INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Rough Set Theory and Its Applications

Dr. Jyoti
Assistant Professor (CE)
YMCA University of Science and Technology, Faridabad, Haryana, India

Abstract:
Similar to data mining, three major web mining operations include clustering, association rule mining, and sequential analysis. Typical clustering operations in web mining involve finding natural groupings of web resources or web users. Researchers have found and pointed at some important and fundamental differences between clustering in conventional applications and clustering in web mining. Moreover, due to variety of reasons inherent in web browsing and web logging, the likelihood of bad and incomplete data is higher. This is where Rough Set Theory can play a crucial role and researchers have been utilizing this in clustering the incomplete data and thus aiding in decision making. This paper aims at understanding the Rough Set Theory and its applications in web mining.

Key words*: Rough Set Theory, Clustering, Fuzzy Clustering, Rough Set and Fuzzy Hybridization*

1. Introduction

The basic objective of clustering is to group data or objects having the similar characteristics in the same cluster and having dissimilarity with other clusters. It has been used in data mining tasks such as unsupervised classification and data summation. It is also used in segmentation of large heterogeneous data sets into smaller homogeneous subsets which are easily managed, separately modelled and analysed [1]. The basic goal in cluster analysis is to discover natural groupings of objects [2]. There may be possibility of uncertainty in certain datasets. Rough set clustering plays a major role under such consitions.

Rough set theory was first introduced by Zdzislaw Pawlak in 1982 [3]. It is a formal approximation of a crisp set in terms of a pair of sets which gives the lower and upper approximation of the original set. Rough Set Clustering (RST) is an approach to aid decision making in the presence of uncertainty [4]. It classifies imprecise, uncertain or incomplete information expressed in terms of data acquired from experience. In RST, a set of all similar objects is called an elementary set, which makes a fundamental atom of knowledge [5]. Any union of elementary sets is called a crisp set and other sets are referred to as rough set. As a result of this definition, each rough set has a boundary-line elements. For example, some elements cannot be definitively classified as members of the set or its complement. In other words, when the available knowledge is employed, boundary-line cases cannot be properly classified. Therefore, rough sets can be considered as uncertain or imprecise. Upper and lower approximations are used to identify and utilize the context of each specific object and reveal relationships between objects. The upper approximation includes all objects that possibly belong to the concept while the lower approximation contains all objects that surely belong to the concept.

1.1. Nomenclature

A rough set, first described by Zdzisław I. Pawlak, is a formal approximation of a crisp set (i.e., conventional set) in terms of a pair of sets which give the lower and the upper approximation of the original set.

Formally, an information system is a pair A = ($\cup$, A) where $\cup$ is a non-empty, finite set of objects called the universe and A is a non-empty, finite set of attributes on $\cup$

With every attribute $a \in A$, a set V_a is associated such that a: $\cup \rightarrow V_a$. The set V_a is called the domain or value set of attribute a.

Indiscernibility is core concept of RST and is defined as equivalence between objects. Objects in the information system about which we have the same knowledge form an equivalence relation.

The equivalence relation has the following properties.

If a binary relation $R \subseteq X * X$

Which is reflexive (i.e. an object is in relation with itself xRx),

Symmetric (if xRy then yRx) and transitive (if xRy and yRz then xRz) is called an equivalence relation.)

Formally any set $B \subseteq A$ there is associated an equivalence relation called B-Indiscernibility relation defined as follows:
$IND_A(B) = \{(x, x') \in U^2 \mid \forall a \in B\ a(x) = a(x')\}$
If $(x, x') \in IND_A(B)$, then objects x and x' are indiscernible from each other by attributes from B.
Equivalence relations lead to the universe being divided into equivalence class partition and union of these sets make the universal set. Target set is generally supposed by the user.
Lower approximation is the union of all the equivalence classes which are contained by the target set. The lower approximation is the complete set of objects that can be positively (i.e., unambiguously) classified as belonging to target set X.
The P-upper approximation is the union of all equivalence classes which have non-empty intersection with the target set. It represents the negative region, containing the set of objects that can be definitely ruled out as members of the target set.

2. Application Areas

2.1. Data Preprocessing
The authors in [6] applied rough set theory to three preprocessing steps: discretization, Feature selection, and Instance selection. And in [7], this theory was used to reduce the attributes of the students in an e-learning system before clustering. [8] describes the application of the RST in feature selection problem. Authors of [9] uses RST to develop 'Interest sets' for various commodities , which characterize customer's interest and utilize an "Interest Map", which displays the interest tendencies of customers using the interest set of each customer, visually.[10] have used RST with heuristics for feature selection.

