International Journal of Research in Advent Technology, Vol.7, No.5, May 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Ensemble based Fuzzy-Rough Nearest Neighbor Approach for Classification of Cancer from Microarray Data

Ansuman Kumar[1], Anindya Halder[2]
[1]*Dept. of Computer Application, North-Eastern Hill University, Tura Campus, Meghalaya 794002, India.*
[2]*Corresponding Author's Email: anindya.halder@gmail.com*

Abstract- Cancer sample classification from gene expression data is one of the challenging areas of research in biomedical engineering and machine learning. In gene expression data, labeled samples are very limited in comparison to unlabeled samples; and labeling of unlabeled data is costly. Therefore, single classifier trained with limited training samples often fails to produce required accuracy. In such situation, ensemble technique can be effective as it combines the results of individual classifier which can improve the cancer classification accuracy. In this article a novel *ensemble based fuzzy-rough nearest neighbour* (EnFRNN) for cancer sample classification from microarray gene expression data is proposed. The proposed method is able to deal the uncertainty, overlapping and indiscernibility generally present in cancer subtype classes of the gene expression data. The proposed ensemble classifier is tested on eight publicly available microarray gene expression datasets. Experimental results suggest that the performance of the proposed ensemble classifier provides better results in comparison to individual classifier for cancer classification from gene expression data. In summary, fuzzy-rough based ensemble learning method turns out to be very effective in cancer sample classification from gene expression data particularly when the individual classifier result is not up to the mark with limited training samples.

Index Terms- Cancer classification; Ensemble technique; Microarray gene expression data; Fuzzy set; Rough set.

1. INRODUCTION

Traditional clinical methods for cancer sample classification rely on the clinical findings and the morphological exhibition of the tumor. These techniques are costly and time consuming. The recent 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 have been 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. Generally, 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 and overlapping in nature. Therefore, the traditional classifiers often fail to achieve required accuracy. In this circumstance, the ensemble technique [7] is supposed to be useful as it judiciously combines the predications of the individual classifiers to produce the final decision which are expected to be better than any individual classifier.

Ensemble technique is the learning model that achieves performance by combining the opinions of multiple base classifiers [7]. Ensemble technique uses many base classifiers, and combines their opinions in such a way that the combination result will improve the performance compared to any individual classifier [7]. The heterogeneity among the base classifiers and diversity in the training data set are the basic ideas to success of ensemble technique. Varieties of popular ensemble algorithms are proposed in the literature, viz., Bagging, Boosting, AdaBoost, and Random Forest [8].

Ensemble methods have the ability to deal with small sample size and high dimensionality. Therefore, ensemble methods have been widely applied to microarray gene expression data. A notable review of ensemble methods applied in bioinformatics may be found in [9]. Several pioneered work to classify cancer from the microarray gene expression data are proposed. Dettling and Buhlmann [10] proposed boosting for tumor classification with gene expression data. Osareh and Shadgar [11] provided an efficient ensemble learning method using RotBoost ensemble methodology. Valentini et al. [12] introduced bagged ensembles of support vector machines for cancer recognition.

However, those ensemble methods are not able to handle the uncertainty, ambiguity, overlappingness and vagueness often present in the gene expression data. Therefore, in this work an ensemble technique using fuzzy-rough nearest neighbour (EnFRNN) is proposed for cancer sample classification from gene expression data (to improve the prediction accuracy of any individual classifier) which can handle the possible presence of uncertainty, ambiguity, vagueness, indiscernibility, overlappingness in the cancer subtype classes.

The remainder of the article is structured as follows. The background theory related to this article is

briefly illustrated in Section 2. Section 3 presents a detailed description of the proposed EnFRNN method. Details of the experiments and analysis of the results are provided in Section 4. Finally, conclusions are given in Section 5.

2. PRELIMINARY STUDY

Fuzzy-rough nearest neighbour classifier uses the concept of fuzzy set and rough set. Thus brief outline of those are provided below.

2.1. Fuzzy set theory

L. Zadeh [13] proposed Fuzzy set theory in the year 1965. It is an expansion 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*.

2.2. Rough set theory

Rough set theory was introduced by Z. Pawlak [14] 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 ordered pair < *X*, *R* >, where *X* is the non-empty universe of discourse and *R* is an equivalence relation defined on *X*. *R* satisfies the reflexive, symmetric and transitive property. For each subset *A* of *X*, the lower approximation is 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*.

2.3. 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 [15]:

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

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

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 the *infimum* and sup represents the *supremum*.

3. ENSEMBLE BASED FUZZY-ROUGH NEAREST NEIGHBOUR CLASSIFIER

As mentioned in the introduction the objective of the present work is to develop a robust ensemble classifier which can handle the ambiguity, vagueness, uncertainty, overlapping and indiscernibility. Therefore in this work fuzzy-rough nearest neighbour (FRNN) [16] is used as a base classifier to judiciously combine the results of prediction to improve the accuracy. Here fuzzy set deals the vagueness, ambiguity and rough set deals the uncertainty, incompleteness and indiscernibility present in the gene expression data. The detailed description of the proposed ensemble technique is given below:

Let, $= \{< l_j \ , d_j > | \ j = 1 \ to \ |L| \ and \ d_j \in C\}$ be the training set which is divided using bootstrap method to generate three subsets L_1, L_2 and L_3. These subsets L_1, L_2 and L_3 act as the training sets for the three individual FRNN classifiers.

