International Research Journal of Engineering And Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 01 | Jan-2016 www.irjet.net p-ISSN: 2395-0072

# Diagnosis of Cancer using Fuzzy Rough Set Theory

**L.Meenachi[1], Dr.S.Ramakrishnan[2], M.Arunithi[3], R.Karthiga[4], S.Karthika[5], P.Nandhini[6]**

*[1]Assistant Professor, Department of Information Technology*
*Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India*
*[2]Head/Professor, Department of Information Technology*
*Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India*
*[3456]Students, Department of Information Technology*
*Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India*

---

**Abstract -** ***Cancer is one of the deadliest diseases. Early diagnosis and treatment at early stage can enhance the outcome of the patient. Our main objective is to classify the different types of cancer data. Our project involves four modules: feature selection, instance selection, classification and performance analysis. It identifies the appropriate set of features by eliminating the irrelevant features to improve the performance of the classifier. The Fuzzy Rough Subset evaluation method is used in conjunction with a Particle Swarm Optimization (PSO) for feature selection. In the second module, the missing values are removed in the data set using RemoveMissing filter. Then the Instance Selection algorithm is used to identify appropriate set of instances by eliminating useless and erroneous instances. Next in the third phase, the Fuzzy-Rough Nearest Neighbor algorithm is utilized to classify the data set obtained from the above steps. Finally the performance of the classifier is evaluated using evaluation metrics.***

***Key Words:*** *FuzzyRoughSubset Evaluator, Particle Swarm Optimization, Fuzzy-Rough Nearest Neighbor.*

## 1. INTRODUCTION

Cancer is a deadly disease, also called as malignant tumor. It causes abnormal cell growth or alteration in the cell genetic structure and they also have potential to spread or invade other parts of the body. Some of the symptoms include: a new lump, unexplained weight loss, abnormal bleeding, a prolonged cough and a change in bowel movements among others. Early diagnosis and treatment of cancer can enhance the outcome of patients. There are over 100 different known cancers that will affect humans. Hence classification of cancer helps to identify cancer at earlier stagewhich helps in determining appropriate treatment and helps to determine the prognosis. When cancer is identified in any of the patients then they can start their treatment and therapy at the earlier stage of cancer.

Initially patient's records are collected and transformed into data sets. And we have to identify the appropriate set of features from which we can classify the cancer. This process involves selecting minimum feature set by eliminating irrelevant features through which we can improve the classification accuracy. Instance selection aims to reduce the number of instances in the data set by either eliminating bad instances or extracting as much instances as possible so that the noise in the original data set can be reduced. It also removes instances that cause conflicts with other instances.

A test pattern related to each type of cancer is developed. The selected instances are matched with the test pattern and classified based on the matches found. The technique used for the above processes are particle swarm optimization for feature selection, Fuzzy Rough Instance Selection method along with weak gamma evaluator as a measure for instance selection and fuzzy rough nearest neighbor classifier for classification process.

The accuracy of a classifier for a given test set is defined as the percentage of test set tuples that are correctly classified by the classifier. The associated class label related to each test tuple is compared with the learned classifier's class prediction for that tuple. If the accuracy of the classifier is acceptable, the classifier can be used to classify future data tuples for which the class label is not known.After completion of the above processes some metrics are calculated like kappa statistics, sensitivity, specificity, f-measure and area under curve only by then we can identify whether ourclassification method is more efficient for classification of data.

## 2. RELATED WORK

There has been a lot of research on the diagnosis of cancer and classification of data with the data set in the literature with a relatively high classification performance.

Fuzzy rough set theory preserves the original meaning of the features even after reduction. This may help the applicationsthat involve datasets with huge number of features, which would be impossible to process

further. When comparing to the results from unreduced dataFRFS [7] is shown to equal or improve classification accuracy. Classifiers that use a lower dimensional set of attributes which are retained by fuzzy-rough reduction outperform those that employ more attributes returned by the existing crisp rough reduction method. In addition, it is shown that FRFS is more powerful than the other FS techniques in the comparative study.

A medical decision making system based on SVM [1] combined with feature selection has been applied for diagnosing breast cancer.To diagnose breast cancer researchers commonly use machine learning methods andexperiments were conducted on different portions of the WBCD. It is observed that the proposed method yields better classification accuracy. They used support vector machine and F-score based feature selection [1] because SVM is used for large and complex data set but their disadvantage is SVM has several key parameters that should be kept correct to achieve the best classification results which is impractical.