2.2. Clustering
In [11], a rough set based hierarchical clustering algorithm for categorical data was proposed. Another algorithm was designed to handle categorical data called Min-Min-Roughness (MMR) based on Rough Set theory. In [12], the authors designed an autonomous knowledge-oriented clustering algorithm to analyze datasets of different attribute types. A variation of the K-means clustering algorithm based on the properties of the rough sets was introduced in [13] where clusters were represented as intervals or rough sets. In [14], a rough approximation based clustering algorithm was introduced to cluster web transactions from web access logs in order to discover web page access patterns. In [15], authors have used RST to cluster web user sessions. Authors of [16] have taken a step ahead in clustering categorical data by finding the categorical similarity measure based on the Euclidian distance so as to better solve the problem of categorical data because of the non-numerical data nature. [17] proposes a new technique called maximum dependency attributes (MDA) for selecting clustering attribute. Their approach is based on rough set theory by taking into account the dependency of attributes of the database. They have analyzed and compared the performance of MDA technique with the bi-clustering, total roughness (TR) and min–min roughness (MMR) techniques based on four test cases. The results establish the better performance of the proposed approach.

2.3. Incomplete Datasets
To solve the problems associated with decision tables with missing attribute values the authors in [18] described a modified version of the rule induction algorithm LEM2 (Learning from Examples Module, version 2), based on the idea of attribute-value pair blocks. A new learning algorithm is introduced in [19], which can simultaneously derive rules from incomplete data sets and estimate the missing values in the dataset. In [20] an imputation technique based on rough set computations was explored to handle missing data attributes.

2.4. The Tolerance Rough Set Model
It differs from the Rough set model in the sense that the semantic relatedness between the documents is established using rough sets. However, instead of equivalence relations, tolerance relations are used. Also, for rough set, the data must be pre-classified. In [21], authors have used the fuzzy sets along with tolerance rough set model as a way of relating data in their semantics. TRSM has also been used to design the algorithms for clustering the hierarchical and non-hierarchical documents [22]. However, it leaves various areas for improvement as to investigate the parameters for TRSM, to incrementally update the tolerance classes of terms and document clusters when new documents are added to the collections.

2.5. Tolerance Rough Set Model For Journals
TRSM is used by the authors of [23] for approximations of subsets of journals. Topics are viewed better using overlapping classes, which can be generated by tolerance relations say I (where I is reflexive and symmetric) in a universe J instead of an equivalence relation (which is reflexive, symmetric and transitive) used in the original rough set model. [23, 24 and25] used k-means clustering algorithm for overlapping clusters.

2.6. Fuzzy Clustering
The main idea in fuzzy clustering is the non-unique partition of the data in a collection of clusters. Fuzzy Clustering is a process that generates the fuzzy membership of objects of various clusters and then using them to assign objects to one or more clusters. In this, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the

strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

- **Fuzzy C-Means Clustering [26]**
  In this, every point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster. Any point x has a set of coefficients giving the degree of being in the kth cluster $w_k(x)$. With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:
  $Ck=\sum_x W_k(x)^m x/\sum_x W_k(x)^m$.
  The degree of belonging, $w_k(x)$, is related inversely to the distance from x to the cluster center as calculated on the previous pass. It also depends on a parameter m that controls how much weight is given to the closest center.

2.7. Rough and Fuzzy Hybridization

There have been two main lines of thought in the hybridization of fuzzy and rough sets,

- The constructive approach and
- The axiomatic approach

A general framework for the study of fuzzy-rough sets from both of these viewpoints is presented in [27]. . For the constructive approach, generalized lower and upper approximations are defined based on fuzzy relations. Initially, these were fuzzy similarity/equivalence relations [28] but have since been extended to arbitrary fuzzy relations [27]. The axiomatic approach is primarily for the study of the mathematical properties of fuzzy-rough sets [29]. Here, various classes of fuzzy-rough approximation operators are characterized by different sets of axioms that guarantee the existence of types of fuzzy relations producing the same operators. In [30], the definition of fuzzy rough sets is given based on the algebraic approach to the rough sets proposed by Iwinski, where a rough set is defined as a pair of subsets from a sub-boolean algebra without reference to universe. The lower and upper bounds of Iwinski rough sets are then fuzzified.

Another approach that blurs the distinction between rough and fuzzy sets has been proposed in [31]. The research was fuelled by the concern that a purely numeric fuzzy set representation may be too precise; a concept is described exactly once its membership function has been defined. It seems as though excessive precision is required in order
to describe imprecise concepts. The solution proposed is termed a shadowed set, which does not use exact membership values but instead employs basic truth values and a zone of uncertainty (the unit interval). This can be thought of as an approximation of a fuzzy set or family of fuzzy sets where elements may belong with certainty (membership of 1), possibility (unit interval) or not at all (membership of 0). This can be seen to be analogous to the rough set definitions for the positive, boundary and negative regions.

3. Conclusion

This paper reviewed Rough Set Theory and the various application areas of rough set theory. This paper also talks about the hybridization of rough sets and the various other techniques of which majorly Fuzzy sets have been reviewed.

