JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.7, Special Issue.1, Jan 2019 E-ISSN: 2347-2693

Imputation of Missing Data using Fuzzy-Rough Hybridization

Pallab kumar Dey[1*]

[1]Department of Computer Science, Kalna College, Kalna, India

[*]*Corresponding Author: pallabkumardey@gmail.com*

Available online at: www.ijcseonline.org

Abstract— Missing data imputation has a significant impact in data mining task. Data mining algorithms cannot be executed effectively due to missing attribute values. Improper handle of missing values affects the data mining and classification accuracy. Imputation based preprocessing approach is very effective technique for handling missing value. In this paper most similar object used to impute missing value. For searching similar object core attributes have to give highest priority after that reduct attributes. In the proposed method to fill missing value concept of core and reduct attributes has been used. Rough set is most suitable to handle discrete data. Fuzzy set can handle continuous data in a better way. Hybridization methodology like fuzzy-rough set are more powerful to deal with imprecision and uncertainty for discrete as well as continuous data. Detail study has been given to impute missing value. Fuzzy rough set based fuzz- rough core reduct based (FRCRB) algorithm has been proposed for missing value imputation.

Keywords— Missing value, Imputation, Rough set, Fuzzy set and fuzzy-rough set, data analysis.

I. INTRODUCTION

With advancement of E-technology large amount of data is generated every day. But these data sets may have missing values. If data sets have missing values then data analysis becomes complicated. So we have to develop efficient techniques that are able to deal with missing value. So by imputing missing value, problems can be solved. There are many reasons for incompleteness of data like unavailable of data or due to time constraints and cost efficiency it is not possible to collect data. Most of the existing data mining algorithms based on complete data. So it is not possible to use these algorithms to those data sets. A pre-processing step to handle missing values is important to use already available data mining algorithms effectively.

In this paper Fuzzy-Rough set approach has been used to incomplete information as pre-processing tool to handle missing values. Rough set is the most important tool for handle uncertainty and impreciseness and it does not need any additional or prior information about data. For searching similar object core attributes have to give highest priority after that reduct attributes and other redundant attributes can be ignored. But Rough set cannot handle continuous data perfectly. Fuzzy sets are also important tool for handle vagueness and can deal with continuous data also. So fuzzy rough set are introduced with hybridization of rough set and fuzzy set. Fuzzy rough set are more power full to deal with imprecision and uncertainty for discrete as well as continuous data. Using Core-Reduct [1] concept, fuzzy rough set based fuzz- rough core reduct based (FRCRB) algorithm has been proposed for missing value imputation. For best result, imputations from similar object can be done. For similarity of objects core attributes have to give highest priority after that reduct attributes considering other attributes as redundant. This concept has been used. Proposed method can be used to impute missing data by most similar object for continuous as well as discrete data.

Section I contains the introduction of Rough set, Fuzzy set and incomplete data. Section II contain the related work of missing value imputation, Section III deals with fuzzy rough set theory related to our proposed method, Section IV explain proposed fuzzy rough set based FRCRB algorithm for missing value imputation.

II. RELATED WORK

By list-wise or pair-wise missing values may be deleted [2] but we lost resources. List-wise deletion can be used if data set is large and missing rate is low. In pair-wise deletion all available information has been considered, but complex due to computation of covariance matrix. Without using pre-processing methods rule may be generated or knowledge can be extracted from incomplete data sets [3]-[6]. Decision tree used to classify new records in C4.5 method [3]. Instance based learning algorithms or extended KNN classifier [6] also classifies new records directly. Modified LEM2 algorithm [4]-[5] may be used by computing block of the attributes with the objects of known values and inducing certain rules with original LEM2 method.

