\underline{\delta}(U) = U = \overline{\delta}(U)$

3. Since $\delta(\mu_x, \mu_x) = 0$, $\max[\delta(\mu_x, \mu_x), \mu_A(x)] = \mu_A(x)$, $\forall x \in U$.
   Hence, $\mu_{\underline{\delta}(A)}(x)\le\mu_A(x)$, $\forall x \in U$.
   Also, $\min[1 - \delta(\mu_x, \mu_x), \mu_A(x)] = \mu_A(x)$.
   So, $\mu_{\overline{\delta}(A)}(x) \ge \mu_A(x)$, $\forall x \in U$.
   Thus, $\underline{\delta}(A) \subseteq A \subseteq \overline{\delta}(A)$, $\forall A \in \mathcal{F}(U)$.
4. By definition, if $A \subseteq B$, then $\mu_A(y) \le \mu_B(y)$, $\forall y \in U$.
   So, $\max[\delta(\mu_x, \mu_y), \mu_A(y)] \le \max[\delta(\mu_x, \mu_y), \mu_B(y)]$ and $\min[\delta(\mu_x, \mu_y), \mu_A(y)] \le \min[\delta(\mu_x, \mu_y), \mu_B(y)]$, $\forall x \in U$.
   It follows that $\mu_{\underline{\delta}(A)}(x))\le\mu_{\underline{\delta}(B)}(x)$ and $\mu_{\overline{\delta}(A)}(x)\le\mu_{\overline{\delta}(B)}(x)$.
   Thus $\underline{\delta}(A) \subseteq \underline{\delta}(B)$ and $\overline{\delta}(A) \subseteq \overline{\delta}(B)$
5. Using property (3), $\underline{\delta}(\hat{\alpha}) \subseteq \hat{\alpha}$. Also, $\mu_{\hat{\alpha}}(y) = \alpha$, $\forall y \in U$.
   Hence, $\max[\delta(\mu_x, \mu_y), \mu_{\hat{\alpha}}(y)] \ge \alpha$, $\forall y \in U$.
   So, $\mu_{\underline{\delta}(\hat{\alpha})}(x) = \inf_{y\in U}\max[\delta](\mu_x, \mu_y), \mu_{\hat{\alpha}}(y)] \ge \mu_{\hat{\alpha}}(x)$.
   Therefore, $\hat{\alpha}\subseteq \underline{\delta}(\hat{\alpha})$. Thus $\underline{\delta}(\hat{\alpha}) = \hat{\alpha} = \underline{\delta}(\hat{\alpha})$, $\forall\alpha \in [0,\ 1]$ The proofs of properties (6) and (7) are straight forward.

Theorem 3.2. *The algebraic properties of the fuzzy rough lower and upper approximations with respect to δ are given below*:

1. ${}_{\delta}(A\cap B) ={}_{\underline{\delta}}(A)\cap{}_{\delta}(B)$
2. $\bar{\delta}(A\cap B)\subseteq\bar{\delta}(A)\cap\bar{\delta}(B)$
3. ${}_{\delta}(A\cup B)\supseteq{}_{\delta}(A)\cup{}_{\delta}(B)$
4. $\bar{\delta}(A\cup B)=\bar{\delta}(A)\cup\bar{\delta}(B)$
5. $\bar{\delta}(A\cap\hat{\alpha})=\bar{\delta}(A)\cap\hat{\alpha}$
6. ${}_{\delta}(A\cup\hat{\alpha}) ={}_{\delta}(A)\cup\hat{\alpha}$

Proof.