A hybrid intelligent classification model for medical data consists of Fuzzy Min–Max Neural Network [11], Regression tree and the Random Forest algorithm [11]. Regression tree [11] is operated by building a number of decision trees at training timebut their algorithm is slow for real time prediction due to large number of trees. A series of empirical studies using three benchmark medical datasets from the UCI Machine Learning Repository, namely Breast Cancer Wisconsin, Liver Disorders, and Pima Indians Diabetes has been used to evaluate the efficiency of the hybrid model. Different experimental configurations have been adopted in order to provide a fair performance comparison with different models reported in the literature. The main contribution of this paper is the hybrid model that possesses three important characteristics for tackling medical decision support tasks such as online learning, high performance, and rule extraction.

Fuzzy-rough NN [6] classification approach performs better under partially exposed and unbalanced domain compared with approach of the crisp NN and fuzzy NN. The result of this fuzzy-rough NN approach will containboth upper and lower membership degree, hence from the output of the new approach more meaningful interpretation can be drawn. This research has some limitations. The simulation for fuzzy-rough NN approach to small data set no larger than 1000 instances thus may restrict its performance estimation in larger data sets and data set that has no missing value or having missing value less than 5 percent of all the instances.

## 3. PROPOSED SYSTEM

In our proposed system, we have to identify the appropriate set of features by eliminating the irrelevant features to improve the performance of the classifier The Fuzzy Rough Subset evaluation method is used in conjunction with a PSO Search algorithm for feature selection.PSO uses a number of agents (particles) that constitute a swarm, which is moving around the search space looking for the best solution.

Next in instance selection module Fuzzy Rough Instance Selection algorithm is used to identify appropriate set of instances by eliminating useless and erroneous instances. In the third phase of the model, the Fuzzy-Rough Nearest Neighbor algorithm is utilized to classify the data set obtained from the above steps.

Finally the performance of the classifier is evaluated using evaluation metrics such as classification accuracy, sensitivity, specificity, F-measure, Area Under Curve (AUC), and Kappa statistics.

### 3.1 Technique Used

Fuzzy-Rough Set Theory

A fuzzy rough set is a derived from the approximation of a fuzzy set.This corresponds to the case where only the decision attribute values are fuzzy; the conditional values are crisp. The upper and lower approximations incorporate the extent to which objects belong to these sets, and are defined as:

$$\mu_{\underline{R}X}([x]_R) = \inf\{\mu_X(x) | x \in [x]_R\}, \quad (1)$$

$$\mu_{\overline{R}X}([x]_R) = \sup\{\mu_X(x) | x \in [x]_R\}, \quad (2)$$

where$\mu_X(x)$ is the degree to whichxbelongs to fuzzy equivalence class X, and each$[x]_R$is crisp. The tuple $<\underline{R}X,\overline{R}X>$is called a rough-fuzzy set.

Rough-fuzzy sets can be generalised to fuzzy-rough sets, where allequivalenceclasses may be fuzzy. When applied to dataset analysis, this means that both the decision values and the conditional values may be fuzzy or crisp.

Fuzzy-rough set-based Feature Selection (FRFS)is based on thenotion of fuzzy lower approximation to reduce the datasets containing realvaluedfeatures. The process becomes a crisp approach when dealing with nominal well-defined features. Positive region is defined as the union of lower approximations.

## 4. MODULES

### 4.1 Feature Selection

It identifies the appropriate set of features by eliminating the irrelevant features to improve the performance of the classifier.It is performed by Fuzzy

Rough subset evaluation in conjunction with Particle Swarm Optimization (PSO) Search algorithm. In PSO search algorithm number of agents is used to make the swarm move around in the looking for the best solution.

In an N-dimensional space each particle is treated as a point which adjusts its "flying" based on its flying experience as well as other flying experience of the particles. Each particle tracks its coordinates which can be associated with the best solution (fitness) that has achieved by the particle, known as personal best, pbest. The best value of the neighborhood particle of the particle is considered as the best value by PSO. And this value is called gbest.

### 4.2 Instance Selection

It selects the appropriate set of instances by eliminating the erroneous and noisy instance. Before that, remove the instances with missing values using RemoveMissing; this filter is available under Pre-processing tab. The robustness of the algorithms can be tested by adding noise to the data set. Next the fuzzy-rough instance selection method is implemented as filters.

Finally implement the Instance Selection algorithm in Pre-process tab and set the measure as Weak Gamma to select the instances using positive region.

### 4.3 CLASSIFICATION

It classifies the data based on the values obtained during training set and in a classifying attribute. Classification is done using the Fuzzy-rough nearest neighbor classifier.

Fuzzy-rough nearest neighbor classifier is the extension of K-nearestneighbor algorithm by using the fuzzy-rough uncertainty.The fuzzy uncertainty concept is used to measure the distancebetween the test pattern and the neighbor. It also helps torepresent the neighbour to be in many classes. Due to lack of features some of the neighbors and the test patterns may be indistinguishable hence the concept of rough uncertainty is used.Theneighborhood structure is artificial, so the roughness emerges.