In pre-processing based imputation approach existing data mining algorithm may be used effectively. In Mean-mode imputation [7],[8] missing values are replaced by mean of all complete values of the attribute for numeric type or by mode of that attribute considering complete data for linguistic attributes. Mean mode can be used considering same class value. There are various statistical methods to impute missing value but due to application criteria, results accuracy and computation complexity cannot be used in all data sets.

k numbers of nearest neighbour are used to impute most similar instance in k nearest neighbour (KNN) imputation method [9] but high computational cost. In Hot deck imputation [10] each missing value is replaced from the similar case. 'Hot deck' and 'cold deck' methods have significant impact on variance estimations as these methods not always fill the same value and no extra value introduce which are not present. Better for large data set but produce poor result for smaller data set as sample variance increased. An iterative procedure Expectation maximization (EM) algorithm [11] used for computing maximum likelihood estimation of incomplete data. EM algorithm has very complex implementation. In Multiple imputation (MI) a model is used in first step of imputation process and process repeat n times. These n complete data sets are used for prediction of missing value using some method [12]-[13]. MI may restore error variance lost from regression based single imputation. For larger number of missing values MI may face difficulties and it also time consuming and not cost efficient.

Rough set theory is very popular to deal with missing values. Rough set concept of indiscernibility relation and discernibility matrix has been used to fill missing values [1], [14]. Core and reduct concept of rough set used to impute missing value [1] from most similar object. Rough set tolerance relation was proposed [15] and some rough set definition has been redefined for incomplete information.
Three approaches to missing attribute values are discussed [16]. Main ideas of these definitions are attribute value blocks. For computation of characteristic relations, characteristic sets, lower-upper approximations and rule induction attribute-value pair block are used in incomplete decision tables [17]. Rough set may deal with discrete data but cannot handle continuous or real value efficiently. Fuzzy-Rough nearest neighbour based [18] tool is also interesting to impute missing value.
Dubois and Prade [19] proposed fuzzy rough set. Its property and axioms has been extended [20],[21]. Fuzzy rough set is an emerging tool for crisp and real valued attribute data sets in attribute reduction and missing value imputation. In Fuzzy-Rough vagueness of fuzzy set and indiscernibility concept of rough set merged. It can deal uncertainty more precisely. Crisp equivalence classes are important idea of rough set, in the same way fuzzy equivalence classes are key concept of fuzzy-rough set [19]. Jensen and Shen[22]-[24] proposed fuzzy-rough quick reduct algorithm using Fuzzy-rough dependency function.

Fuzzy-rough quick reduct algorithm may not be convergent [25-26]. Fuzzy-rough based quick reduct algorithm on uncertainty degree has been proposed [27]. A variable precision fuzzy-rough set was proposed [28], [29] and used to handle noise of misclassification. Shannon's Information entropy used to find conditional and decisions attributes dependency in fuzzy rough sets[30], [31]. Fuzzy-rough set hybrid data reduction model based on for granular computing proposed [32]. Using the concept of feature significance dimensionality reduction method proposed [33] on fuzzy-rough set that simultaneously selects attributes and extracts features. Proper reduct cannot be fetched by these algorithms an over reduct or sub reduct are fetched to save running time. But this loss of information may not be acceptable in many field. To find proper reduct discernibility matrix based approach is most popular. Rough sets discernibility matrix extended to used in fuzzy rough sets [27], [34]-[37] as fuzzy discernibility matrix.

III. METHODOLOGY

Rough set based approach for imputation and attribute reduction can be used efficiently for discrete values. Fuzzy rough set based approach better suited for real valued data also.

Fuzzy equivalence class are fundamental concept fuzzy rough set as crisp equivalence class are fundamental concept of rough set [19]. Let R be a fuzzy binary relation on a non empty universe U. R is a fuzzy similarity relation if

R is

reflexive R(x,x)=1,

symmetric (R(x,y)= R(y,x))

and sup-min transitive

$$\mathrm{R(x,\,y)} \geq \sup_{z\in U} \min\{R(x,z), R(z,y)\} \quad \text{..Eq. (1)}$$

