An Efficient Classification based Fuzzy Rough Set Theory using ID3 Algorithm

Suchita Raghuwanshi
M.Tech. Scholar
CSE Department
S.A.T.I. Vidisha M.P.

Ramratan Ahirwal
Assistant Professor
CSE Department,
S.A.T.I. Vidisha M.P.

ABSTRACT
Machine learning is a concerned with the design and development of algorithms. Machine learning is a programming approach to computers to achieve optimization. Classification is the prediction approach in data mining techniques. Decision tree algorithm is the most common classifier to build tree because of it is easier to implement and understand. Attribute selection is a concept by which more significant attributes are selected in the given datasets. The proposed novel hybrid approach with combine of fuzzy set, rough set and ID3 algorithm called FuzzyRoughSetID3 classifier which is used to deal with uncertainties, vagueness and ambiguity associated with datasets. In this approach the selected significant attributes based on hybridization of fuzzy set and rough set theory by using fuzzy positive region and degree of dependency and take reduct as an input to ID3 classifier for constructing the decision tree which is more efficient and scalable approach as compare to previous methods such as ID3, FID3.

Keywords
Decision tree, ID3 algorithm, Rough set, Fuzzy set

1. INTRODUCTION
The major objective of this paper is to compare the classification algorithms for decision trees for data analysis. Classification problem is important task in data mining. Because today databases are becoming rich with cryptic information that can be used for making intelligent business decisions. To understand that information, data classification is a form of data analysis that can be used to extract models describing important data classes or to predict future data trends.

1.1 Decision Tree Classification
Decision tree is a very practical and popular approach in the machine learning domain for solving classification problems. In decision tree approach ID3 algorithm is the most popular and used in this area. Decision tree is one of the classification methods and supervised learning algorithms, supervised learning approach is chosen to unsupervised learning approach because its former knowledge of the class labels of data records makes feature selection simple and this leads to good prediction/classification accuracy. Decision tree classifiers are used successfully in many miscellaneous areas such as radar signals classification, character recognition, remote sensing, speech recognition, medical diagnosis, expert system etc. Perhaps, the most important feature of decision tree classifiers is their capability to break down a sophisticate decision-making process into a collection of easier decisions, thus conferring a solution which is often simpler to interpret. The like benefits capacitate to handle both numerical and categorical data, performs well with large data in a short time and so on makes decision trees superior to many data mining methods.

1.2 Lung Cancer Dataset
- Data was published in:

- Hong, Z.Q. and Yang, J.Y. "Optimal Discriminant Plane for a Small Number of Samples and Design Method of Classifier on the Plane", Pattern Recognition, Vol. 24, No. 4, pp. 317-324, 1991.
- Donor:Stefan Aeberhard, stefan@coral.cs.jcu.edu.au
- Date: May, 1992[6]

In the lung cancer dataset there are 32 numbers of instances and 57 number of attributes (1 class attribute, 56 predictive). Missing attribute values are 5 and 39.

There are so many applications of rough set theory in real world i.e. Financial investment, Prediction of business failure, Bioinformatics and medicine, Spatial and meteorological pattern, Music and acoustics, Fault diagnosis.

2. LITERATURE SURVEY
Various classification methods have been proposed over the year's e.g. Naive Bayesian approach, neural networks, decision trees, nearest-neighbor method, genetic algorithms etc. In this paper, ***Sanjay Kumar Malik et al [1]*** attention is restricted to decision tree method after considering all its advantages compared to other methods. There exist a large number of algorithms for inducing decision trees like CHAID, FACT, C4.5, CART etc. In this paper, these five decision tree classification algorithms are considered – ID3, SLIQ, SPRINT, PUBLIC and RAINFOREST.

Aimed at the problem of large computation, big tree size and over-fitting of the testing data for multivariate decision tree (MDT) algorithms, ***Dianhong Wang et al [2]*** proposed a novel rough set-based multivariate decision trees (RSMDT) method. In this paper, the positive region degree of condition features with respect to decision features in rough set theory is used for selecting features in multivariate tests. And a new concept of extended generalization of one equivalence relation corresponding to another one is proposed and used for construction of multivariate tests. They experimentally test RSMDT algorithm in terms of data classification accuracy, tree size and computing time, using the whole 36 UCI machine learning repository data sets chosen by Weka platform, and compare it with C4.5 algorithm, classification and regression trees with linear combinations (CART-LC), classification and regression trees (CART), Quick Unbiased Efficient Statistical Trees (QUEST), Oblique Classifier 1 (OC1). The experimental results indicate that proposed RSMDT algorithm significantly outmatches the comparison

classification algorithms with improved classification accuracy, comparatively small tree size, and shorter computing time.

