IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Cancer Classification from Gene Expression data using Fuzzy-Rough techniques: An Empirical Study

Ansuman Kumar[1], Anindya Halder[2*]

[1,2] Dept. of Computer Application, North-Eastern Hill University, Tura Campus, Meghalaya, India

[*]*Corresponding Author: anindya.halder@gmail.com*

Available online at: www.ijcseonline.org

***Abstract*—** Cancer classification from gene expression data is one of the most challenging research areas in the field of computation biology, bioinformatics and machine learning as the number of clinically labeled samples are very few compared to number of genes present. Also the cancer subtype classes are often highly overlapping, imprecise, and indiscernible in nature. Various machine techniques have been developed and applied on gene expression data for cancer sample classification. Here in this article, an empirical study of cancer classification from microarray gene expression data is performed using fuzzy-rough nearest neighbour techniques where performance of four different types of classifiers viz., Fuzzy nearest neighbour, Fuzzy-rough nearest neighbour, Vaguely quantified fuzzy-rough nearest neighbour and Ordered weighted average based fuzzy-rough nearest neighbor are investigated. The experiments are carried out on eight publicly available real life microarray gene expression cancer datasets. To assess the results of the classifiers percentage accuracy, precision, recall, macro averaged F_1 measure, micro averaged F_1 measure and kappa are used. The comparative study of the investigated methods is also done using paired *t*-test. Fuzzy-rough nearest neighbour method is found to be better for most of the data sets for cancer classification.

***Keywords*—**Cancer Classification, Fuzzy-Rough set, Vaguely Quantified, Ordered Weighted Average, Microarray Gene Expression data.

I. INTRODUCTION

Classification of cancer sub-types classes is of immense significance in early stage diagnosis and drug discovery. Cancer diagnosis by traditional methodologies depends on the clinical findings and morphological appearance of the tumor. These methods are usually costly, some time inaccurate and time taking. Also traditional methods are usually restricted due to expert's observation in differentiating different cancer subtypes classes as the most cancers are highly related to the specific biological perception. In this context, the latest development of microarray technology [1] has enabled biologists to specify thousands of genes in a single experiment in order to produce comparatively low-cost diagnosis and prediction of cancer at early stage.

Different machine learning techniques were being applied for microarray gene expression data analysis using supervised (i.e., classification) [2], unsupervised (i.e., clustering) [3], semi-supervised clustering [4], and semi-supervised classification [5] mode.

Usually, the number of samples present in microarray gene expression data is very less compared to the number of genes [6], and the classes present in data are often vague, indiscernible and overlapping in nature. Hence, it is necessary to test the different types of classifiers particularly the fuzzy and rough set based classifiers on cancer datasets to handle the overlapping, vague and indiscernible subtype classes of microarray gene expression datasets. Motivated from these issues in this article, an empirical investigation is done using different fuzzy-rough based classification techniques, viz., Fuzzy nearest neighbour, Fuzzy-rough nearest neighbour, Vaguely quantified fuzzy-rough nearest neighbour and Ordered weighted average based fuzzy-rough nearest neighbor for cancer classification from microarray gene expression data.

The remainder of the article is structured as follows. The background theory pertinent to this article is briefly described in Section 2. Section 3 provides a detailed description of methodologies presented. Details of the experiments and analysis of the results are given in Section 4, and finally, conclusions and future scope of work are drawn in Section 5.

II. BACKGROUND STUDY

Fuzzy-rough nearest neighbour classifier is an amalgamation of fuzzy set and rough set thus brief outline of those are provided below.

A. *Fuzzy set theory*

Fuzzy set theory was proposed by L. A. Zadeh [7] in 1965. It is an extension of crisp sets to handle vague and imprecise data. Fuzzy set *A* uses mapping from the universe X to the interval [0, 1]. The value *A(x)* for $x \in X$ is called the membership degree of *x* in *A*.

