International Journal of Information Processing, 9(2), 36-46, 2015
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

# Ensemble of Soft Decision Trees Using Multiple Approximate Fuzzy-Rough Set Based Reducts

G Kishor Kumar[a], P Viswanath[b] and A Ananda Rao[c]

[a]Department of Information Technology, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, Andhra Pradesh, India. Contact: kishorgulla@yahoo.co.in

[b]Machine Learning Research Laboratory, Department of Computer Science and Engineering, Madanapalle Institute of Technology and Science, Madanapalle, Andhra Pradesh, India. Contact: viswanath.p@ieee.org

[c]Department of Computer Science and Engineering, JNTUA, Anantapuramu, Andhra Pradesh, India. Contact:akepogu@gmail.com

Fuzzy-Rough set theory is a potential approach to find a subset of features called "reduct" which is decision equivalent to the entire feature set. It is a generalized approach of classical rough set theory for feature selection and reduction. The paper, with an aim to achieve better classification accuracy, proposes to use several approximate reducts with several decision trees, where the final classification decision is derived through a consensus function. We also extended the existing heuristic to find the goodness of attributes in the decision tree construction based on rough sets. Based on the experimental study, it is shown that the proposed method achieves better classification accuracy than the existing ensemble method where each component was derived using a C4.5 decision tree on a cluster obtained by applying k-prototype clustering method. Since the existing method considers all attributes for each component and also it could not resolve the vagueness present in the data, we proposes this ensemble method.

**Keywords :** Decision Tree Classification, Ensemble Technique, Fuzzy-Rough Sets, Rough Sets.

## 1. INTRODUCTION

The Rough set theory was introduced by Pawlak [1], which is a mathematical approach to resolve the vagueness and uncertainty present in information [2]. The main idea of rough set theory is on an assumption of every instance or object is associated some information. The objects whose characteristics are defined as same, they are referred as indiscernible (or precise or similar) with respect to available information. The indiscernible objects in the set can be formed as a basic granule called *elementary set*. Since, available information has a granular structure, some objects can be framed as indiscernible whereas other objects can be vague which means these objects whose characteristics are not defined as same from available information. To resolve vagueness the concepts in rough set theory are *lower approximation* and *upper approximation*.

This concept works well if the data is a qualitative data, where each attribute can have limited number of distinct values. But if the data is quantitative, where attributes are continuous valued like length, age or speed *etc.*, then the indiscernibility of instances can be measured based on closeness of its values. By applying discretization [3][4] on continuous valued attributes they appear to grade discernibility between instances. On the other hand fuzzy rough sets [5] can also be applied with the use of fuzzy relations [6] to find indiscernibility between instances and also to derive fuzzy reduct. Fuzzy rough set theory is a generalized approach of rough set theory for feature selection and reduction. Research on fuzzy rough

Table 1
Accuracy of Data sets with various Threshold Distance

| *Dataset* | *Threshold Distance* | | | | |
|---|---|---|---|---|---|
| | *0.05* | *0.15* | *0.25* | *0.35* | *0.45* |
| *IRIS* | 0.931 | 0.973 | **0.991** | 0.974 | 0.946 |
| *KDDCUP* | 0.884 | 0.928 | 0.962 | **0.992** | 0.971 |
| *NAD1998* | 0.921 | 0.943 | 0.978 | **0.991** | 0.976 |
| *SPAM* | 0.913 | 0.932 | 0.969 | **0.982** | 0.968 |

over all classified instances labeled as *normal* by the method.

- Accuracy is the percentage of all instances labeled as either *normal* or any *attack* correctly classified by the method.

Figure 2, Figure 3, Figure 4 and Figure 5 illustrates the performance of the proposed method and the existing method over the data sets 1999 kddcup, Iris, NAD-1998 and Spam mail data sets respectively. It is shown that the performance of the proposed method is better than the existing method in terms of TPR, FPR, precision and accuracy over the data sets 1999 kddcup, NAD-1998 and Spam mail. But the performance of the proposed ensemble method is almost equal to the existing method in terms of FPR over the Iris data set, but the other measures such as TPR, precision and accuracy are favor to the proposed method than the existing method.

In this paper, multiple approximate fuzzy reducts are derived using a threshold distance $\tau$, to construct an ensemble of decision trees. Table 1 shows the accuracies obtained on the data sets IRIS, KDDCUP, NAD1998 and SPAM MAIL with various threshold distance, where the high accuracy is shown in bold for each data set.