1.
$$\begin{aligned}\mu_{\underline{\delta}(A\cap B)}(x) &= \inf_{y\in U}\max[\delta(\mu_x,\mu_y),\mu_{(A\cap B)}(y)]\\ &= \inf_{y\in U}\max[\delta(\mu_x,\mu_y),\min(\mu_A(y),\mu_B(y)]\\ &= \inf_{y\in U}\min\{\max[\delta(\mu_x,\mu_y),\mu_A(y)],\max\\ &\quad[\delta(\mu_x,\mu_y),\mu_B(y)]\}\\ &= \min\{\inf_{y\in U}\max[\delta(\mu_x,\mu_y),\\ &\quad\mu_A(y)],\inf_{y\in U}\max[\delta(\mu_x,\mu_y),\mu_B(y)]\}\\ &= \mu_{\underline{\delta}(A)\cap\underline{\delta}(B)}(x),\quad\forall x\in U.\end{aligned}$$

Therefore, $\underline{\delta}(A\cap B) = \underline{\delta}(A)\cap\underline{\delta}(B)$

2.
$$\begin{aligned}\mu_{\overline{\delta}(A\cap B)}(x) &= \sup_{y\in U}\min[1-\delta(\mu_x,\mu_y),\mu_{(A\cap B)}(y)]\\ &= \sup_{y\in U}\min[1-\delta(\mu_x,\mu_y),\min(\mu_A(y),\mu_B(y)]\\ &= \sup_{y\in U}\min\{\min[1-\delta(\mu_x,\mu_y),\\ &\quad\mu_A(y)],\min[1-\delta(\mu_x,\mu_y),\mu_B(y)]\}\le\min\\ &\quad\{\sup_{y\in U}\min[1-\delta(\mu_x,\mu_y),\mu_A(y)],\\ &\quad\sup_{y\in U}\min[1-\delta(\mu_x,\mu_y),\mu_B(y)]\}\\ &= \mu_{\overline{\delta}(A)\cap\overline{\delta}(B)}(x),\quad\forall x\in U.\end{aligned}$$

Therefore, $\overline{\delta}(A\cap B)\subseteq\overline{\delta}(A)\cap\overline{\delta}(B)$

3.
$$\begin{aligned}\mu_{\underline{\delta}(A\cup B)}(x) = &\inf_{y\in U}\max[\delta(\mu_x,\mu_y),\mu_{(A\cup B)}(y)]\\ &= \inf_{y\in U}\max[\delta(\mu_x,\mu_y),\max(\mu_A(y),\mu_B(y)]\\ &= \inf_{y\in U}\max\{\max[\delta(\mu_x,\mu_y),\mu_A(y)],\\ &\max[\delta(\mu_x,\mu_y),\mu_B(y)]\}\\ &\ge\max\{\inf_{y\in U}\max[\delta(\mu_x,\mu_y),\mu_A(y)],\\ &\inf_{y\in U}\max[\delta(\mu_x,\mu_y),\mu_B(y)]\}\\ &= \mu_{\underline{\delta}(A)\cup\underline{\delta}(B)}(x),\quad\forall x\in U.\end{aligned}$$

Therefore, $\underline{\delta}(A\cup B)\supseteq\underline{\delta}(A)\cup\underline{\delta}(B)$.

4.
$$\begin{aligned}\mu_{\overline{\delta}(A\cup B)}(x) = &\sup_{y\in U}\min[1-\delta(\mu_x,\mu_y),\mu_{(A\cup B)}(y)]\\ &= \sup_{y\in U}\min[1-\delta(\mu_x,\mu_y),\max(\mu_A(y),\mu_B(y)]\\ &= \sup_{y\in U}\max\{\min[1-\delta(\mu_x,\mu_y),\mu_A(y)],\\ &\min[1-\delta(\mu_x,\mu_y),\mu_B(y)]\}\\ &= \max\{\sup_{y\in U}\min[1-\delta(\mu_x,\mu_y),\mu_A(y)],\\ &\sup_{y\in U}\min[1-\delta(\mu_x,\mu_y),\mu_B(y)]\}\\ &= \mu_{\overline{\delta}(A)\cup\overline{\delta}(B)}(x),\quad\forall x\in U.\end{aligned}$$

Therefore, $\overline{\delta}(A\cup B) = \overline{\delta}(A)\cup\overline{\delta}(B)$