B. *Rough set theory*

Rough set theory was introduced by Z. Pawlak [8] in early 1980s. It can handle uncertainty, indiscernibility and incompleteness in the datasets. It starts with the idea of an approximation space, which is a pair < *X*, *R* >, where *X* is the non-empty universe of discourse and *R* is an equivalence relation defined on *X*, where *R* satisfies the reflexive, symmetric and transitive property. For each subset *A* of *X*, the lower approximation defined as the union of all the equivalence classes which are fully included inside the class *A*, and the upper approximation is defined as the union of equivalence classes which have non-empty intersection with the class *A*.

C. *Fuzzy-rough set theory*

Fuzzy set theory can handle vague information, while rough set theory can handle incomplete information. These two theories are complementary to each other. Hybridization of these two concepts yields the idea of the fuzzy-rough set which is the pair of lower and upper approximations of a fuzzy set *A* in a universe *X* on which a fuzzy relation *R* is defined. The fuzzy-rough lower and upper approximations of *A* are defined respectively as follows [9]:

$$(R \downarrow A)(x) = \inf_{y \in X} I(R(x,y), A(y)) \quad (1)$$

$$(R \uparrow A)(x) = \sup_{y \in X} T(R(x,y), A(y)) \quad (2)$$

where *I* is the Lukasiewicz implicator, *T* is the Lukasiewicz *t*-norms and $R(x,y)$ is the valued similarity of patterns x and y, inf is infimum and sup represents supremum.

III. METHODOLOGY

Methods investigated for cancer classification from microarray gene expression data in the present study are described briefly in this section below.

A. *Fuzzy k- Nearest Neighbour Classifier*

Fuzzy *k*-Nearest Neighbour (FKNN) [10] is an extension of the *k*-Nearest Neighbour (KNN) classifier. In *KNN* algorithm, equal weightage is given to all the *k*-nearest neighbours to calculate the predicted class of a test data. FKNN algorithm assigns fuzzy membership of a test pattern in each class. That class is taken to be the predicted class (of that test pattern) for which the fuzzy-membership is maximum. Microarray gene expression data have a very high dimension which contains thousands of genes. However, number of samples present in the microarray gene expression data is often very less and sometimes subtype classes have overlapping and indiscernibility. In these cases fuzzy *k-NN* algorithm is expected to provide better result than *k-NN*.

B. *Fuzzy-Rough Nearest Neighbour Classifier*

Fuzzy-Rough Nearest Neighbour (FRNN) [11] classifier is the combination of fuzzy and rough theories. It uses the concept of upper and lower approximations to assign the class label information to the test pattern. The values of lower and upper approximations of a decision class are computed based on the *k*-nearest neighbours of test pattern.

Detailed procedure of FRNN Classifier is provided below:

1. Compute the *k*-nearest neighbour (*kNN*) labeled patterns closest to each of the test pattern (t) based on the Euclidean distance (compute the distance from the labeled pattern to test pattern).
2. The values of lower and upper approximations of test pattern (t) for belonging to each class *C* is calculated respectively as follows:

$$(R \downarrow C)(t) = \inf_{y \in kNN} I(R(t,y), C(y)) \quad (3)$$

$$(R \uparrow C)(t) = \sup_{y \in kNN} T(R(t,y), C(y)) \quad (4)$$

where *I* is the Lukasiewicz implicator, *T* is the Lukasiewicz *t*-norms and $R(t,y)$ is computed as:

$$R(t,y) = \frac{\sum_{y \in kNN} (\|t-y\|)^{\frac{2}{m-1}}}{(\|t-y\|)^{\frac{2}{m-1}}}; \quad (5)$$

where $\|t-y\|$ is the distance of the test pattern (t) from the labeled pattern $y \in kNN$ (*k*-nearest neighbour labeled pattern of test pattern t) and m $(1 < m < \infty)$ is the fuzzifier. $C(y)$ is computed as:

$$C(y) = \begin{cases} 1, & if \quad y \in C; \\ 0, & Otherwise. \end{cases} \quad (6)$$

3. The test pattern (t) is assigned to a particular class for which the average value of lower and upper

approximations is highest. The assigned $ClassLabel(t)$ of test pattern (t) is determined as follows:

$$ClassLabel(t) = \arg\max_j \left(\frac{(R \downarrow C_j)(t) + (R \uparrow C_j)(t)}{2} \right); \forall t \quad (7)$$

C. Vaguely Quantified Fuzzy-rough Nearest Neighbour Classifier

Vaguely Quantified Fuzzy-rough Nearest Neighbour (VQFRNN) [11] is an extension of FRNN method. It uses vague quantifier to replace the Lukasiewicz implicator (*I*) and the Lukasiewicz *t*-norms (*T*) of traditional lower and upper approximations of a rough set.