In the fuzzy-rough nearest neighbour algorithm, the initial memberships of training patterns will be crisp, constrained fuzzy or may have possible values. The algorithm differsfrom fuzzy counterpart because it is not necessary to fix the K parameter. With the help of a fuzzy-rough ownership function the above mentioned uncertainties in the usual K-nearestneighbor algorithm are found.

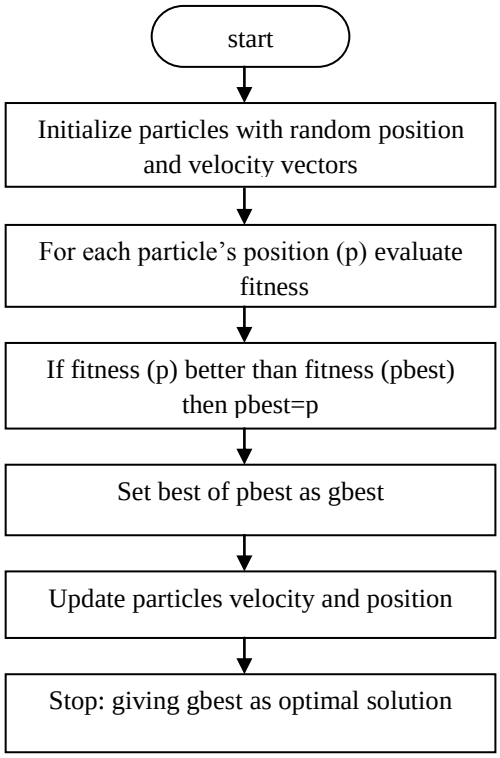


**Fig -1**: PSO Search Algorithm

The fuzzy-rough ownership function is used to identify thefuzzy-roughness present in the data. K parameter of the algorithm may be either fixed or it may be made proportionalto the inverse of the average distance between all neighbours andthe test pattern.

### 4.4 PERFORMANCE ANALYSIS

Performance is measured based on some evaluation metricslike classification accuracy, sensitivity, specificity,F-measure, Area Under Curve and Kappa statistics.

#### 4.4.1 Classification accuracy

Classification accuracy is one of the most popular metricsin classifier evaluation. It is the proportion of the number of truepositivesand true negatives obtained through the classification algorithmsin the total number of instances.

$$\text{Accuracy} = \frac{TN+TP}{TP+FP+FN+TN} \quad (3)$$

#### 4.4.2 Sensitivity

Another common metric for evaluation of classifiers is the sensitivityof algorithm. Sensitivity represents the true positiverate and it is calculated by the divisionof true positive classificationby true positive and false negative classification.

$$\text{Sensitivity} = \frac{TP}{TP+FP} \quad (4)$$

#### 4.4.3 Specificity

Specificity of the classification algorithm is also used for evaluation. It is defined as the number of true negative classification divided by all other negative classifications.

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (5)$$

#### 4.4.4 F-measure

F-measure is the harmonic mean of precision and recall where Precision is the proportion of the true positives against all thepositive results (i.e. both true positives and false positives) and Recall is the proportion of true positives against true positivesand false negatives.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (7)$$

$$F-\text{measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (8)$$

#### 4.4.5 Kappa statistics

Kappa statistics is based onthe difference between the actual agreement in the error matrixand chance agreement. The values range from 0 to1.

$$\text{Kappa} = \frac{Pr(a)-Pr(e)}{1-Pr(e)} \quad (9)$$

#### 4.4.6 Area under curve (AUC)

Area under curve (AUC) is another metric for evaluating the classifiers. It equals to the probability of which a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. It takes on values from 0 to 1.

## 5. RESULT

Initially the dataset which is used for our analysis is loaded into the tool. Soon after loading the dataset we can view the attributes, instances and other information about the attributes that is count of the attribute, missing values, distinct values and type of the attribute. Feature selection is employed using the select attribute tab. For attribute evaluator Fuzzy rough subset evaluator is chosen and Particle Swarm Optimization(PSO) is chosen for select attribute. Consider a breast-cancer dataset collected from Wisconsin Breast Cancer Dataset (WBCD) repository which has 10 features.

**Table -1:** Before Feature Selection

| *S.No* | *Features* |
|---|---|
| 1 | age |
| 2 | menopause |
| 3 | tumor-size |
| 4 | inv-nodes |
| 5 | node-caps |
| 6 | deg-malig |
| 7 | breast |
| 8 | breast-quad |
| 9 | irradiat |
| 10 | class |

After applying PSO search algorithm only 8 features are selected which is shown in Table II. Number of features reduced after using PSO algorithm for other different types of cancer datasets is also shown in Table III.Attribute selection mode is set as full training set. When start button is clicked attribute selection will be done and output is shown. PSO search algorithm will generate the subset and calculate the merit value for each subset. All the generated subset and merit value for each subset is shown in the output. Finally the optimal subset with high merit value is chosen as the best subset of attribute and the attributes are displayed.