5.
$$\begin{aligned}\mu_{\overline{\delta}(A\cap\hat{\alpha})}(x) &= \sup_{y\in U}\min[1-\delta(\mu_x,\mu_y),\mu_{(A\cap\hat{\alpha})}(y)]\\ &= \sup_{y\in U}\min[1-\delta(\mu_x,\mu_y),\min(\mu_A(y),\alpha)]\\ &= \sup_{y\in U}\min\{\min[1-\delta(\mu_x,\mu_y),\mu_A(y)],\alpha\}\\ &= \min\{\sup_{y\in U}\min[1-\delta(\mu_x,\mu_y),\mu_A(y)],\alpha\}\\ &= \mu_{\overline{\delta}(A)\cap\hat{\alpha}}(x),\quad\forall x\in U.\end{aligned}$$

Therefore, $\overline{\delta}(A\cap\hat{\alpha}) = \overline{\delta}(A)\cap\hat{\alpha}$

6.
$$\begin{aligned}\mu_{\underline{\delta}(A\cup\hat{\alpha})}(x) &= \inf_{y\in U}\max[\delta(\mu_x,\mu_y),\mu_{(A\cup\hat{\alpha})}(y)]\\ &= \inf_{y\in U}\max[\delta(\mu_x,\mu_y),\max(\mu_A(y),\alpha]\\ &= \inf_{y\in U}\max\{\max[\delta(\mu_x,\mu_y),\mu_A(y)],\alpha]\}\\ &= \max\{\inf_{y\in U}\max[\delta(\mu_x,\mu_y),\mu_A(y)],\alpha)]\}\\ &= \mu_{\underline{\delta}(A)\cup\hat{\alpha}}(x),\quad\forall x\in U.\end{aligned}$$

Therefore, $\underline{\delta}(A\cup\hat{\alpha}) = \underline{\delta}(A)\cup\hat{\alpha}$

The following theorem discusses the monotonic property of the divergence based fuzzy rough approximations with respect the divergence measures.

Theorem 3.3. *If δ_1 and δ_2 are two divergence measures of fuzzy sets such that $\delta_1(A,B)\le\delta_2(A,B),\ \ \forall A,B\in\mathcal{F}(P)$, then ${}_{-}\delta_1(A)\le{}_{-}\delta_2(A)$ and $\bar{\delta_1}(A)\ge\bar{\delta_2}(A)$.*

Proof. Given that $\delta_1(A,B)\le\delta_2(A,B),\ \ \forall A,B\in\mathcal{F}(P)$.

So, $\delta_1(\mu_x, \mu_y)\le\delta_2(\mu_x, \mu_y)$, $\forall x, y\in U$.
Hence, $\max[\delta_1(\mu_x, \mu_y), \mu_A(y)]\le\max[\delta_2(\mu_x, \mu_y), \mu_A(y)]$, $\forall y\in U$.
By the property of infimum, ${}_{-}\delta_1(A)\le{}_{-}\delta_2(A)$.
Also, $1-\delta_1(\mu_x, \mu_y)\ge1-\delta_1(\mu_x, \mu_y)$.
Hence, $\min[1-\delta_1(\mu_x, \mu_y), \mu_A(y)]\ge\min[1-\delta_2(\mu_x, \mu_y), \mu_A(y)]$, $\forall y\in U$.
By the property of supremum, $\overset{1}{\delta}(A)\ge\overset{2}{\delta}(A)$.

4. Feature selection using fuzzy positive region

In this section, an application of divergence based fuzzy rough sets to feature selection is described. An algorithm using the fuzzy positive regions of the decision classes is presented.

Let (U, C, D) be a fuzzy information system, where $U=\{x_1, x_2, \ldots, x_n\}$, $C=\{a_1, a_2, \ldots, a_m\}$ and $D=\{d_1, d_2, \ldots, d_k\}$. If the conditional attributes are real valued functions on U, then they can be converted into fuzzy sets on U by applying the transformation $a_i^*(x) = \frac{x-a}{b-a}$ where, $a=\min_{x\in U}a_i(x)$ and $b=\max_{x\in U}a_i(x)$.

