ROUGH SET FOR PRE-PROCESSING OF DATA

Pallab Kumar Dey

Department of Computer Science, Kalna College
Kalna-713409, India

Corresponding author E-mail: pallabkumardey@gmail.com

Abstract

Data mining algorithms may hamper due to missing attribute values. Classification accuracy of the data mining may hamper due to improper handling of missing values. For data processing Rough set is an important tool. It can handle uncertainty and impreciseness without any additional or prior information about data. Crisp equivalence classes are the main concept of Rough set. In this paper it has been shown that Rough set is an important tool for data pre processing in applications like imputation of missing values, attributes reduction, etc.

Keywords*: Rough set, Attribute Reduction, Missing value Imputation*

INTRODUCTION

By the use of E-technology raw data collection has been increased rapidly. Now to identify important information from such a huge data is a new challenge as maximum data mining algorithms are based on ideal data. Due to missing attribute values and extraneous information data mining algorithms may hamper. Improper handling of missing attribute values and extraneous information may fully change the data mining and classification accuracy. Rough set is an important technique for handling missing attribute values and extraneous information.

As a pre-processing tool, Rough set technique [1] can be used to handle missing attribute values and extraneous information. Main reason behind rough sets popularity is that it does not require any additional or prior information about data[2-8]. In Rough set, crisp equivalence classes are the main concept. Rough set is an important tool for applications of data modeling .

Organization of the paper is as follows: Section 2 is devoted to overview of Rough set. In section 3 applications of Rough set have been discussed.

OVERVIEW OF ROUGH SETS

Pawlak's rough set theory [1] has a significant importance for analyzing inexact, uncertain and vague knowledge. To deal with vague and imprecise data, rough sets theory can be used. With respect to available information objects that have the same information are indiscernible.

Information system (U,A,V,f) can be described as four tuple where non empty finite set of objects represented by U, non empty finite set of attributes denoted by A , $\forall$ a$\in$ A: Va is the domain of attribute a, the domain of A is V= $\cup$ Va and f is a mapping f: U×A$\rightarrow$V, f(x, a) $\in$ Va denotes value that x holds on a. Indiscernibility relation is a binary relation denoted by I(B) where B is any subset of A can be defined as:

$$I(B)=\{(x i, x j)\in U\times U| \forall a\in B, a(xi)=a(xj)\} \quad (1)$$

So, x_i and x_j are indiscernible by attributes from B then (x i, xj) $\in$ I(B). $[x_i]_B$ denotes equivalence classes of indiscernibility relation B and it can be computed as

$$[xi]B =\{ x j \in U|(x i, x j)\in I(B)\} \quad (2)$$

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

$$\overline{B} X=\{ x i |[xi]B \cap X \neq \emptyset \} \quad (3)$$

$$\underline{B} X=\{ x i |[xi]B \subseteq X\} \quad (4)$$

Rough set of X can be described as order pair of ($\underline{B}$ X, $\overline{B}$ X) .

Now P and Q are the equivalence relations on U. So positive region may be described as:

$$POSP(Q)=\cup X\in U/Q \ \underline{P} X \quad (5)$$

All objects that can be classified by the information of P to classes of U/Q belong to positive region. Q depends on P in a degree k (0$\leq$ k $\leq$ 1) where P, Q $\subseteq$ A, may be computed by

$$k= \gamma P(Q) = \frac{|POS_P(Q)|}{|U|} \quad (6)$$

Total dependency of Q on P can be said if k=1. Partial dependency of Q on P can be said if 0<k<1. No dependency of Q on P can be said if k=0. Significance of the attribute can be obtained using change in dependency degree when an attribute is removed from conditional attributes.

Significance of the attribute a on Q where $a \in P$ can be obtained by

$$\sigma P(Q,a) = \gamma P(Q) - \gamma P\text{-}a(Q) \quad (7)$$

An important application of Rough Set is the reduction of dimensionality of data sets. Attributes are redundant if removing of these attributes does not affect classification of the dataset.

Attribute 'a' is dispensable if $a \in B$ and I(B)=I(B-{a}) otherwise 'a' is indispensable.

Reduct is the minimal subset of attributes that maintain same partition as whole set of attributes. Core is the set of all indispensable attributes of the Universe and it can be described as intersection of all reduct.

Discernible matrix can be used to compute reduct and core. Discernible matrix M(B) of B is a $n \times n$ matrix and can be described as :

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

Object xi and xj are discern by the set of all attributes C_{ij} .

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

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

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

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

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

From discernibility function by simplification using boolean laws of boolean algebra, all minimal reduct of a system can be found. Intersections of all reducts denotes the Core.

Conditional attribute set ‘C’ and decision attribute set ‘D’ with no intersection form the decision system with attributes set ‘A’.

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

Minimum subset with $\gamma c'(D)=1$ denotes the minimal reduct. Minimal reduct can be fetched by calculating dependency of all probable subset of C but it is not appropriate for large data set.

When data set have discrete attribute values then number of attributes can be reduced using Rough set as a form of reduct[2-12]. When data set have continuous attribute values then also Rough set can be used to calculate Reduct but continuous values have to change to discrete values by discretization method.