Detailed procedure of VQFRNN classifier is provided below:

1. Compute the *k*-nearest neighbour (*kNN*) labeled patterns closest to each of the test pattern (t) based on Euclidean distance (to compute the distance from the labeled pattern to test pattern).
2. The values of lower and upper approximations of test pattern (t) for belonging to each class *C* is calculated respectively as follows:

$$(R \downarrow_{Q_l} C)(t) = Q_l \left(\frac{\sum_{y \in kNN} \min(R(t,y), C(y))}{\sum_{y \in kNN} R(t,y)} \right) \quad (8)$$

$$(R \uparrow_{Q_u} C)(t) = Q_u \left(\frac{\sum_{y \in kNN} \min(R(t,y), C(y))}{\sum_{y \in kNN} R(t,y)} \right) \quad (9)$$

where $R(t,y)$ and $C(y)$ are computed using Equation (5) and (6) respectively. $Q_l = Q_{(0.1,0.0)}$ and $Q_u = Q_{(0.2,1.0)}$ are computed respectively as follows:

$$Q_{(\alpha,\beta)}(x) = \begin{cases} 0, & x \le \alpha \\ \frac{2(x-\alpha)^2}{(\beta-\alpha)^2}, & \alpha \le x \le \frac{\alpha+\beta}{2} \\ 1 - \frac{2(x-\beta)^2}{(\beta-\alpha)^2}, & \frac{\alpha+\beta}{2} \le x \le \beta \\ 1, & \beta \le x. \end{cases} \quad (10)$$

$$for \quad 0 \le \alpha < \beta \le 1.$$

3. Test pattern (t) is assigned to a particular class for which the sum of lower and upper approximation value is highest.

$$ClassLabel(t) = \arg\max_j \left((R \downarrow_{Q_l} C_j)(t) + (R \uparrow_{Q_u} C_j)(t) \right); \forall t \quad (11)$$

D. Ordered Weighted Average based Fuzzy-rough Nearest Neighbour Classifier

Ordered Weighted Average based Fuzzy-rough Nearest Neighbour (OWAFRNN) [12] classifier uses $OWA_{\min}$ and $OWA_{\max}$ weight to replace the infimum and supremum operators of traditional lower and upper approximations .

Detailed procedure of OWAFRNN classifier is provided below:

1. Compute the *k*-nearest neighbour (*kNN*) labeled patterns closest to each of the test pattern (t) based on Euclidean distance (to compute the distance from the labeled pattern to test pattern).
2. The values of lower and upper approximations of test pattern (t) for belonging to each class *C* is calculated respectively as follows:

$$(R \downarrow_{OWA} C) = OWA_{\min} I(R(t,y), C(y)); \quad (12)$$

$$(R \uparrow_{OWA} C) = OWA_{\max} T(R(t,y), C(y)); \quad (13)$$

where *I* is the Lukasiewicz implicator, *T* is the Lukasiewicz *t*-norms and $R(t,y)$ and $C(y)$ are computed using Equation (5) and (6) respectively. $OWA_{\min}$ and $OWA_{\max}$ weights are computed as:

$$OWA_{\min} = \left\langle \frac{2}{p(p+1)}, \frac{4}{p(p+1)}, \ldots\ldots, \frac{2p}{p(p+1)} \right\rangle \quad (14)$$

$$OWA_{\max} = \left\langle \frac{2p}{p(p+1)}, \frac{2(p-1)}{p(p+1)}, \ldots\ldots, \frac{2}{p(p+1)} \right\rangle \quad (15)$$

where $p = |kNN|$.