Cuiru et al [3] proposed algorithm for decision tree construction based on rough set theory. They proposed a novel and effective algorithm in which knowledge reduction of rough set theory is applied to reduce irrelevant information from the decision table. In this paper, first of all degree of dependency of all condition feature on decision feature is determined. The condition features which have highest degree of dependency are selected as splitting feature. In case if there is more than two feature which have same degree of dependency then β-dependability is used to select splitting feature. They applied weather dataset for experimental result and compared this result to the ID3 decision tree algorithm. The decision tree generated in consist limited node and produce simple and efficient decision tree.

Baoshi et al [4] developed FID3 (Fixed Information Gain) algorithm. In the FID3, a new parameter fixed information gain is used to select splitting feature. In FID3, feature is reduced by calculating degree of dependency and then fixed information gain of each feature is selected the feature which have highest information gain is selected as splitting feature. FID3 algorithm removes the drawback of ID3 in which feature is selected as splitting feature which have different feature values. There is one more drawback of ID3 is the instability of the decision tree built by information gain is removed by using fixed information gain.

3. PRELIMINARIES

For the better comprehension of the proposed work, firstly some basic concepts of fuzzy rough sets is introduced. The fuzzy positive region and evaluation function are defined as

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

$$\mu_{POS_P(Q)}(x) = \sup_{X\in U/Q} \mu_{\underline{P}X}(x) \quad (2)$$

$$\gamma_P(Q) = \frac{\left|\mu_{POS_P(Q)}(x)\right|}{|U|} \quad (3)$$

Fuzzy lower approximation based feature selection is a new technique based on fuzzy similarity relations by hybridization of fuzzy-rough sets.In fuzzy rough set model the fuzzy P-lower and P-upper approximations, fuzzy positive region and dependency function are defined as

$$\mu_{\underline{R_P}X}(x) = \inf_{y\in U} I\left(\mu_{R_P}(x,y), \mu_X(y)\right) \quad (4)$$

$$\mu_{\overline{R_P}X}(x) = \sup_{y\in U} T\left(\mu_{R_P}(x,y), \mu_X(y)\right) \quad (5)$$

$$\mu_{POS_{R_P}(Q)}(x) = \sup_{X\in U/Q} \mu_{\underline{R_P}X}(x) \quad (6) \qquad \gamma_P(Q) = \frac{\sum_{x\in U} \mu_{POS_{R_P}(Q)}(x)}{|U|} \quad (7)$$

Here, I is a fuzzy implicator and T a *t*-norm. R_P is the fuzzy similarity relation motivated by the subset of features P.

$$\mu_{R_P}(x,y) = \cap_{a\in P}\{\mu_{R_a}(x,y)\} \quad (8)$$

Where $\mu_{R_a}(x,y)$ is the degree to which objects *x* and *y* are similar for feature *a*. Many fuzzy similarity relations can be created for this purpose, for example

$$R_a(x,y) = max\left(\min\left(\frac{a(y)-a(x)+\sigma_a}{\sigma_a}, \frac{a(x)-a(y)+\sigma_a}{\sigma_a}\right), 0\right) \quad (9)$$

3.1 ID3 (Iterative Dichotomiser 3) Algorithm

ID3 is a simple decision learning algorithm developed by J. Ross Quinlan (1986). ID3 builds decision tree by employing a top-down, greedy search through the given sets of training data to test each feature at every node. It uses statistical property call information gain to choose which feature to test at each node in the tree. The information gain measures how well a given feature separates the training examples according to their target classification [5].