APPLICATIONS OF ROUGH SETS

Now a model is required for imputation of missing attribute values based on rough set theory for incomplete information system. According to our previous discussion it is clear that core attributes are the main or primary feature of a data set and reduct attributes are secondary necessary feature of a data set. Other attributes may be discard as these attributes does not have any effect on data sets decision attribute. According to this concept, core attributes are important than others attribute and remaining reduct attributes need to be consider also for imputation of missing data.

Core Reduct Based (CRB) algorithm [3] can be used to impute missing attributes values by near similar object. So almost same object can be fetched based on core and reduct attributes values similarity. This concept has been used in CRB algorithm. CRB algorithm has been used to table 1.

cases	temperature	headache	cough	Flu
X1	'High'	'*'	'no'	'yes'
X2	'veryhigh'	'yes'	'yes'	'yes'
X3	'*'	'no'	'no'	'no'
X4	'High'	'yes'	'yes'	'yes'
X5	'High'	'yes'	*'	'no'
X6	'Normal'	'*'	'*'	'no'
X7	'Normal'	'*'	'yes'	'no'
X8	'*'	'yes'	'yes'	'yes'
X9	'veryhigh'	'*'	'*'	'yes'
X10	'Normal'	'no'	'no'	'no'

Table 1: Incomplete Information System

Result has been shown in table 2. It can be said from table 2 that CRB algorithm able to fill missing value if similar object values exist otherwise keeps it as missing without mislead the imputation process. CRB algorithm is reliable than other methods. CRB algorithm has almost perfect classification ability. Rough set based algorithms accuracy is better than other methods. So, Rough set based algorithm can be used as an imputation method.

cases	temperature	headache	cough	Flu
X1	'High'	'yes'	'no'	'yes'
X2	'veryhigh'	'yes'	'yes'	'yes'
X3	'Normal'	'no'	'no'	'no'
X4	'High'	'yes'	'yes'	'yes'
X5	'High'	'yes'	'*'	'no'
X6	'Normal'	'no'	'no'	'no'
X7	'Normal'	'no'	'yes'	'no'
X8	'veryhigh'	'yes'	'yes'	'yes'
X9	'veryhigh'	'yes'	'yes'	'yes'
X10	'Normal'	'no'	'no'	'no'

Table 2: Information System after Imputation

CONCLUSION

In this paper, Rough set approach for pre-processing of missing data has been discussed. Rough set is the most effective technique to deal with incomplete data set. Using Core and Reduct attributes, efficiency of filling missing data has been enhanced by considering most suitable object. Attribute reduction and imputation of missing value has been discussed using fuzzy rough set approach. For different type of applications with different data sets, Rough set method may be used. It is essential to pre processed the Real life data sets such that it can be used for most of the data mining algorithms. Rough set based pre processing approach may be the better way for pre processing of data like data imputation, data reduction etc. This work can be enhanced for the both applications like imputation and feature reduction method to achieve more suitable data for data mining in less time

REFERENCES:

1. Pawlak Z. Rough Sets: “Theoretical Aspects of Reasoning about Data”, Vol.9, Kluwer Academic Publishers, Dordrecht (1991).
2. Dubois D., Prade H.. “Rough fuzzy sets and fuzzy rough sets”, International Journal of General Systems, Vol.17, No.1, pp. 191–209(1990).
3. Dey P. K. and Mukhopadhyay S.. “Core reduct based preprocessing approach to incomplete data”, International Journal of Intelligent Engineering and Systems, Vol. 10, No. 5, pp. 19-28 (2017).
4. Dey P. K. and Mukhopadhyay S.. “Fuzzy-Rough set Approach to Attribute Reduction”, International Journal of Applied Engineering Research, Vol. 12, No. 16, pp. 5637-5642(2017).
5. 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).
6. Ismail A.R. et al., “Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare” Journal of Robotics and Control (JRC) Vol.3, No.2, , pp. 143–152, ISSN: 2715-5072 DOI: 10.18196/jrc.v3i2.13133 143(2022) .
7. Wu W., Zhang W.. “Constructive and axiomatic approaches of fuzzy approximation operators”, Information Sciences, Vol.159, pp. 233–254(2004).
8. Cock M. D., Cornelis C., kerre E. E.. “Fuzzy rough sets: the forgotten step”, IEEE Transactions on Fuzzy Systems, Vol.15, No.1, pp. 121–130(2007).
9. Hong T.P., Liou Y.L. and Wang S.L.. “Fuzzy rough sets with hierarchical quantitative attributes”, Int. J. Expert Systems with Applications, Vol.36, No.3, pp. 6790–6799(2009).

10. Yeung D.S., Chen D.G., Tsang E.C.C., Lee J.W.T. and Wang X.Z.. “On the generalization of fuzzy rough sets”, IEEE Trans. Fuzzy Systems, Vol.13, No.3, pp. 343–361(2005).
11. Hu Q., Zhang L., An S., Zhang D. and Yu D.. “On robust fuzzy rough set models”, IEEE Transactions on Fuzzy Systems, Vol.20, No.4, pp. 636–651(2012).
12. Jensen R., Shen Q.. “Fuzzy-rough attribute reduction with application to web categorization”, Fuzzy Sets and Systems, Elsevier, Vol.141, pp. 469–485(2004).
13. Charm, S.E. and Kurland, G.S. “Blood Rheology in Cardiovascular Fluid Dynamics”, Academic Press, London (1965)