4. References

1. Z. Huang, Data Mining and Knowledge Discovery 2 (3) (1998) 283–304.
2. R. Johnson, W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, New York, 2002.
3. Zdzislaw Pawlak, Rough Sets- Theoretical Aspects of Reasoning About Data. Norwell: Kluwar Academic Publishers, 1992.
4. Z. Pawlak, Rough set approach to knowledge-based decision support, European Journal of Operational research 99 (1) (1997) 48-57.
5. Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11 (5) (1982) 341-356.
6. F. Coaquira and E. Acuna, “Applications of rough sets theory in data preprocessing for knowledge discovery,” in proceedings of the World Congress on Engineering and Computer Science WCECS 2007, San Francisco, USA, 2007.
7. L. Shuai-dong and C. Shi-hong, “Clustering of web learners based on rough set,” Wuhan University Journal of Natural Sciences, vol. 9, pp. 542–546, 2004.
8. L. Shuai-dong and C. Shi-hong, “Clustering of web learners based on rough set,” Wuhan University Journal of Natural Sciences, vol. 9, pp.542–546, 2004.
9. Akihiro Ogino, Toshikazu Kato, “A Modeling Method of Interest using Rough Set Theory and Its Application”, http://www.hm.indsys.chuo-u.ac.jp/wp-content/uploads/2011
10. Ning Zhong, Juzhen Dong, Setsuo Ohsuga, “Using Rough Sets with Heuristics for Feature Selection”, Journal of Intelligent Information Systems, Kluwer Academic Publishers, 16, 199–214, 2001.
11. D. Chen, D. wu Cui, C. xue Wang, and Z. rong Wang, “A rough set-based hierarchical clustering algorithm for categorical data,” International Journal of Information Technology, vol. 12, pp. 149–159, 2006.
12. C. Bean and C. Kambhampati2, “Autonomous clustering using rough set theory,” International Journal of Automation and Computing, vol. 5, pp. 90–102, 2008.

13. P. Lingras and C. West, "Interval set clustering of web users with rough k-means," J. Intell. Inf. Syst., vol. 23, no. 1, pp. 5–16, 2004
14. S.K. De, P.R. Krishna, "Clustering web transactions using rough approximation", Fuzzy sets and systems, Vol 148, no. 1, pp-131-138, 2004
15. Jyoti, A K Sharma, Amit Goel, "A novel approach for clustering web user sessions using RST", In the Proc. Of International Journal on Computer Science and Engineering, Vol.2(1), 56-61, 2009.
16. Duo Chen, Du-Wu Cui, Chao-Xue Wang, Zhu-Rong Wang, "A Rough Set-Based Hierarchical Clustering Algorithm for Categorical Data", In the Proc. Of International Journal of Information Technology, Vol.12, No.3, 2006.
17. Tutut Herawan , Mustafa Mat Deris, Jemal H. Abawajy, "A rough set approach for selecting clustering attribute", In the Proc. Of Knowledge Based Systems, Vol 23, Issue 3, pp- 220-231, April 2010
18. J. W. Grzymala-busse and S. Siddhaye, "Rough set approaches to rule induction from incomplete data," in Proceedings the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, 2004, pp. 923–930.
19. T.-P. Hong, L.-H. Tseng, and S.-L. Wang, "Learning rules from incomplete training examples by rough sets," Expert Systems with Applications, vol. 22, no. 4, pp. 285 – 293, 2002.
20. F. V. Nelwamondo and T. Marwala, "Rough sets computations to impute missing data," CoRR, vol. abs/0704.3635, 2007
21. Hsuan-Shih Lee, Pei-Di Shen, Wen-Li Chyr, Wei-Kuo Tseng, "Mining quantitative data based on tolerance rough set model, In the Proc. of Knowledge-Based Intelligent Information and Engineering Systems Lecture Notes in Computer Science Volume 3681, pp 359-364, 2005
22. Tu Bao Ho, Ngoc Binh Nguyen, "Nonhierarchical Document Clustering Based on a Tolerance Rough Set Model, in the Proc. of International journal of Intelligent systems, VOL. 17, 199–212 (2002)
23. T. B. Ho and N. B. Nguyen, "Nonhierarchical document clustering based on a tolerance rough set model," International Journal of Intelligent Systems, vol. 17, pp. 199–212, 2002.
24. T. B. Ho, S. Kawasaki, and N. B. Nguyen, "Documents clustering using tolerance rough set model and its application to information retrieval," pp. 181–196, 2003.
25. G. Peters, "Some refinements of rough k-means clustering," Pattern Recogn., vol. 39, no. 8, pp. 1481–1491, 2006.
26. http://en.wikipedia.org/wiki/Fuzzy_clustering
27. D.S. Yeung, D. Chen, E.C.C. Tsang, J.W.T. Lee, and W. Xizhao, "On the Generalization of Fuzzy Rough Sets," IEEE Transactions on Fuzzy Systems, vol. 13, no. 3, pp. 343–361, 2005.
28. D. Dubois and H. Prade, "Putting Rough sets and fuzzy sets together", Intelligent Decision Support, pp. 203–232, 1992.
29. W.Z. Wu and W.X. Zhang, "Constructive and axiomatic approaches of fuzzy approximation operators," Information Sciences, vol. 159, no.3-4, pp. 233–254, 2004.
30. S. Nanda and S. Majumdar, "Fuzzy rough sets," Fuzzy Sets and Systems, vol. 45, pp. 157–160, 1992.
31. W. Pedrycz, "Shadowed sets: bridging fuzzy and rough sets," in the Proc of Rough fuzzy hybridization, pp. 179–199,199