Volume 2, Issue 1, 2025
Intuitions & Insights
An Interdisciplinary Research Journal
ISSN: 3048-6793

Application of Fuzzy-Rough Set for Pre-Processing of Data

Pallab Kumar Dey

Department of Computer Science, Kalna College, Kalna-713409, India
E-mail: pallabkumardey@gmail.com

Abstract: For processing of health data Rough set is an important tool. It can handle uncertainty and impreciseness without any additional or prior information about data. Though, it cannot deal with continuous data perfectly. Fuzzy sets can handle vagueness and it can also deal with continuous data perfectly. By hybridization of Fuzzy set and Rough set new concept Fuzzy-Rough set has been introduced. It can deal with imprecision and uncertainty for discrete as well as continuous data more precisely. To handle uncertainty more perfectly, vagueness of Fuzzy set and Indiscernibility concept of Rough set are merged in Fuzzy-Rough sets. Crisp equivalence classes are the main concept of Rough set. Fuzzy equivalence classes are the major ideas of Fuzzy-Rough set. In this paper it has been shown that Fuzzy-Rough set is an emerging tool for pre processing of data in applications like attributes reduction, missing value imputation etc.

Keywords: Rough set, Fuzzy set, Fuzzy-Rough set, Attribute Reduction, Missing value Imputation.

***Corresponding author:** Pallab Kumar Dey

Received: 25.11.2024; **Accepted**: 29.01.2025; **Published**: 10.02.2025

1. Introduction

Collections of raw data have been increased rapidly in current era of E-technology but finding useful information from such data collections is a challenging issue since most of the data mining algorithm is based on ideal data. Most of the data modelling algorithms hamper due to missing attribute values, extraneous data and attributes. Inadequate treatments of these data seriously affect the data mining and classification accuracy. Fuzzy-Rough set approach is a useful technique to handle missing attribute values, extraneous data and attributes.

Rough set approach [1] can be used as pre-processing tool to handle missing values, extraneous data and attributes. The main feature of rough set is that it does not require any additional or prior information about data. But it cannot handle continuous data perfectly. To handle vagueness, Fuzzy sets are also important tool and it can deal with continuous data. So new concept Fuzzy Rough set [2] are introduced by hybridization of Fuzzy set and Rough set and it is more power full to deal with imprecision and uncertainty for discrete as well as continuous data. Indiscernibility concept of Rough set and vagueness of Fuzzy set are merged in Fuzzy-Rough sets to deal with uncertainty more precisely [3-4].

A directed fuzzy rough set model has been proposed to better captures of inherent uncertainty for distribution of sample comparing the traditional models [5]. It has been shown that a Fuzzy rough set is a valuable tool for feature selection. A fuzzy rough set model has been introduced which prevents samples misclassification. A greedy forward algorithm for feature selection has been design using fuzzy rough set [6]. It has been shown that fuzzy rough set model widely used in medical diagnosis for image processing, information fusion, and classification problems as it can handle continuous and uncertain information [7]. Overlap function-based variable precision fuzzy rough set (OVPFRS) has been proposed and used for practical tumor classification. In Rough set, crisp equivalence classes are the main concept. In hybridization process of Fuzzy-Rough set, Fuzzy equivalence classes are key concept. Fuzzy Rough set is an emerging tool for data modelling applications

Organization of the paper is as follows: Section 2 is devoted to overview of Rough set. Section 3 deals with Fuzzy-Rough set theory. In section 4 it has been shown that applications of Fuzzy-Rough set have been discussed.

2. OVERVIEW OF ROUGH SETS

Pawlak's rough set theory [1] is an important tool important to analyze inexact, uncertain and vague knowledge. The rough sets theory can be used to deal with vague and imprecise data. Objects with the same information are indiscernible considering available information.

Information system can be presented as four tuple (U,A,V,f) , here U is a non empty finite set of objects, A is a non empty finite set of attributes, $\forall a \in A$: Va is the domain of attribute a, V= $\cup$ Va is the domain of A, f is a mapping f: $U \times A \rightarrow V$, f(x, a) $\in$ Va is the value that x holds on a. Any subset B of A determines a binary relation I(B) called indiscernibility relation defined as:

$$I(B)=\{(x_i, x_j)\in U\times U \mid \forall a\in B, a(x_i)=a(x_j)\} \qquad (1)$$