3. The test pattern (t) is assigned to a particular class for which the sum of lower and upper approximation value is highest.

$$ClassLabel(t) = \arg\max_j \left((R \downarrow_{OWA} C_j)(t) + (R \uparrow_{OWA} C_j)(t) \right); \forall t \quad (16)$$

IV. Results and Discussion

In this section, we provide the details of microarray gene expression cancer datasets used for the experiments followed by the performance evaluation measures. Finally, Experimental results and analysis of the results are summarized.

A. Description of Datasets

In this article, we have used eight real life microarray gene expression cancer datasets namely, Colon Cancer, Brain tumor, SRBCT, Lymphoma, Prostate Cancer, Ovarian Cancer, Leukemia, Lung Cancer datasets. These datasets are publicly available at www.stat.ethz.ch/dettling/bagboost.html

[13] and http://datam.i2r.astar.edu.sg/datasets/krbd/index.html [14]. The dataset is a collection of the samples and each sample consist of gene expression values and their class label information. Brief descriptions of the used datasets are provided below.

Colon Cancer dataset contains 40 samples of cancerous patients and 22 samples of normal patients. There are 2000 genes in each sample.

Brain Tumor dataset is having 42 samples distributed in 5 classes of brain tumor viz., medulloblastomas, malignant gliomas, atypical teratoid/rhabdoid tumors, primitive neuroectodermal tumors, human cerabella. Numbers of samples for these classes are 10, 10, 10, 8 and 4 respectively. The expression profile contains 5597 genes.

Small round blue cell tumors (SRBCT) dataset consists of 63 samples. Among them, 12 samples of neuroblastoma (NB), 20 samples of rhabdomyosarcoma (RS), 8 samples of Burkitt's lymphoma (BL) and 23 samples of Ewing's sarcoma (ES). Each sample comprises of 2308 genes expression values.

Lymphoma dataset consists of 62 samples and each sample is having 4026 genes. There are 3 classes of lymphoma viz., diffuse large B-cell lymphoma, follicular lymphoma and chronic lymphocytic leukemia.

Prostate cancer dataset comprises of 102 samples in which 52 observations are from prostate cancer tissues and 50 are from normal patients. Each observation contains expression values for 6033 genes.

Ovarian cancer dataset contains 203 samples in which 91 samples are normal and 162 samples are cancerous. There are 15154 genes in each sample.

Leukemia dataset is having 72 samples distributed in two classes namely, lymphoblastic leukemia and myeloid leukemia. Number of genes present in each sample is 3571.

Lung Cancer dataset contains 203 samples in which 139 samples of lung adenocarcinomas, 20 samples of pulmonary carcinoids, 21 samples of squamous cell lung carcinomas, 6 samples of small-cell lung carcinomas and 17 normal lung samples. Each sample contains expression values of 12600 genes.

The summary of the datasets used for the experiments is provided in Table 1.

Table 1. Summary of eight microarray gene expression datasets used for the experiments.

Datasets	**Samples (patterns)**	**Genes**	**Subtype (classes)**
Colon Cancer	62	2000	2
Brain Tumor	42	5597	5
SRBCT	63	2308	4
Lymphoma	62	4026	3
Prostate cancer	102	6033	2
Ovarian cancer	253	15154	2
Leukemia	72	3571	2
Lung Cancer	203	12600	5

B. Performance evaluation measures

Six different kinds of validity measures are used to assess the performance of the presented classifiers namely, (i) percentage accuracy, (ii) precision, (iii) recall, (iv) macro averaged F_1 measure, (v) micro averaged F_1 measure [15] and (vi) kappa [16].

C. Experimental set up

In this article, we have repoted the average results of 10 simulation runs of all the methods performed on eight real life microarray gene expression datasets. All the methods used in this article are implemented in MATLAB and executed in Windows 7 machine with processor speed 2.40 GHz and main memory 4 GB. Training set consists of two samples from each class present in the dataset and test set comprises of the total samples available (in the datasets) excluding the training samples.