Definition 4.1. Let $\delta(A, B)$ be a normalized divergence measure of fuzzy sets. The *divergence matrix* of U with respect to $P\subseteq C$ is the $n\times n$ matrix $D_P = [\delta_{ij}]_{n\times n}$, whose entries are given by

$$\delta_{ij} = \delta(\mu_{x_i}, \mu_{x_j}), i, j = 1, 2, \ldots, n \tag{8}$$

The decision attributes d_i may be crisp or fuzzy. The membership functions of the decision classes of crisp attributes are the characteristic functions of the corresponding equivalence classes. The membership functions of the fuzzy decision attributes act as the membership functions of their decision classes.

Definition 4.2. The *fuzzy positive region* with respect to $P\subseteq C$ is the fuzzy set POS_P on U given by

$$\mu_{POS_P}(x) = \sup_{X\in U/D}\mu_{\underline{\delta_P}(X)}(x) \tag{9}$$

where $\mu_{\delta P(X)}(x)$ is given by "Eq. (6)".

Definition 4.3. The *degree of dependency* of D on $P \subseteq C$ is defined as

$$\gamma_P(D) = \frac{\sum_{x \in U} \mu_{POS_P}(x)}{|U|} \tag{10}$$

Algorithm 4.1. The algorithm for finding the dependency measure of D with respect to $P \subseteq C$.

```
1.  Input the decision table and $P \subseteq C$
2.  Find the decision classes $U/D=\{X_1, X_2, \ldots, X_r\}$
3.  Find the divergence matrix $D_P$
4.  For $l$=1, 2, ..., $r$, $i$=1, 2, ..., $n$, compute $\mu_{\delta P(X_l)}(xi) = \inf_{x_j \in U} \max[\delta_{ij}, X_l(y)]$
5.  Compute $\mu_{POS_P}(x_i) = \sup_l \mu_{\delta P(X_l)}(x)$ for each $x_i \in U$
6.  Compute $\gamma_P(D) = \frac{\sum_{x_i \in U} \mu_{POS_P}(x_i)}{|U|}$.
7.  Return $\gamma_P(D)$
```

Algorithm 4.2. The following is the algorithm to find the set of features to be selected for the decision table reduction.

```
1.Input the fuzzy decision system
2.Initialise $C \leftarrow \{a_1, a_2, \ldots, a_m\}$, $R=\emptyset$
3.For each $a_i \in C-R$, generate the divergence matrix with respect to
  $R \cup \{a_i\}$
4.Calculate the dependency degree $\gamma_{R \cup \{a_i\}}(D)$ for each $a_i \in C-R$
5.Find the attribute $a_k$ that makes $\gamma_{R \cup \{a_i\}}(D)$ the maximum.
6.When $\gamma_{R \cup \{a_i\}}(D) > \gamma_R(D)$, $C \leftarrow C-R$ and $R \leftarrow R \cup \{a_k\}$
7.Return $R$
```

4.1. Example

Consider the fuzzy decision system given in Table 1, which is a part of the ser Knowledge Modelling Data taken from UCI repository of databases [23]. All the conditional attributes are fuzzy sets. There are 5 fuzzy conditional attributes and one crisp decision attribute. The description of the attributes are

a_1 – STG (The degree of study time for goal object materials)
a_2 – SCG (The degree of repetition number of user for goal object materials)
a_3 – STR (The degree of study time of user for related objects with goal object)
a_4 – LPR (The exam performance of user for related objects with goal object)
a_5 – PEG (The exam performance of user for goal objects)
d– UNS (The knowledge level of user)

The attribute UNS is the decision attribute. The values of the attributes corresponding to 8 objects belonging to different classes are considered. Here the divergence measure [11]

$$\delta(A, B) = \sup_{a \in P} |A(a) - B(a)| \tag{11}$$

is used.