**Table -2:** After Feature Selection

| S.NO | FEATURES |
|---|---|
| 1 | age |
| 2 | menopause |
| 3 | tumor-size |
| 4 | node-caps |
| 5 | deg-malig |
| 6 | Breast |
| 7 | breast-quad |
| 8 | Irradiat |

Remaining attribute which are irrelevant will be easily identified. Next in the pre-processor tab all the attributes which are irrelevant to our system is selected and removed using remove button. Next in the pre-processor tab filter is chosen. Instances with missing values are removed using the Remove Missing filter.

**Table -3:** Feature Selection

| Data set | Before Feature Selection | After Feature Selection |
|---|---|---|
| Breast cancer | 10 | 7 |
| GCM | 16064 | 16 |
| Leukemia | 7130 | 32 |
| Lung-cancer | 57 | 6 |

Since class labels are not known unsupervised filters are used. After removing the missing values Instance Selection filter is used to remove the instances which are duplicate and causing conflict with other instances.

**Table -4:** Instance Selection

| Data set | Before Instance Selection | After Instance Selection |
|---|---|---|
| Breast cancer | 699 | 647 |
| GCM | 46 | 45 |
| Leukemia | 34 | 33 |
| Lung-cancer | 32 | 22 |

**Table -5:** Classification Using Fuzzy Rough NN

| Dataset | Lung Cancer | GCM | Leukemia | Breast Cancer |
|---|---|---|---|---|
| Accuracy | 99.976 | 99.984 | 99.995 | 97.658 |
| F-measure | 0.986 | 1 | 0.967 | 0.916 |
| Sensitivity | 1 | 0.997 | 1 | 0.916 |
| Specificity | 1 | 1 | 1 | 1 |
| Kappa Statistics | 1 | 1 | 0.983 | 0.979 |

## 6. CONCLUSION

This paper is about classification method to classify the cancer data. The two necessary pre-processing steps done before classification is feature selection and instance selection. Particle Swarm Optimization technique is used for feature selection to reduce the number of features by eliminating the irrelevant feature which is not important during the classification of data. The selected features are used for further process. In instance selection process noisy data, inconsistent data and missing data's are removed by Fuzzy Rough Instance Selection technique. In this process weak gamma evaluator is used as a measure. The above two pre-processing steps will reduce the complexity in dataset and improve the efficiency. Fuzzy Rough Nearest Neighbor classifier is used for classification of cancer data from the given dataset. After classification, the efficiency of classified data is measured by calculating some evaluation metrics like kappa statistics, sensitivity, area under curve (AUC), F-measure.

## 7. REFERENCES

[1] Akay, M. F., "Support vector machines combined with feature selection for breast diagnosis". *Expert Systems with applications,*36,3240-3247,2009.

[2] archive.ics.uci.edu/ml

[3] Aytug Onan, "A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer".*Expert Systems with Applications*,42, 6844–6852, 2015

[4] Bhardwaj, A., & Tiwari, A., "Breast cancer diagnosis using genetically optimized neural network model". *Expert Systems with Applications*, 42,4611–4620, 2015.

[5] Bonyadi, Mohammad reza; Michalewicz, Z, "An analysis of the velocity updating rule of the particle swarm optimization algorithm",2014

[6] Haiyun Bian, Lawrence Mazlack,"Fuzzy-Rough Nearest-Neighbor Classification Approach". *University of Cincinnati*.

[7] Keller J.M., Gray M.R. and Givens J.A., "A Fuzzy K-Nearest Neighbor Algorithm," *IEEE Transactions on Systems, Man and Cybernetics*, 15(4), 1985

[8] Nele Verbiest,"Fuzzy Rough and Evolutionary Approaches to Instance Selection".*Ghent University*,2007.

[9] Richard Jensen ,"Combining Rough and Fuzzy Sets for Feature Selection". *University of Edinburgh*,2005.

[10] Seera, M., & Lim, C. P., "A hybrid intelligent system for medical data classification", *Expert Systems with Applications*,41, 2239–2249, 2014.

[11] http://www.thearling.com/text/dmtechniques/dmtechniques.html

[12] www.cs.waikato.ac.nz/ml/weka

[13] Z. Pawlak, "Rough Sets - Theoretical Aspects of Reasoning about Data".*Kluwer Academic Publishers, Boston, London, Dordrecht*, P.229, 1991.

[14] Z. Pawlak, "Rough Sets and Data mining".*Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, ul. Baltycka 5, 44 100 Gliwice, Poland,*1992