D. Experimental Results and Analysis

The average experimental results of 10 simulation runs (on random selection of labelled / training patterns) in terms of percentage accuracy, precision, recall, macro F_1, micro F_1 and kappa obtained by all the methods (viz., FKNN, FRNN, VQFRNN and OWAFRNN) performed on eight microarray gene expression datasets are reported in Table 2. Best results are shown in bold font in the Table 2. The standard deviations of accuracies of 10 simulations are also shown using ± sign corresponding to each percentage accuracy in Table 2

It is seen from the Table 2, that the FRNN method performed better in terms all the validity measures (viz., accuracy, overall precision, overall recall, macro averaged F_1 measure, micro averaged F_1 measure and kappa) over other methods namely, FKNN, VQFRNN and OWAFRNN for five datasets (viz., Brain tumor, SRBCT, Lymphoma, Prostate Cancer and Leukemia). Whereas, VQFRNN method achieved better accuracies over other methods (viz., FKNN, FRNN and OWAFRNN) for three datasets namely, Colon Cancer, Ovarian Cancer and Lung Cancer.

Table 2. Summary of the average experimental results (in terms of accuracy, precision, recall, macro F_1 , micro F_1 and kappa) of 10 simulations obtained by different methods viz., FKNN, FRNN, VQFRNN and OWAFRNN performed on eight microarray gene expression datasets.

Datasets	Methods	Accuracy (%)	Overall Precision	Overall Recall	Macro F_1	Micro F_1	Kappa
Colon Cancer	FKNN	80.69 ± 8.28	0.8467	0.8237	0.8029	0.8350	0.6255
	FRNN	90.86 ± 4.74	0.9078	**0.9128**	0.9006	0.9098	0.8040
	VQFRNN	**91.03** ± 5.73	**0.9257**	0.9054	**0.9061**	**0.9153**	**0.8153**
	OWAFRNN	90.52 ± 6.20	0.9158	0.9056	0.9001	0.9105	0.8039
Brain Tumor	FKNN	67.81 ± 7.66	0.6692	0.7901	0.6433	0.7224	0.5812
	FRNN	**82.77** ± 8.24	**0.8227**	**0.8648**	**0.7914**	**0.8423**	**0.7772**
	VQFRNN	77.71 ± 4.79	0.7632	0.8279	0.7555	0.7940	0.7103
	OWAFRNN	80.05 ± 4.19	0.7836	0.8387	0.7692	0.8091	0.7407
SRBCT	FKNN	71.45 ± 4.37	0.7918	0.7727	0.7140	0.7818	0.6239
	FRNN	**83.09** ± 5.56	**0.8586**	**0.8197**	**0.8129**	**0.8386**	**0.7678**
	VQFRNN	77.82 ± 7.26	0.8299	0.7879	0.7603	0.8081	0.7038
	OWAFRNN	81.09 ± 6.37	0.8566	0.8040	0.7947	0.8294	0.7450
Lymphoma	FKNN	96.25 ± 1.01	0.9786	0.9218	0.9474	0.9493	0.9202
	FRNN	**97.33** ± 1.26	**0.9875**	**0.9431**	**0.9630**	**0.9647**	**0.9431**
	VQFRNN	94.29 ± 2.02	0.9340	0.8873	0.9044	0.9099	0.8785
	OWAFRNN	96.43 ± 1.19	0.9833	0.9261	0.9516	0.9538	0.9241
Prostate cancer	FKNN	67.55 ± 10.89	0.6736	0.7444	0.6425	0.7047	0.3471
	FRNN	**86.12** ± 7.96	**0.8613**	**0.8738**	**0.8594**	**0.8675**	**0.7224**
	VQFRNN	80.92 ± 8.60	0.8076	0.8550	0.7985	0.8304	0.6163
	OWAFRNN	85.00 ± 5.61	0.8499	**0.8738**	0.8472	0.8615	0.6997
Ovarian cancer	FKNN	87.07 ± 7.50	0.8563	0.8704	0.8525	0.8626	0.7100
	FRNN	90.76 ± 7.04	0.9149	0.9145	0.9027	0.9144	0.8101
	VQFRNN	**92.25** ± 4.36	**0.9242**	**0.9230**	**0.9166**	**0.9232**	**0.8347**
	OWAFRNN	89.40 ± 4.92	0.9095	0.8944	0.8887	0.9015	0.7814
Leukemia	FKNN	75.59 ± 5.77	0.7879	0.7668	0.7482	0.7772	0.5162
	FRNN	**81.76** ± 11.95	**0.8408**	**0.8356**	**0.8106**	**0.8381**	**0.6425**
	VQFRNN	65.44 ± 13.95	0.6836	0.6878	0.6351	0.6846	0.3366
	OWAFRNN	76.91 ± 7.73	0.7395	0.7956	0.7290	0.7637	0.4770
Lung Cancer	FKNN	61.81 ± 8.43	0.7895	0.6061	0.6070	0.6852	0.4414
	FRNN	68.94 ± 7.46	**0.8300**	**0.6364**	**0.6613**	**0.7195**	0.5147
	VQFRNN	**69.42** ± 11.53	0.7854	0.6339	0.6384	0.6999	**0.5243**
	OWAFRNN	65.76 ± 14.41	0.7751	0.6064	0.6133	0.6787	0.4903