Suppose S is the set of example set, and the number of equivalence class built by indiscernibility relation is n then entropy is defined as:

$$Entropy(S) = -\sum_{i=1}^{n} p_i log_2\, p_i \quad (10)$$

Where $Entropy(S)$ the entropy of is set S and pi is the proportion of the number of elements in class i to the number of elements in set S.

where $pi = \frac{Si}{|S|}$, $|s|$ is the number of example set S. Given a feature $A \in C$, is the set of condition features the domain of A is denoted as V_A, then the expected information of the entropy is given as follows:

$$Info_A(S) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i) \quad (11)$$

where $\frac{|S_i|}{|S|}$ is the proportion of the number of elements in i to the number of elements in set S.

$Entropy(S_i)$ is the entropy of subset i.

Hence, the information gain on $A \in C$ is defined as

$$Gain(S,A) = Entropy(S) - Info_A(S) \quad (12)$$

Compute the information gain of each conditional feature, and the feature with the maximum information gain is the most informative feature.

3.2 Pseudo code

ID3 (Examples, Target Feature, Features)

Examples are the training examples. Target Feature is the feature whose predicted value by the tree. Feature is the list of features which may be tested by the learned decision tree. returns a decision tree that correctly classifies the given Examples.

(1) Create a root node for the tree

(2) If all examples are +ve, Return the single-node tree Root, with label = +

(3) If all examples are -ve, Return the single-node tree Root, with label = -

(4) If number of predicting features is empty, Return the single node tree Root, with label = most common value of the target feature in the examples

(5) Otherwise begin

A ←The Feature that best classifies examples

Decision Tree feature for Root ←A

For each positive value, vi, of A,

Add a new tree branch below Root, corresponding to the test A = vi

Let Examples, be the subset of examples that have the value vi for A

If Examples(vi) is empty

Then below this new branch add a leaf node with label = most general target value in the examples

Else below this new branch add the sub tree ID3 (Examples(vi), Target Feature, Features – {A})

End

(6) Return Root

4. PROPOSED METHODOLOGY

4.1 FuzzyRoughSetID3 Algorithm

Now the proposed algorithm to generate a decision tree in the following way:

Input: An information system IS= (U, A)

Output: A decision tree *T*.

Step 1: Create an initial node of tree based on maximum Information Gain of attribute of fuzzy rough- reduct. Judge that whether the samples are all of the same class. If they are, then turns the node into a leaf and return the leaf labeled with that class.

Step 2: For each attribute in P_i, Calculate Information Gain according to, Choose the attribute A_i with the maximum value of Information Gain as the root node. Where *Pi* is the set of fuzzy rough- reduct.

Step3: Constructing the branches according to different values of attribute *Pi* so that the samples are partitioned accordingly.

Step 4: If samples in a certain value are all of the same class, then generate a leaf node and is labeled with that class.

Step 5: Otherwise use the same process recursively to form a decision tree for the samples at each partition.

Step 6: Build nodes and branches repeat until any one of the following conditions is satisfied:

All samples for a selected node belong to the same class, return a leaf labeled with that class.

There are no more training samples to be classified, we can create a leaf belong to the class in majority among samples.

Step 6: Output the decision tree T.

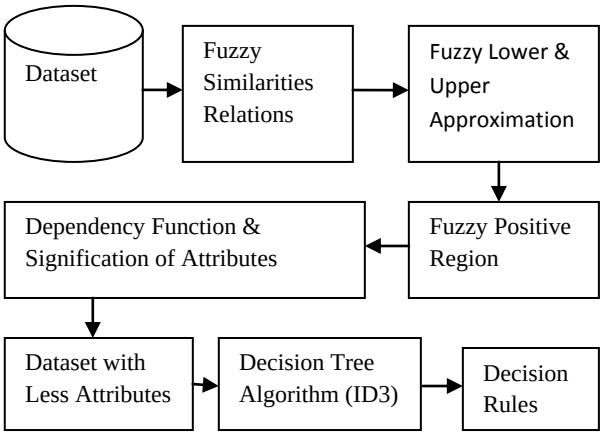

Figure 1: Proposed Methodology

5. EXPERIMENTAL EVALUATION

Accuracy of proposed classification algorithm is a measurement of total accurate identified instances over the given samples. The accuracy of the classification can be evaluated on following datasets [6] using Use training set test mode.