If $(x_i, x_j) \in I(B)$, then x_i and x_j are indiscernible by attributes from B. The equivalence classes of indiscernibility relation B are denoted by $[x_i]_B$ and computed as:

$$[x_i]_B =\{ x_j \in U \mid (x_i, x_j)\in I(B)\} \qquad (2)$$

Let $X\subseteq U$, B-lower and B-upper approximation of a set can be defined as:

$$\underline{B}X=\{ x_i \mid [x_i]_B \subseteq X\} \qquad (3)$$

$$\overline{B}X=\{ x_i \mid [x_i]_B \cap X\neq \emptyset \} \qquad (4)$$

Order pair $(\underline{B}X, \overline{B}X)$ is called rough set of X. If P and Q be equivalence relations over U, then positive region can be defined as:

$$POS_P(Q)=\cup_{X\in U/Q}\ \underline{P}X \qquad (5)$$

Positive region contain all objects that can be classified to classes of U/Q using the information of P. If $P, Q \subseteq A$, then Q depends on P in a degree k ($0\leq k \leq 1$), determined by

$$k= \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} \qquad (6)$$

k=1 signify, total dependency of Q on P. 0<k<1 signify, partial dependency of Q on P. k=0 signify, no dependency of Q on P. If an attribute is removed from conditional attributes then by using change in dependency degree, significance of the attribute can be obtained. If an attribute $a \in P$, significance of the attribute a on Q is determined by

$$\sigma_P(Q,a)= \gamma_P(Q) - \gamma_{P-a}(Q) \qquad (7)$$

Reduction of dimension of data sets is an important application of Rough Set. If removing of some attributes does not affect classification of the dataset then these attributes are redundant. If attribute $a \in B$ and $I(B)=I(B-\{a\})$ then ‘a’ is dispensable, otherwise ‘a’ is indispensable. In a subset containing minimum number of attributes that maintain same partition as whole set of attributes is called Reduct. Set of all indispensable attributes of the Universe is called Core; also it may be defined as intersection of all reduct.

To compute reduct and core discernible matrix may also be used. Discernible matrix M(B) of B is a $n \times n$ matrix defined as :

$$C_{ij} = \{a \in B: a(x_i) \neq a(x_j)\} \quad \text{for } i, j = 1, 2, ..n. \tag{8}$$

C_{ij} is the set of all attributes that discern object xi and xj.

Core is the set of all single element entries of the discernible matrix M(B).

$$Core(B) = \{a \in B, C_{ij} = \{a\}, \text{for some } i, j\} \tag{9}$$

Discernibility function [5] f is a boolean function of m boolean variables $a_1^*, \ldots, a_m^*$ (corresponding to the attributes $a_1, \ldots, a_m$) computed from discernibility matrix M(B) , may defined as:

$$f(a_1^*, \ldots, a_m^*) = \wedge \{\vee\, c_{ij}^* | 1 \leq j \leq i \leq |U|, C_{ij} \neq \emptyset \} \tag{10}$$

where $c_{ij}^* = \{a^* | a \in C_{ij}\}$.

All minimal reduct of a system may be found by simplification of discernibility function, using boolean laws of boolean algebra. Core may also be found by intersections of all reduct.

Suppose a decision system with ‘A’ as the attributes set where ‘C’ is the conditional attribute set and ‘D’ is the decision attribute set with no intersection.

If $C' \subseteq C$ is a D-reduct of C then C' is minimal subset of C such that $\gamma(C,D)=\gamma(C',D)$.

Minimum subset with $\gamma_{c'}(D)=1$ is the minimal reduct. After computation of dependency for all possible subset of C, it may be possible to determine minimal reduct but for large data set it is not applicable.

If data sets have discrete attribute values then rough set may be used for attribute reduction using the concept of reduct. If data sets have real attribute values then it cannot be handled by rough set. This problem may be solved by discretization but information will be lost. Due to noise, same value may differ slightly but it will be treated as different value in rough set.

3. OVERVIEW OF FUZZY-ROUGH SETS

In current era, Fuzzy rough set is an emerging tool for data analysis. It is frequently used in pattern recognition and machine learning. Dubois and Prade [2] first proposed the concept of fuzzy rough set. Fuzzy rough set theory extended further [5-15] in terms of property and axioms. Fuzzy rough set has significant applications in attribute reduction for crisp as well as real valued attribute data sets. Vagueness of fuzzy set and Indiscernibility of Rough set have been merged in Fuzzy-Rough sets to deal with uncertainty more precisely.