## 8. CONCLUSIONS

In this paper, to make effective use of possible information of the given information system and with an aim to achieve better classification accuracy, we proposed "an ensemble of soft decision trees using multiple approximate fuzzy-rough set based reducts". And also an existing rough set based heuristic is extended to find the goodness of attributes in the decision tree construction on each reduct. Finally an ensemble method is derived for the final classification decision of a given query pattern through a consensus function. Experimental results over four standard data sets have shown favor to the proposed ensemble method than the existing method.

## REFERENCES


1. Z Pawlak. Rough Sets, *International Journal of Computer and Information Sciences*, 11(5):341–356, 1982.
2. Zdzislaw Pawlak. Rough Set Approach to Knowledge-based Decision Support, *European Journal of Operational Research,* pages 48–57, 1997.
3. H S Nguyen. Approximate Boolean Reasoning: Foundations and Applications in Data Mining, *Transactions on Rough Sets, Lecture Notes in Computer Science, Springer*, pages 334–506, 2006.
4. Z Pawlak, A Skowron. Rough Sets and Boolean Reasoning, *Information Sciences* 177:41–73, 2007.
5. D Dubois, H Prade. Rough Fuzzy Sets and Fuzzy Rough Sets, *International Journal of General Systes,* 17:191–209, 1990.
6. L A Zadeh. Fuzzy Sets, *Information and Control*, 8:338–353, 1965.
7. Z Pawlak. Rough Set Approach to Multi-Attribute Decision Analysis, *European Journal of Operational Research* , 72:443–459, 1994.
8. J W GrZymala-Busse and W Ziarko. Data Mining and Rough Set Theory, *Communications of ACM,* 43:108–109, 2000.
9. Z Pawlak, S K M Wang and W Ziakro. Rough Sets: Probabilistic Versus Deterministic Approach, *International Journal of Man-Machine Studies,* 299(1):81–95, 1988.
10. Jiawei Han and Micheline Kamber. Data Mining Concepts and Techniques, *Morgan Kaufmann*, 2000.
11. Richard O Duda, Peter E Hart and David G Stork. Pattern Classification, *John Wiley and Sons, A Wiley-interscience Publication*, 2000.