There are 4 decision classes namely X_1={1, 7}, X_2={2, 4, 8}, X_3={3, 5} and X_4={6} corresponding to the decision attribute values high, low, medium and very low respectively. Also,

$$\delta(\mu_x, \mu_y) = \sup_{a \in P} |\mu_x(a) - \mu_y(a)| \tag{12}$$

At first the dependency values of the single feature subsets are to be determined. The divergence matrix corresponding to the attribute 'a_1' is presented in Table 2.

The divergence based fuzzy rough lower approximations of the decision classes are computed using the formula, $\mu_{\delta(X_l)}(x_i) = \inf_{x_j \in U} \max[\delta_{ij}, \mu_A(y)]$. Then the fuzzy positive region, $\mu_{POS_D}(x_i) = \sup_l \mu_{\delta(X_l)}(x_i)$ is determined. The values are given in Table 3.

Therefore, the dependency degree of D on the attribute a_i is given by $\gamma_a(D) = \frac{\sum_i \mu_{POS_D}(x_i)}{n} = \frac{0.172}{8}$.

Similarly the dependency degree of D on attributes a_2, a_3, a_4 and a_5 are computed as $\gamma_{a_2}(D) = \frac{0.75}{8}$, $\gamma_{a_3}(D) = \frac{0.65}{8}$, $\gamma_{a_4}(D) = \frac{0.91}{8}$ and $\gamma_{a_5}(D) = \frac{1.02}{8}$ respectively. The highest value is $\gamma_{a_5}(D)$ and hence the attribute a_5 is selected.

Now the dependency degrees of each attributes in combination with the selected attribute a_5 are to be determined, and the pair with highest dependency value is to be selected. For this, first consider the pair {a_4, a_5}.

The divergence matrix corresponding to the attribute set {a_4, a_5} is computed as given in Table 4.

The values of the divergence based fuzzy rough lower approximations of each of the decision classes with respect to {a_4, a_5} and the positive region are given in Table 5.

Therefore, the dependency degree of D on the attribute set {a_4, a_5} is given by $\gamma_{\{a_4,a_5\}}(D) = \frac{\sum_i \mu_{POS_D}(x_i)}{n} = \frac{2.19}{8}$.

Similarly, the dependency degree of D on attribute sets are computed as $\gamma_{\{a_3,a_5\}}(D) = \frac{1.61}{8}$, $\gamma_{\{a_2,a_5\}}(D) = \frac{1.535}{8}$ and $\gamma_{\{a_1,a_5\}}(D) = \frac{1.186}{8}$. Thus the highest value is $\gamma_{\{a_4,a_5\}}(D)$ and hence the attribute set {a_4, a_5} is selected.

In a similar way, the different dependency degrees are calculated as $\gamma_{\{a_3,a_4,a_5\}}(D) = \frac{2.34}{8}$, $\gamma_{\{a_2,a_4,a_5\}}(D) = \frac{2.56}{8}$ and $\gamma_{\{a_1,a_4,a_5\}}(D) = \frac{2.19}{8}$. Thus the highest value is $\gamma_{\{a_2,a_4,a_5\}}(D)$ and hence the attribute set {a_2, a_4, a_5} is selected.

Now the dependency degree of sets in the next level are $\gamma_{\{a_2,a_3,a_4,a_5\}}(D) = \frac{2.77}{8}$ and $\gamma_{\{a_1,a_2,a_4,a_5\}}(D) = \frac{2.56}{8}$. Thus the highest value is $\gamma_{\{a_2,a_3,a_4,a_5\}}(D)$ and hence the attribute set {a_2, a_3, a_4, a_5} is selected. The dependency degree of the whole set C is

Table 1
Fuzzy information system.

Object	a_1	a_2	a_3	a_4	a_5	d
1	0.08	0.08	0.1	0.24	0.9	High
2	0.06	0.06	0.05	0.25	0.33	Low
3	0.1	0.1	0.15	0.65	0.3	Medium
4	0.08	0.08	0.08	0.98	0.24	Low
5	0.09	0.15	0.4	0.1	0.66	Medium
6	0.15	0.02	0.34	0.4	0.01	Very low
7	0.24	0.75	0.32	0.18	0.86	High
8	0.276	0.255	0.81	0.27	0.33	Low

Table 2
Divergence matrix w.r.t. {a_1}.