The fuzzy equivalence (similarity) class $[x]_R$ with $x \in U$ can be defined as:

$$[x]_R(y) = R(x,y), \forall y \in U \qquad \text{……… .Eq. (2)}$$

Dubois and prade first proposed fuzzy rough set [19]. Let us denote non empty universe as U and R asfuzzy binary relation on U and if F(U) denote fuzzy power set of U then (R*(F), R*(F)) is the fuzzy rough set on U, where Fuzzy lower R*(F) and fuzzy upper R*(F) approximation was defined for every $x \in U$ as:

$$\mathrm{R}_*(\mathrm{F})(\mathrm{x}) = \inf_{y \in U} \max\{1 - R(x,y), F(y)\} \text{ ..Eq. (3)}$$

$$\mathrm{R}^*(\mathrm{F})(\mathrm{x}) = \sup_{y \in U} \min\{R(x,y), F(y)\} \text{ ….Eq. (4)}$$

Let $U = \{x_1,.....,x_n\}$, R is the family of fuzzy equivalence relations on conditional attributes and D denotes decision attribute, then (U, R∩D) is called fuzzy decision system. If Sim(R) = ∩ {R: R ∈R} then Sim(R) is also a fuzzy equivalence relation.

Positive region of D with respect to R may be defined as :

PosSim(R) D= $\bigcup_k$ Sim(R)*(Di) ………….Eq. (5)

where i=1,..k and (U/D)={D1,D2,….Dk}.

If PosSim(R) D= PosSim(R-{R}) D then R is dispensible relative to D in R. If P is the minimal subset of R with PosSim(R) D= PosSim(P) D then P⊂ R is the reduct of R.

Let $U = \{x_1,....,x_n\}$, then MD(U,R) is the n× n discernibility matrix(c_{ij}) of (U, R∪D) such that

1). $c_{ij} = \{\mathrm{R} : 1\text{-}\,\mathrm{R}(\mathrm{x_i}, \mathrm{x_j}) \ge \lambda_{\mathrm{i}}\}$,

Where λ_i =Sim(R)$_*([x_i]_D)(x_i)$,

λ_j =Sim(R)$_*([x_j]_D)(x_j)$, and if $\lambda_j < \lambda_i$;

………….Eq. (6)

2) $c_{ij} = \phi, otherwise$

…………Eq. (7)

MD(U,R) may not be symmetric and $c_{ii} = \phi$.

If (U, R∪D) is a boolean function with m Boolean variable $\overline{R_1},...\overline{R_m}$ with corresponding fuzzy attributes $R_1,...R_m$, then $f_D(U,R)$ ia a discernibility function and is defined as follows:

$$f_D(U,R)\ \ (\overline{R_1},...\overline{R_m}) = \wedge\{\vee(c_{ij}) : c_{ij} \neq \phi\}$$

………. …Eq. (8)

where $\vee(c_{ij})$ is the disjunction of all variable $\overline{R}$ such that $R \in c_{ij}$.

Core D(R) represented by single element entries c_{ij} . After deletion of $c_{ij} \neq \phi$ and c_{ij} with non empty overlap with core we have to compute

$$f_D(U,R) = \wedge\{\vee(c_{ij})\} \qquad \text{……………Eq. (9)}$$

and all reduct as:

$$\text{Core}\ _{\mathrm{D}}(\mathrm{R}) \wedge f_D(U,R)\ . \qquad \text{……….Eq. (10)}$$

IV. Results and Discussion

All method cannot be applicable for all type of data. The proposed method is applicable for incomplete data set where some instances are complete.

Fuzzy rough based method has been proposed for attribute reduction and imputation of incomplete data. First incomplete data set will be transformed to fuzzy decision table with appropriate fuzzy membership functions. Then from this decision table complete data instances will be fetched to form a subset of decision table with complete data. From this complete decision table core and reduct attributes will be determined according to method attribute reduction using Fuzzy Rough Set for complete data discussed in section III.

This core and reduct attributes will be used to impute missing value in fuzzy decision table according to FRCRB algorithm discussed later.

Rough set based CRB algorithm [1] can be used only for discrete data or have to perform discritization method for continuous data which may introduce error. Proposed fuzzy rough set based approach can easily solve this problem. Fuzzy rough discernibility matrix based approaches have been used for computing core and other reduct. By discernibility matrix based approach, it is possible to compute reduct sets which can be used for imputing missing value. But no need to impute missing value for all attribute.