12. J R Quinlan. Introduction of Decision Trees, *Machine Learning,* 3:81–106, 1986.
13. J R Quinlan. C4.5: Programs for Machine Learning, *Morgan Kaufmann* 1993.
14. L Breiman, J H Friedman, R A Olshen and C J Stone. Classification and Regression Trees, *Wadsworth International Monterey*, 1984.
15. Sang Wook Han and Jae-Yearn Kim. Rough Set-based Decision Tree Algorithm using the Core Attributes Concept, *International Conference ICICIC 07*, pages 298–301, 2007.
16. Lin Zhou and Feng Jiang. A Rough Set Based Decision Tree Algorithm and Its Application in Intrusion Detection, *4th International Conference*, pages 333-338, 2011.
17. Xiangpeng Li and Min Dong. An Algorithm for Constructing Decision Tree based on Variable Precision Rough Set Model, *4th International Conference on Natural Computation*, pages 280–283, 2008.
18. Baoshi Ding, Yongqing Zheng and Shaoyu Zang. A New Decision Tree Algorithm Based on Rough Set Theory, *Asia-Pasific Conference on Information Processing*, pages 326–329, 2009.
19. Jin-Mao Wei. Rough Set based Approach to Selection of Node, *International Journal of Computational Cognition,* 1(2):25–40, 2003.
20. Padraig Cunningham. Ensemble Techniques, *Technical Report UCD-CSI-2007-5*, 2007.
21. Lior Rokach. Ensemble-based Classifiers, *Artificial Intelligence Review,* 33(1):1–39, 2010.
22. G Kishor Kumar, P Viswanath and A Ananda Rao. Intrusion Detection Using an Ensemble of Decision Trees, *in Proceedings of the 5th Indian International Conference on Artificial Intelligence(IICAI)*, pages 382–392, 2011.
23. Amir Ahmad and Lipika Dey. A $k$-mean Clustering Algorithm for Mixed Numeric and Categorical Data, *IEEE Transactions on Data and Knowledge Engineering*, 63:503–507, 2007.
24. Q Shen and A Chouchoulas. A Fuzzy-Rough Approach for Generating Classification Rules, *Pattern Recognition,* 35(11):341–354, 2002.
25. D Dubois and H Prade. Putting Rough Sets and Fuzzy Sets Together, *Intelligent Decision Support,* pages 203-232, 1992.
26. H Thiele. Fuzzy Rough Sets Versus Rough Fuzzy SetsAn Interpretation and a Comparative Study using Concepts of Modal Logics, *Technical Report*, 1998.
27. G C Y Tsang, D Chen, E C C Tsang, J W T Lee and D S Yeung. On Attributes Reduction with Fuzzy Rough Sets, *In Proceedings of IEEE International Conference on Systems, Manual Cybernatics,* 03:2775–2780, 2005.
28. Richard Jensen and Qiang Shen. New Approaches to Fuzzy-Rough Feature Selection, *IEEE Transactions on Fuzzy Systems*, 17(4):824–838, 2009.
29. L A Zadeh. The Concept of a Linguistic Variable and its Application to Approximate ReasoningI, *Information Sciences* , 8:199–249, 1975.
30. Radaideh Q A, Sulaiman M N, Selamat M H and Ibrahim H. Approximate Reduct Computation by Rough Sets based Attribute Weighting, *International Conference on Granular Computing*, 2:383–386, 2005.
31. Feng Jiang, Yuefei Sui and Cungen Cao. An Incremental Decision Tree Algorithm based on Rough Sets and its Application in Intrusion Detection, *Artificial Intelligence Review*, 40(4):517–530, 2013.
32. Richard Jensen and Qiang Shen. Fuzzy Rough Attribute Reduction with Application to Web Categorization, *Fuzzy Sets and Systems*, 2003.
33. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu and Ali A Ghorbani. A Deatiled Analysis of the KDD CUP 99 Data Set, *Proceedings of 2009 IEEE Symposium on Computer Intelligence in Security and Defense Applications(CISDA)*, 2009.
34. R Lippmann and R Cunningham. Improving Intrusion Detection Performance using Keyword Selection and Neural Networks, *Computer Networks*, 34(4):579–603, 2000.
35. UCI Machine Learning Repository. Spam Mail Data Sets, *http://www.ics.uci.edu/ mlearn/MLRepository.html*

**G Kishor Kumar** is pursuing Ph.D in Computer Science and Engineering from JNTUA, Anantapuramu, Andhra Pradesh, India. He received his MTech in Computer Science and Engineering from the same University. He received BTech Degree in Computer Science and Engineering from JNTUH, Hyderabad, India. At present, he is working as Assistant Professor at Rajeev Gandhi Memorial College of Engg and Technology(Autonomous), Nandyal, India. His re-

search interest includes Pattern Recognition, Data Mining.

**Viswanath Pulabaigari** did his MTech in Computer Science and Engineering at IIT Madras, India in 1996. He received his Ph.D from IISc Bangalore, India in 2005. He received best thesis award from IISc Bangalore for his Ph.D Thesis. He worked as a faculty member at IIT Guwahati, India. At present he is working as a Professor and Dean CSE in the Department of Computer Science and Engineering, he is also coordinator for the Machine Learning Research Laboratory, Department of Computer Science and Engineering, Madanapalle Institute of Technology and Science – Madanapalle, India. His research interest includes Pattern Recognition, Data Mining and Algorithms.

**Ananda Rao Akepogu** received BTech Degree in Computer Science and Engineering from University of Hyderabad, Andhra Pradesh, India and MTech Degree in AI and Robotics from University of Hyderabad, Andhra Pradesh, India. He received Ph.D Degree from Indian Institute of Technology Madras, Chennai, India. He is Professor of Computer Science and Engineering Department and currently working as Director, Academic and Planning of JNTUA College of Engineering, Jawaharlal Nehru Technological University, Andhra Pradesh, India. Dr. Rao published more than 100 publications in various National and International Journals/Conferences. He also received Best Educationist Award, Bharat Vidya Shiromani Award, Rashtriya VidyaGaurav Gold Medal Award, Best Computer Teacher Award and Best Teacher Award from the Andhra Pradesh Chief Minister for the year 2014. His main research interest includes Software Engineering and Data Mining.