Crisp equivalence classes are key concept of rough set. In similar way fuzzy equivalence classes are the key concepts of fuzzy-rough set [2-4]. A fuzzy binary relation of a fuzzy set R on a non empty set U, is called fuzzy equivalence relation if it satisfies

1) Reflexivity: $\mu_R(x,x)=1 \ \forall x \in U$

2) Symmetry: $\mu_R(x,y)=\mu_R(y,x) \ \forall x,y \in U$

3) Transitivity: $\mu_R(x,z) \geq sup_y \, min\{\mu_R(x,y), \mu_R(y,z)\} \ \forall x,y,z \in U$

From fuzzy equivalence relation, fuzzy equivalence class $\mu_{[X]_R}$ for objects close to x may be defined as:

$$\mu_{[X]_R}(y) = \mu_R(\mathrm{x},\mathrm{y}), \ \mathrm{y} \in \mathrm{U} \qquad (11)$$

The fuzzy P-lower and P upper approximation are defined as:

$$\mu_{\underline{P}X}(F_i) = \inf_{\mathrm{X}} \max\{1-\mu_{F_i}(\mathrm{x}), \mu_X(\mathrm{x})\} \ \forall \mathrm{i} \qquad (12)$$

$$\mu_{\overline{P}X}(F_i) = \sup_{\mathrm{X}} \min\{\mu_{F_i}(\mathrm{x}), \mu_X(\mathrm{x})\} \ \forall \mathrm{i} \qquad (13)$$

Where F_i denotes a fuzzy equivalence class belonging to U/P. According to above definition crisp upper and lower approximations are same as individual object's membership to the approximations are not explicitly available. So fuzzy lower and upper approximations are herein redefined [16] as:

$$\mu_{\underline{P}X}(X) = \sup_{F\in U/P} \min\{\mu_F(x), \inf_{y\in U} \max\{1-\mu_F(y), \mu_X(y)\}\} \tag{14}$$

$$\mu_{\overline{P}X}(X) = \sup_{F\in U/P} \min\{\mu_F(x), \sup_{y\in U} \min\{\mu_F(y), \mu_X(y)\}\} \tag{15}$$

The tuple ⟨ $\underline{P}X$, $\overline{P}X$ ⟩ is called a fuzzy-rough set. If only crisp equivalence classes are present then these definitions degenerate to traditional rough set.

4. Applications of fuzzy-Rough SETS .

cases	temperature	headache	cough	Flu
X_1	'High'	'*'	'no'	'yes'
X_2	'veryhigh'	'yes'	'yes'	'yes'
X_3	'*'	'no'	'no'	'no'
X_4	'High'	'yes'	'yes'	'yes'
X_5	'High'	'yes'	*'	'no'
X_6	'Normal'	'*'	'*'	'no'
X_7	'Normal'	'*'	'yes'	'no'
X_8	'*'	'yes'	'yes'	'yes'
X_9	'veryhigh'	'*'	'*'	'yes'
X_{10}	'Normal'	'no'	'no'	'no'

Table 1: Incomplete Information Discrete value

Object	a	b	c	d
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
7	?	-0.3	-0.5	No
8	-0.4	?	-0.1	Yes
9	0.2	-0.3	?	Yes
10	-0.1	-0.4	?	No

Table 2: Incomplete Information with Real value

In real world application, databases may have extraneous attributes and objects. Irrelevant attributes or objects may degrade performance of machine learning and pattern recognition. Irrelevant attributes or objects can hamper the task of data mining by increasing time and misguiding information. Noisy information can be removed to improve overall performance. It is also expensive to analyze and store unwanted data. Redundant attributes can be removed by maintaining same classification efficiency.

Table 1 shows incomplete information which can be changed to complete information using Rough set or Fuzzy Rough set but Table 2 shows real valued data set for which Fuzzy Rough set approach is the better choice.

Fuzzy-rough dependency function and fuzzy-rough quick reduct algorithm for computing reduct first proposed by Jensen and Shen[13-15]. Using fuzzy positive region of a fuzzy set (where A= C ∪D), membership of an object X∈U may be defined as:

$$\mu_{POS_C(D)}(X) = \sup_{X \in U/D} \{\mu_{\underline{C}X}(x)\} \qquad (16)$$

and dependency function may be defined as:

$$\gamma'_p(Q) = \frac{|\mu_{POS_C(D)}(X)|}{|U|} = \frac{\sum_{X \in U} \mu_{POS_C(D)}(X)}{|U|} \qquad (17)$$

In crisp case, *U/P* contains set of objects grouped together that are indiscernible. In traditional rough set, if *R(⊆A)* is a reduct of attribute set *A* then *γ(A)* and *γ (R)* are identical and equal to 1 if data set is consistent. But in fuzzy-rough set it may not be as object may belongs to many fuzzy equivalence classes, so reduce total dependency. With this issue fuzzy-rough quick reduct algorithm [16] has been proposed. It starts with an empty set. It add one by one attribute that effect maximum increasing of fuzzy-rough dependency function *γ'*. Algorithm was tested and shown better performance with some real life data sets.