To reduce time complexity, missing values for reduct set are imputed.

Core may be treat as essential feature of a data set and other reduct attributes have also influence on data set but other extraneous attributes may misguide or have no importance in data mining or classification task. So other attributes have no importance. So core attributes have great importance to impute missing ('?') data with considering other reduct attributes.

In incomplete fuzzy decision table, if $x_i \in U$ then the missing set of fuzzy attributes w.r.t. object x_i may be defined as:

$$\mathrm{MFA_i} = \{\mathrm{j; a_j (x_i) = ?, j = 1,2..m\, and\, reduct\, attr}\}$$

.... ……...Eq. (11)

and missing fuzzy object set by:

$$\mathrm{MFO} = \{\mathrm{i; MFA_i} \neq \phi, \mathrm{i = 1,2..n}\}$$

………………..Eq. (12)

$(x_i, x_j) \in I(D)$, denotes objects x_i and x_j belongs to same decision class

For fuzzy reduct attribute am, significance relation $SF_m(i, j)$ can predict similarity of object x_i and x_j. Core attributes have been given maximum priority by significance value 3, after that priority have been given to other reduct attributes by value 1 to consider similarity. Again any missing reduct attribute values insignificance or negative influence has been consider by value -1 as its value may be unequal with comparing object. 3,1,-1 these values used as priority of attributes to impute missing data for prediction of most similar object.

Now fuzzy decision table with reduct attributes, priority significance relation $\mathrm{PF(i, j)}$ may be defined to predict most suitable object to fill missing attribute using $SF_m(i, j)$

$$\mathrm{PF(i, j)} = \begin{cases} 0 & \text{if } \mathrm{MA_i \subseteq MA_j \vee (x_i, x_j) \notin I(D)} \\ \sum_{\mathrm{a_m} \in A} \mathrm{SF_m\ (i, j)} & \text{otherwise} \end{cases}$$

……..……Eq. (13)

Where

$$SF_m(i, j) = \begin{cases} 3 & if a_m(x_i) = a_m(x_j) \neq '?' \wedge a_m core \\ 1 & if a_m(x_i) = a_m(x_j) \neq '?' \wedge a_m oth.reduct \\ -1 & fa_m(x_i) = a_m(x_j) = '?' \end{cases}$$

…………....Eq. (14)

If $j \in \max_k(PF(i, j))$ then it can be said x_i is most similar with x_j and attributes value of object x_j may be used to fill missing values of x_i .

According to above discussion fuzz- rough core reduct based (FRCRB) algorithm has been proposed as Algorithm 1.

Algorithm 1: FRCRB algorithm

```
Input :   Incomplete fuzzy decision table with core
          and other reduct attributes only.

Output:  Complete reduct set.

    Step 1. Compute MFAi and MFO ;
    Step 2.    Compute PF(i,j);
    Step 3.    for ( each object i ∈ MO )
    Step 4.      $if\,(\max_k(PF(i,k)) > 0)$
    Step5.       $j \leftarrow k;\ for\ \max_k(PF(i,k)) exist;$
    Step 6.  $\mathrm{for\ (each\ attribute\ m \in c)}$
    Step 7.  $\mathrm{v'(i, m)} = \begin{cases} \mathrm{v(j, m)} & \mathrm{if\ v(i, m) = ?} \\ \mathrm{v(i, m)} & \mathrm{if\ v(i, m) \neq ?} \end{cases}$
    Step 8.    end for step 6
    Step 9.      end if step 4
     Step 10.   end for step 3
    Step 11.    Stop
```

V. Conclusion and Future Scope

In this paper imputation using fuzzy rough set approach has been proposed. Proposed method can be used effectively for discrete as well as continuous data. Fuzzy discernibility matrix has been used for computation of core and other reduct attributes which are the main features of data sets. These main features are used in fuzzy incomplete decision table to select most similar object and it is used to impute missing values. If similar complete instances does not exist then proposed algorithm does not misguide, but keep missing value as missing. No extraneous data has been introduced, so there is no chance to generate misleading information. For small and medium size data this method produce better result. For large data set, core and reduct attributes computations have to change.