D	1	2	3	4	5	6	7	8
1	0	0.02	0.02	0	0.01	0.07	0.16	0.196
2	0.02	0	0.04	0.02	0.03	0.09	0.18	0.216
3	0.02	0.04	0	0.02	0.01	0.05	0.14	0.176
4	0	0.02	0.02	0	0.01	0.07	0.16	0.196
5	0.01	0.03	0.01	0.01	0	0.06	0.15	0.186
6	0.07	0.09	0.05	0.07	0.06	0	0.09	0.126
7	0.16	0.18	0.14	0.16	0.15	0.09	0	0.036
8	0.196	0.216	0.176	0.196	0.186	0.126	0.036	0

$\gamma_{\{a_1,a_2,a_3,a_4,a_5\}}(D) = \frac{2.77}{8}$. Hence there is no increase in dependency by adding a_1. Thus the feature set selected is $\{a_2, a_3, a_4, a_5\}$.

4.2. Algorithmic complexity

For a data set having n features, the dependency function is evaluated n times corresponding to each single attribute. After selecting the feature with highest dependency value, the process is repeated by considering pairs of the selected feature with the remaining $(n-1)$ features. In the worst case, this process is terminated when the whole feature set has been exhausted. Therefore, the maximum number of evaluations of the dependency function for a particular data set is $n+(n-1)+(n-2)+\cdots+1=(n^2+n)/2$. But in many cases the process will be terminated when a reduct having a smaller number of features with the same dependency value as that of the entire feature set is reached. Thus, the maximum time complexity of the proposed algorithm is $o(n^2)$. Also, at the initial stage, n divergence matrices are computed and stored corresponding to each individual features. The space for all the subsequent matrices and local variables can be reused. Thus, the space complexity of the proposed algorithm is $o(n)$.

Table 3
Fuzzy positive region w.r.t $\{a_1\}$.

δ	1	2	3	4	5	6	7	8
X_1	0	0	0	0	0	0	0.036	0
X_2	0	0.02	0	0	0	0	0	0.036
X_3	0	0	0.02	0	0.01	0	0	0
X_4	0	0	0	0	0	0.05	0	0
POS_{a_1}	0	0.02	0.02	0	0.01	0.05	0.036	0.036

Table 4
Divergence matrix w.r.t $\{a_4, a_5\}$.

D	1	2	3	4	5	6	7	8
1	0	0.57	0.6	0.74	0.24	0.89	0.06	0.57
2	0.57	0	0.4	0.73	0.33	0.32	0.53	0.02
3	0.6	0.4	0	0.33	0.55	0.29	0.56	0.38
4	0.74	0.73	0.33	0	0.88	0.58	0.8	0.71
5	0.24	0.33	0.55	0.88	0	0.65	0.2	0.33
6	0.89	0.32	0.295	0.58	0.65	0	0.85	0.32
7	0.06	0.53	0.56	0.8	0.2	0.85	0	0.053
8	0.57	0.02	0.38	0.71	0.33	0.32	0.53	0

Table 5
Positive region w.r.t $\{a_5, a_1\}$.

δ	1	2	3	4	5	6	7	8
X_1	0.24	0	0	0	0	0	0.2	0
X_2	0	0.32	0	0	0	0	0	0.32
X_3	0	0	0.29	0	0.2	0	0	0
X_4	0	0	0	0	0	0.29	0	0
$POS_{\{a_4,a_5\}}$	0.24	0.32	0.29	0.33	0.2	0.29	0.2	0.32

5. Experimentation

This section presents the results from the experimental study of the feature selection method using the divergence based fuzzy positive region. Eleven data sets from the UCI Machine Learning repository [24] and one from the website of Milano Chemometrics and QSAR Research Group [25] have been used for the experimentation. The data sets consist of real valued features ranging from 5 to 166 in number, objects ranging from 120 to 4898 and decision classes ranging from 2 to 34. The description of the data sets is given in Table 6.