Table 1. Accuracy Comparisons between FID3 and FuzzyRoughSetID3

Datasets	Instances	Attributes	FID3 Accuracy (%)	FuzzyRoughSetID3 Accuracy (%)
Weather nominal	14	4	84.20%	100%
Irish	150	4	66.50%	100%
Wine	178	14	74.70%	100%
Lung Cancer	32	56	83.33%	100%

The comparative accuracy of two algorithms are given using Table 1 shows the better performance of FuzzyRoughSetID3 than FID3 algorithm. According to the evaluated results the performance of the proposed algorithm is much better as compared to another algorithm.

The amount of time consumption required to developing data model using proposed algorithm is as on following datasets. Time consumption means time complexity of the algorithm on various datasets.

Table 2. Time Consumption of FID3 and FuzzyRoughSetID3

Datasets	Instances	Attributes	FID3 Time Consumption (In Seconds)	FuzzyRoughSetID3 Time Consumption (In Seconds)
Weather nominal	14	4	0	0

Irish	150	4	0.03	0.02
Wine	178	14	0.03	0.05
Lung Cancer	32	56	0.03	0

The comparative time complexity of algorithms is given using Table 2 shows the better performance of FuzzyRoughSetID3 than FID3 algorithm.

6. CONCLUSION

Decision tree algorithm is classical approach of supervised machine learning and data mining. There are a number of decision tree algorithms are available such as ID3, C4.5 and others. The decision tree algorithms are able to develop a transparent and reliable data model. In order to maintain the transparency and relativity between attributes decision tree algorithms are computationally expensive in terms of memory and time. Therefore, a number of approaches are developed in recent years by which the classifiers are claimed to provide much efficient classification accuracy in less complexity. To overcome these computationally expensive in our proposed approach.

In this presented work the fuzzy set and rough set hybridization is done for feature selection and decision tree is constructed by ID3 Classifier. By combining this approach, a new FuzzyRoughSetID3 algorithm is proposed and implemented. The proposed algorithm is enhancing classification accuracy of datasets, reducing the size of tree and minimizing the redundancy in data.

The proposed model is implemented using WEKA 3.7.2 [7] and NetBeans IDE 8.1 and the comparative study is performed with respect to the FID3 algorithm and FuzzyRoughSetID3 algorithm. The comparison among these algorithms is performed in case of accuracy and time complexity. The comparative performance is as following in table 3

Table 3. Comparative Performance

S.No.	**Parameters**	**Proposed FuzzyRoughSetID3**	**FID3**
1	Accuracy	High	Low
2	Time Consumed	Low	High

The proposed algorithm, FuzzyRoughSetID3 produces high accuracy, low error rate and consumes less time as compared with FID3 algorithm and. Thus proposed algorithm provides efficient and effective results for classification of datasets.

7. FUTURE WORK

The proposed algorithm is efficient and accurate which provides effective results as compared to the traditional algorithms. In future it will optimize the performance of classification in terms of memory consumption and training time. In future this algorithm can be used in parallel for analysis of big data.

8. ACKNOWLEDGMENT

I would like to thank the anonymous referees for their helpful guidance that have improved the quality of this paper.

9. REFERENCES

[1] Sanjay Kumar Malik, Sarika Chaudhary, "Comparative Study of Decision Tree Algorithms for Data Analysis", International Journal of Research in Computer Engineering and Electronics, Page 1, ISSN 2319-376X, VOL :2 ISSUE: 3 (June: 2013)

[2] Dianhong Wang, Xingwen Liu, Liangxiao Jiang, Xiaoting Zhang, Yongguang Zhao, "Rough Set Approach to Multivariate Decision Trees Inducing", Journal of Computers, VOL. 7, NO. 4, APRIL 2012

[3] Cuiru Wang and Fangfang OU, "An Algorithm for Decision Tree Construction Based on Rough Set Theory," International Conference on Computer Science and Information Technology, IEEE, pp. 295- 299, 2008.

[4] Baoshi Ding, Yongqing Zheng, Shaoyu Zang, "A New Decision Tree Algorithm Based on Rough Set Theory", Asia-Pacific Conference on Information Processing, IEEE, pp. 326-329, 2009.

[5] R. Quinlan, "Induction of decision trees," Machine Learning, vol.1, No. 1, pp. 81–106, 1986.

[6] www.ics.uci.edu/-mlearnlMLRepository.html

[7] https://sourceforge.net/projects/weka/files/documentation/3.7.x/WekaManual-3-7-2.pdf/download?use_mirror=tenet&download