REFERENCES

[1] P. K. Dey and S. Mukhopadhyay, "*Core reduct based preprocessing approach to incomplete data*", International Journal of Intelligent Engineering and Systems, Vol. 10, No. 5, pp. 19-28, 2017.

[2] A. C. Acock, "*Working with Missing Values",* Journal of Marriage and Family, Vol.67, No.4, pp. 1012-1028, 2005.

[3] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[4] J. W. Grzymala-Busse and A. Y. Wang, "*Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values*", In: Proc. of Fifth international Workshop on Rough Sets and Soft Computing (RSSC'97), Third Joint Conference on Information Sciences (JCIS'97), pp.69-72, 1997.

[5] J. W. Grzymala-Busse, "*Data with missing attribute values: Generalization of idiscernibility relation and rule induction*", Transactions on Rough Sets, Lecture Notes in Computer Science Journal Subline, Springer-Verlag, vol. 1 , pp. 78–95, 2004.

[6] D. Aha, D. Kibler and M. Albert, "*Instance-based learning algorithms*", Machine Learning, Vol.6, pp. 37-66, 1991.

[7] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2001.

[8] P. K. Dey and S. Mukhopadhyay, "*Modified Deviation Approach to Deal with Missing Attribute values in Data Mining with Different percentage of Missing Values*", International Journal of Computer Applications, Vol.73, No. 5, pp. 1-6, 2013.

[9] J. V. Hulse and T. M. Khoshgoftaar, "*Incomplete-case nearest neighbor* imputation in software measurement data", Information *Sciences,* Vol. 259, No. 2, pp. 596-610, 2014.

[10] R. R. Andridge and R. J. A. Little, "*A review of hot deck imputation for survey non- response*", International Statistical Review, Vol.78, No.1, pp. 40-64, 2010.

[11] A. Dempster, N. Laired and D. Rubin, "*Maximum likelihood from incomplete data via the EM algorithm*", Journal of the Royal Statistical Society, Vol.39, No.1, pp. 1-38, 1977.

[12] D. B. Rubin, Multiple Imputation for Nonresponse in Surveys, Wiley, New York, 1987.

[13] J. L. Schafer, "*Multiple imputation: a primer*", Statistical Methods in Medical Research, Vol. 8, pp. 3–15, 1999.

[14] W. Zhou, W. Zhang and Y. Fu , "*An incomplete data analysis approach using rough set theory*", Intelligent Mechatronics and Automation, *IEEE*, pp.332-338, 2004.

[15] M. Kryszkiewicz, "*Rough set approach to incomplete information systems*", Information Sciences, Elsevier, Vol. 112, pp.39-49, 1998.

[16] J. W. Grzymala-Busse, "*Three Approaches to Missing Attribute Values— A Rough Set Perspective*", In: Proc. of the Workshop on Foundations of Data Mining, associated with the fourth IEEE International Conference on Data Mining, Brighton, UK, November 1–4, 2004.

[17] J. W. Grzymala-Busse and S. Siddhaye, "*Rough Set Approaches to Rule Induction from Incomplete Data*", In: Proceedings of the *IPMU'2004,* the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Perugia, Italy, July 4–9, vol. 2, 923–930, 2004.

[18] M. Amiri and R. Jensen, "*Missing data imputation using fuzzy-rough methods*", Neurocomputing, Elsevier, Vol.205, pp. 152-164, 2016.

[19] D. Dubois, H. Prade, "*Rough fuzzy sets and fuzzy rough sets*", International Journal of General Systems, Vol.17, No.1, pp. 191–209, 1990

[20] D.S. Yeung, D.G. Chen, E.C.C. Tsang, J.W.T. Lee and X.Z. Wang, "*On the generalization of fuzzy rough sets*", IEEE Trans. Fuzzy Systems, Vol.13, No.3, pp. 343–361, 2005.