The data pre-processing step includes conversion of the real valued features to fuzzy ones. The dependency values corresponding to each single attribute sets are computed first and the attribute with maximum dependency value is selected. Then, pairs of the selected feature with the remaining features are considered and the pair having the maximum value of dependency is selected. This process is repeated unless there is no further increase in the dependency value. In the worst case, the process is terminated when the whole feature set has been exhausted. The feature selection process is conducted using an OCTAVE program and the results are presented in Table 7.

It is clear from the data presented in Table 7, that the size of the feature set is reduced significantly in almost all the cases. The algorithm converges even for data sets consisting of around 5000 objects. The run time given includes time taken for the data pre-processing step also. However, the time taken to run the program increases considerably with increase in the number of objects and the number of features. Using effective optimization

Table 7
Experimental results.

Dataset	Objects	Features	Reduct	Dependency	Run time(s)
Olitos	120	26	17	0.240	19.677
Sonar – mines/rocks	208	61	31	0.341	113.647
Glass	214	10	7	0.098	16.878
Knowledge	258	6	5	0.129	11.872
Ionosphere	351	35	29	0.414	156.838
Musk	476	166	79	0.843	2465.641
Energy efficiency	769	10	4	0.374	154.942
Plant leaves	1600	65	40	0.032	16063.745
Steel plates faults	1941	28	15	0.206	3640.639
Segment	2310	20	11	0.363	3638.397
Statlog	4435	37	22	0.103	24838.111
Wine quality – white	4898	12	10	0.05	20956.464

Table 6
Dataset description.

Dataset	Objects	Features	Decision classes	Description
Olitos [26]	120	26	4	Chemical analysis
Sonar – mines/rocks	208	61	2	Mine/rock recognition
Glass	214	10	7	Glass identification
Knowledge [23]	258	6	4	Knowledge level classification
Ionosphere	351	35	2	Structure analysis
Musk	476	166	2	Musk/non musk classification
Energy efficiency [27]	768	10	2	Energy analysis
Plant leaves [28]	1600	65	34	Plant leaves identification
Steel plate faults	1941	28	7	Steel plates fault diagnosis
Segment	2310	20	7	Image segmentation
Statlog	4435	37	7	Landsat satellite data
Wine quality – white [29]	4898	12	7	Wine quality analysis

strategies and some programming techniques the run times may be reduced further.

6. Conclusion

The fuzzy rough set theory has been found to be a very effective tool for data mining, especially for feature selection. In this paper, the fuzzy rough lower and upper approximations of a fuzzy set in a fuzzy information system have been defined using the divergence measure of fuzzy sets. Divergence measures are fuzzy measures that express the extent to which two fuzzy sets differ from each other. The properties of the proposed approximations were examined. Further, an algorithm for feature selection using the fuzzy positive region has been presented. The proposed method was illustrated by taking a part of the ser Knowledge Modelling Data from UCI repository of databases. To support the performance of the feature selection method using the divergence based fuzzy positive region, an experimental study was conducted. The results have shown that the number of selected features in each case was considerably less than the number of features in the original data sets. Future work includes optimization of the algorithm so as to reduce the run times. Also, the classification accuracy for the reduced data sets needs to be studied. Further, a comparison of the method using different divergence measures is to be done.

References

[1] S. Sumathi, S.N. Sivanandam, Introduction to Data Mining and Its Applications, Springer, Berlin, 2006.

[2] L.A. Zadeh, Fuzzy sets, Inf. Control 8 (3) (1965) 338–353.

[3] Z. Pawlak, Rough set theory and its applications, J. Telecommun. Inf. Technol. 3 (2002) 341–356.

[4] S. Vluymans, L. D'eer, Y. Saeys, C. Cornelis, Applications of fuzzy rough set theory in machine learning: a survey, Fundam. Inform. (2015) 1–34.