5. CONCLUSION

In this paper, Fuzzy-Rough set approach for pre-processing of data has been discussed. Fuzzy Rough set is the most effective technique to deal with all type of data sets. Attribute reduction and imputation of missing value has been discussed using fuzzy rough set approach. Fuzzy-Rough set can be effectively used for discrete as well as continuous data. Fuzzy-Rough set method can be used for different type of applications with any data sets. Real life data sets have to be pre processed to use most of the data mining algorithms so Fuzzy-Rough set based pre processing approach is the best choice.

Acknowledgment: The author is thankful to the editor and the referee for their valuable suggestions and comments.

References:

[1] Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning about Data, Vol.9,* Kluwer Academic Publishers, Dordrecht. https://doi.org/10.1007/978-94-011-3534-4

[2] Dubois, D., & Prade, H. (1990). "Rough fuzzy sets and fuzzy rough sets", International Journal of General Systems, 17(1), 191–209. https://doi.org/10.1080/03081079008935107

[3] Dey, P., K., & Mukhopadhyay, S. (2017). Core reduct based preprocessing approach to incomplete data, *International Journal of Intelligent Engineering and Systems,* 10(5), 19-28 . https://inass.org/2017/2017103103.pdf

[4] Dey, P., K., & Mukhopadhyay, S. (2017). Fuzzy-Rough set Approach to Attribute Reduction, *International Journal of Applied Engineering Research*, 12(16), 5637-5642. https://www.ripublication.com/ijaer17/ijaerv12n16_30.pdf

[5] Wang,C. Wang,C. An,S., Ding, W., & Qian,Y.(2025). Feature Selection and Classification Based on Directed Fuzzy Rough Sets, in *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 55(1),699-711, doi: 10.1109/TSMC.2024.3492337.

[6] Wang, C. , Qi, Y., Shao, M. Hu, Q.,Chen, D., Qian, Y., & Lin, Y. (2017). A Fitting Model for Feature Selection With Fuzzy Rough Sets. IEEE Transactions on Fuzzy Systems.25(4), 741 - 753. https://doi.org/10.1109/TFUZZ.2016.2574918

[7] Zhang, X., Ou,Q., Wang, J.(2024) Variable precision fuzzy rough sets based on overlap functions with application to tumor classification, *Information Sciences*, 666, 120451,ISSN 0020-0255, https://doi.org/10.1016/j.ins.2024.120451.

[8] Jensen, R., & Shen, Q. (2009). New Approach to Fuzzy- Rough Feature Selection, *IEEE Transactions on Fuzzy Syatems*, 17(4), 824-838. https://doi.org/10.1109/TFUZZ.2008.924209

[9] Morsi, N.N., & Yakout, M., M., (1998). Axiomatics for fuzzy-rough sets", *Fuzzy Sets and Systems*, 100, 327–342. https://doi.org/10.1016/S0165-0114(97)00104-8

[10] Wu, W., Mi J., & Zhang, W. (2004). Generalized fuzzy-rough sets, *Information Sciences*, 151, 263–282. https://doi.org/10.1016/S0020-0255(02)00379-1

[11] Wu, W., & Zhang, W. (2004). “Constructive and axiomatic approaches of fuzzy approximation operators”, *Information Sciences*, 159, 233–254. https://doi.org/10.1016/j.ins.2003.08.005

[12] Cock, M. D., Cornelis, C., & kerre, E., E., (2007). Fuzzy rough sets: the forgotten step, *IEEE Transactions on Fuzzy Systems*, 15(1),121–130. http://dx.doi.org/10.1109/TFUZZ.2006.889762

[13] Hong, T., P., Liou, Y.,L. & Wang, S.,L. (2009). Fuzzy rough sets with hierarchical quantitative attributes, *Int. J. Expert Systems with Applications*, 36(3), 6790–6799. https://doi.org/10.1016/j.eswa.2008.08.069

[14] Yeung, D.,S., Chen, D.,G., Tsang, E.,C.,C., Lee, J.,W.,T., & Wang, X., Z. (2005). On the generalization of fuzzy rough sets, *IEEE Trans*. *Fuzzy Systems*, 13(3), 343–361. https://doi.org/10.1109/TFUZZ.2004.841734

[15] Hu, Q., Zhang, L., An, S., Zhang, D., & Yu, D. (2012). On robust fuzzy rough set models, *IEEE Transactions on Fuzzy Systems*, 20(4), 636–651. https://doi.org/10.1109/TFUZZ.2011.2181180

[16] Jensen, R., Shen, Q. (2004). Fuzzy-rough attribute reduction with application to web categorization, *Fuzzy Sets and Systems, Elsevier*, 141, 469–485. https://doi.org/10.1016/S0165-0114(03)00021-6