[21] W. Wu and W. Zhang, "*Constructive and axiomatic approaches of fuzzy approximation operators*", Information Sciences, Vol.159, pp. 233–254, 2004.

[22] R. Jensen, Q. Shen, "*Fuzzy-rough attribute reduction with application to web categorization*", Fuzzy Sets and Systems, Elsevier, Vol.141, pp. 469–485, 2004.

[23] Q. Shen, R. Jensen, "*On robust fuzzy rough set models*", Pattern Recognition, Vol. 37(7), pp. 1351–1363, 2004.

[24] R. Jensen, Q. Shen, "*Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approaches*", IEEE Transactions on Knowledge and Data Engineering, Vol.16, pp. 1457-1471, 2004.

[25] R. B. Bhatt, M. Gopal, "*On fuzzy-rough sets approach to feature selection*", Pattern Recognition Letter, Vol.26, No.7, pp. 965–975, 2005.

[26] R.B. Bhatt, M. Gopal, "*On the compact computational domain of fuzzy rough sets*", Pattern Recognition Letter, Vol.26, No.11, pp. 1632–1640, 2005.

[27] R. Jensen, Q. Shen, "*New approaches to fuzzy-rough feature selection*", IEEE Transactions on Fuzzy Systems, Vol.17, No.4, pp. 824-838, 2009.

[28] A. Mieszkowicz-Rolka and L. Rolka, "*Variable precision fuzzy rough sets*", Transactions on Rough sets, Springer, Vol. LNCS-3100, pp. 144-160,2004.

[29] S. Zhao, E. C. C. Tsang, and D. Chen, "*The model of fuzzy variable precision rough sets*", IEEE Transactions on Fuzzy Systems, Vol.17, No2, pp. 451–467, 2009.

[30] Q. H. Hu, D. R. Yu, Z. X. Xie, and J. F. Liu, "*Fuzzy probabilistic approximation spaces and their information measures*", IEEE Transactions on Fuzzy Systems, Vol.14, No.2, pp. 191–201, 2006.

[31] Q. H. Hu, D. R. Yu and Z. X. Xie , "*Information-preserving hybrid data reduction based on fuzzy-rough techniques*", Pattern Recognition Letters, Vol.27, No.5, pp. 414–423, 2006.

[32] Q. Hu., Z. Xie and Daren Yu, "*Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation*", Pattern Recognition, Vol.40, pp. 3509 – 3521, 2007.

[33] P. Maji and P. Garai, "*Fuzzy Rough simultaneous Attribute selection and Feature Extraction Algorithm*", IEEE Transactions on Cybernetics, Vol.43, No.4, pp. 1166–1177, 2013.

[34] E. C. C. Tsang, D. G. Chen, D. S. Yeung and X. Z. Wang, "*Fuzzy probabilistic approximation spaces and their information measures*", IEEE Transactions on Fuzzy Systems, Vol.16, No.5, pp. 1130–1141, 2008.

[35] D. Chen, L. Zhang, S. Zhao, Q. Hu and P. Zhu, "*A noval algorithm for finding reducts with fuzzy rough sets*", IEEE Transactions on Fuzzy Systems, Vol.20, No.2, pp. 385–389, 2012.

[36] C. C. Nghia and N. L. Giang, "*Fuzzy Rough set based Attribute Reduction in Fuzzy Decision Tables*", International Journal of Computer Applicationss, Vol.132, No.4, pp. 32–37, 2015.

[37] E. C. C. Tsang, D. Chen, D. S. Young, X. Wang and J. Lee, "*Attribute Reduction Using Fuzzy Rough Sets*", IEEE Transactions on Fuzzy Systems, Vol.16, No.5, pp. 1130-1140, 2008.

Authors Profile

Pallab kumar Dey, PhD, MCA, is an Assistant professor of Department of Computer Science, Kalna College, Kalna West Bengal, India. He has teaching experience of 8 years. His major research interests are Artificially Intelligence, Data Mining, SoftComputing etc. He has many publications in different national and foreign journals and Conferences.