[5] A. Nakamura, Fuzzy rough sets, Note Multiple-Valued Logic Jpn. 3 (1988) 1–8.

[6] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. Gen. Syst. 17 (1990) 191–209.

[7] A. Radzikowska, E.E. Kerre, A comparative study of fuzzy rough sets, Fuzzy Sets Syst. 126 (2002) 137–155.

[8] W.Z. Wu, J.S. Mi, W.X. Zhang, Generalized fuzzy rough sets, Inf. Sci. 151 (2003) 263–282.

[9] J.S. Mi, Y. Leung, T. Feng, Generalized fuzzy rough sets determined by a triangular norm, Inf. Sci. 178 (2008) 3203–3213.

[10] J. Klir, B. Yuan, Fuzzy Sets and Fuzzy Logic, Prentice Hall, New Jersey, 1995.

[11] S. Montes, I. Couso, P. Gil, C. Bertoluzza, Divergence measure between fuzzy sets, Int. J. Approx. Reason. 30 (2002) 91–105.

[12] Z. Pawlak, Rough Sets – Theoretical Aspect of Reasoning About Data, Kluwer Academic Publishers, The Netherlands, 1991.

[14] L. D'eer, N. Verbiest, C. Cornelis, L. Godo, A comprehensive study of implicator-conjunctor-based and noise-tolerant fuzzy rough sets: definitions, properties and robustness analysis, Fuzzy Sets Syst. 275 (2015) 1–38.

[15] R. Jensen, Q. Shen, Fuzzy-rough attribute reduction with application to web categorization, Fuzzy Sets Syst. 141 (3) (2004) 469–485.

[16] R. Jensen, Q. Shen, Semantics preserving dimensionality reduction: rough and fuzzy rough based approaches, IEEE Trans. Knowl. Data Eng. 16 (12) (2004) 1457–1471.

[17] R. Jensen, Q. Shen, Fuzzy-rough data reduction with ant colony optimization, Fuzzy Sets Syst. 149 (1) (2005) 5–20.

[18] R. Jensen, Q. Shen, Fuzzy-rough sets assisted feature selection, IEEE Trans. Fuzzy Syst. 15 (1) (2007) 73–89.

[19] R. Jensen, Q. Shen, New approaches to fuzzy-rough feature selection, IEEE Trans. Fuzzy Syst. 17 (4) (2009) 824–838.

[20] N.M. Parthalain, R. Jensen, Q. Shen, Fuzzy entropy assisted fuzzy-rough feature selection, in: IEEE International Conference on Fuzzy Systems Canada, July, (2006), pp. 1499–1506.

[21] C.C. Eric, D. Chen, D.S. Yeung, X.Z. Wang, W.T. John, Attributes reduction using fuzzy rough sets, IEEE Trans. Fuzzy Syst. 16 (5) (2008) 1130–1141.

[22] N.M. Parthalain, Q. Shen, R. Jensen, A distance measure approach to exploring the rough set boundary region for attribute reduction, IEEE Trans. Knowl. Data Eng. 2 (3) (2010) 305–317.

[23] H.T. Kahraman, S. Sagiroglu, I. Colak, Developing intuitive knowledge classifier and modeling of users' domain dependent data in web, Knowl. Based Syst. 37 (2013) 283–295.

[24] M. Lichman, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2013, http://archive.ics.uci.edu/ml.

[25] http://michem.disat.unimib.it/chm.

[26] C. Armanino, R. Leardi, S. Lanteri, G. Modi, Chemom. Intell. Lab. Syst. 5 (1989) 343–354.

[27] A. Tsanas, A. Xifara, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, Energy Build. 49 (2012) 560–567.

[28] C. Mallah, J. Cope, J. Orwell, Plant leaf classification using probabilistic integration of shape, texture and margin features, signal processing, Pattern Recognit. Appl. (2013).

[29] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, Modeling wine preferences by data mining from physicochemical properties, Decis. Support Syst. 47 (4) (2009) 547–553, Elsevier.