Results of investigation (in terms of percentage accuracy) are also statistically validated using the paired *t*-test [17] performed with the best method FRNN versus other methods at 5% level of significance. Paired *t*-test results obtained by the FRNN method versus other methods in terms of *p*-score are reported in Table 3. Statistically significant results of paired *t*-test are marked as bold for *p*-score values those are less than 0.05 (at 5% label of significance) indicating that the null hypothesis is rejected. That means there exists statistically significant difference in the results (in terms of accuracy) obtained by the two methods. Significant improvement of the FRNN method compared to the other methods is indicated by up-arrow (↑) in the Table 3.

Table 3. Result of paired *t*-test performed on accuracies obtained by the FRNN method versus other methods in terms of *p*-score for eight microarray gene expression datasets.

Datasets	FRNN Vs. FKNN	FRNN Vs. VQFRNN	FRNN Vs. OWAFRNN
Colon Cancer	**0.0100**↑	0.9543	0.8693
Brain Tumor	**0.0033**↑	0.0880	0.3828
SRBCT	**0.0004**↑	0.1252	0.4528
Lymphoma	0.0500	**0.0012**↑	0.1729
Prostate cancer	**0.0004**↑	0.2184	0.7731
Ovarian cancer	0.0934	0.5138	0.5321

Leukemia	0.1541	**0.0049**↑	0.2807
Lung Cancer	0.1217	0.9140	0.5449

V. CONCLUSION AND FUTURE SCOPE

Cancer subtype classes are usually overlapping and indiscernible in nature that can be handled by the fuzzy-rough set theory. Therefore, in this article an empirical study of cancer classification from microarray gene expression data using different types of fuzzy-rough nearest neighbour based classifiers is presented. The effectiveness of the presented methods are tested using eight real life microarray gene expression cancer datasets in terms of different validity measures viz., accuracy, precision, recall, F_1-measures and kappa. It is observed from the experimental results that the FRNN method performed better in terms all the validity measures (viz., accuracy, overall precision, overall recall, macro averaged F_1 measure, micro averaged F_1 measure and kappa) for five datasets namely, Brain tumor, SRBCT, Lymphoma, Prostate cancer and Leukemia. Whereas, VQFRNN method performed better in terms of accuracy for three datasets namely, Colon cancer, Ovarian cancer and Lung cancer.

The encouraging results obtained from the presented methods motivate us to develop some more classifier based on fuzzy-rough in future. The presented methods may also be tested on other microarray gene expression cancer datasets in future.

REFERENCES

[1] D. Stekel, *"Microarray Bioinformatics",* 1st ed., Cambridge, Cambridge University Press, UK, 2003.

[2] M. Dettling and P. Buhlmann, *"Boosting for tumor classification with gene expression data"*, Bioinformatics, Vol. 19, Issue. 9, pp.1061–1069, 2003.

[3] D. Jiang, C. Tang and A. Zhang, *"Cluster analysis for gene expression data: A survey"*, IEEE Transactions on Knowledge and Data Engineering, Vol.16, Issue.11,pp.1370–1386, 2004.

[4] R. Priscilla and S. Swamynathan, *"A semi-supervised hierarchical approach: two-dimensional clustering of microarray gene expression data"*, Frontiers of Computer Science, Vol.7, Issue.2, pp. 204–213, 2013.

[5] A. Halder and S. Misra, *"Semi-supervised fuzzy k-NN for cancer classification from microarray gene expression data"*, in Proceedings of the 1st International Conference on Automation, Control, Energy and Systems (ACES 2014) (IEEE Computer Society Press) pp.1–5,2014.

[6] D. Du, K. Li, X. Li, and M. Fei, *"A novel forward gene selection algorithm for microarray data,"* Neurocomputing, vol. 133, pp. 446–458, 2014.

[7] L. Zadeh, *"Fuzzy sets"*, Information and Control, Vol.8, Issue.3, pp.338–353, 1965.

[8] Z. Pawlak, *"Rough sets"*, International Journal of Computer and Information Science, Vol.11, Issue.5, pp.341–356, 1982.

[9] A.M. Radzikowska and E.E. Kerre, *"A comparative study of fuzzy rough sets"*, Fuzzy Sets and Systems, Vol.126, pp.137–156, 2002.

[10] J.M. Keller, M.R. Gray and J.A. Givens, *"A fuzzy K -nearest neighbor algorithm"*, IEEE Transactions on Systems, Man and Cybernetics, Vol.15, Issue.4, pp. 580–585, 1985.

[11] R. Jensen and C. Cornelis, *"A new approach to fuzzy-rough nearest neighbour classification"*, in: Proceedings of the 6th International Conference on Rough Sets and Current Trends in Computing, pp. 310–319, 2008.

[12] R. R. Yager, *"On ordered weighted averaging aggregation operators in multicriteria decision making"*, IEEE Transaction on Systems, Man and Cybernetics, Vol.18, pp.183-190, 1988.

[13] M. Dettling, *"Bagboosting for tumor classification with gene expression data",* Bioinformatics, Vol.20, Issue.18, pp.583–593, 2004.

[14] Technology Agency for Science and Research. Kent ridge bio-medical dataset repository.

[15] A. Halder, S. Ghosh, and A. Ghosh. "*Aggregation pheromone metaphor for semi-supervised classification*", Pattern Recognition, Vol.46, Issue.8, pp.2239–2248, 2013.

[16] J. Cohen, "*A coefficient of agreement for nominal scales*", Educational and Psychological Measurement, Vol.20, Issue.1 pp. 37–46, 1960.

[17] E. Kreyszig, *"Introductory Mathematical Statistics"*, 1st ed, John Wily, 1970.

Authors Profile

Mr.Ansuman Kumar is presently working as an Assistant Professor in Department of Computer Applications, North-Eastern Hill University (NEHU), Tura Campus, Meghalaya, India. He is also presently working towards Ph.D. from NEHU. He received the M. Tech in Information Technology from the Tezpur University (A Central University), Assam, India, in 2008. He has 6 years of teaching experience and 2 years of research experience.

Dr. Anindya Halder is presently working as an Assistant Professor in Department of Computer Applications, North-Eastern Hill University, Tura Campus, Meghalaya, India. He received the Master of Computer Application (MCA) from University of Kalyani, India, in 2005 and Ph.D. in Engineering from Jadavpur University, Kolkata, India, in 2012. He worked (towards Ph.D.) as a Research Scholar at Center of Soft Computing Research, Indian Statistical Institute (ISI), Kolkata, India during August 2007 to July 2012. Dr. Halder was a Visiting Scientist at Remote Sensing Laboratory, University of Trento, Italy during January 2010 to June 2010. He has published a number of research articles in internationally reputed journals and refereed conferences. His research interests include, machine learning, pattern recognition, swarm intelligence, soft computing, remote sensing image analysis, and bioinformatics.