The details of the base classifier FRNN are described 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:

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

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

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

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

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:

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

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:

$ClassLabel \ (t) =$

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

The individual base classifiers are tested with test set $T = \{t_i \ \forall i = 1 to \ |T|\}$. The predicted class labels of each test pattern (t_i) are then combined using majority voting process to get the final prediction.

The complete procedure of the proposed ensemble based fuzzy-rough nearest neighbour (EnFRNN) method is shown in Figure 1.

4. RESULTS AND DISCUSSION

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

4.1 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 [17] and http://datam.i2r.astar.edu.sg/datasets/krbd/index.html [18]. The dataset is a collection of the samples and each sample consists of gene expression values and their class label information. Brief descriptions of the used datasets are provided below.

Colon Cancer dataset is having 40 samples of cancerous patients and 22 samples of normal patients. Each sample contains expression values of 2000 genes.

Brain Tumor dataset contains 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. There are 5597 genes in each sample.

Small round blue cell tumors (SRBCT) dataset consists of 63 samples. Among them, 12 samples are of neuroblastoma (NB), 20 samples are of rhabdomyosarcoma (RS), 8 samples are of Burkitt's lymphoma (BL) and 23 samples are 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 contains 102 samples in which 52 observations are from prostate cancer

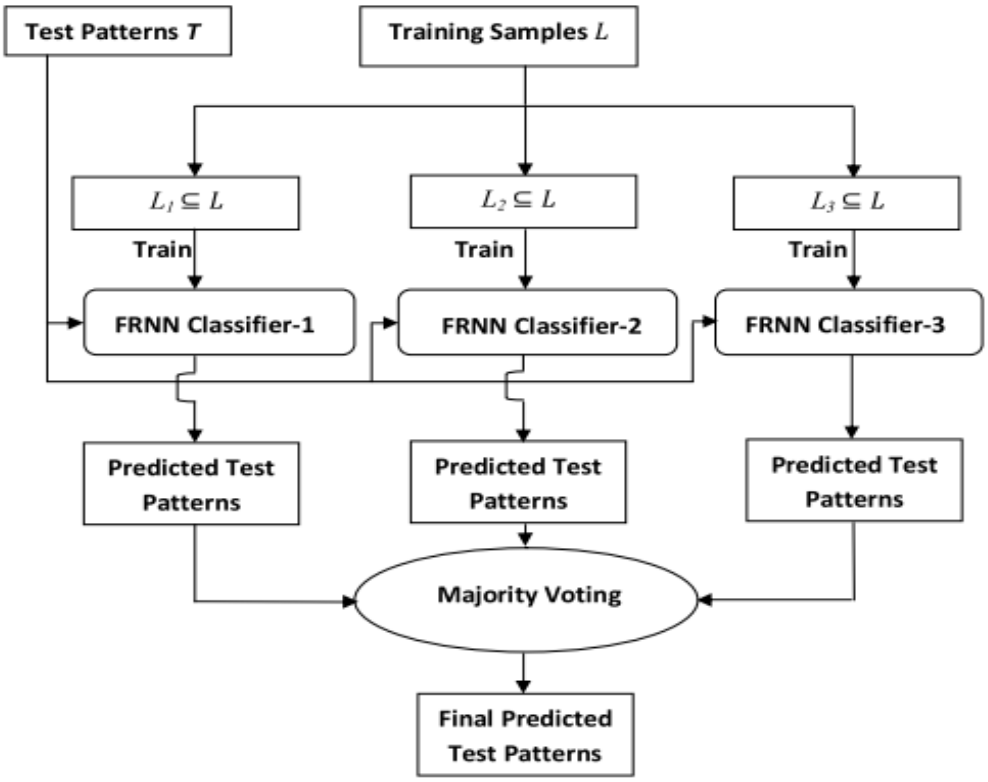

Figure 1. Block diagram of the ensemble based fuzzy-rough nearest neighbour (EnFRNN) method.

tissues and 50 are from normal patients. The expression profile contains 6033 genes.

Ovarian cancer dataset consists of 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	Genes	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

4.2 *Comparison with others methods*

The performance of the proposed EnFRNN method is compared with two methods namely, Fuzzy *k*-Nearest Neighbour (FKNN) [19] and Fuzzy-Rough Nearest Neighbour (FRNN) [16] classifiers.

Fuzzy k- Nearest Neighbour Classifier

Fuzzy *k*-Nearest Neighbour (FKNN) [19] 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 vague. In these cases fuzzy *k-NN* algorithm is expected to provide better result than *k-NN*.

Fuzzy-Rough Nearest Neighbour Classifier

Fuzzy-Rough Nearest Neighbour (FRNN) [16] classifier is the combination of fuzzy and rough sets 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 a test pattern.

4.3 Performance evaluation measures

In this article, we have used six different kinds of validity measures namely, (i) percentage accuracy, (ii) precision, (iii) recall, (iv) macro averaged F_1 measure, (v) micro averaged F_1 measure [20] and (vi) kappa [21] to assess the performance of the methods.

4.4 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 and the proposed EnFRNN) 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 $\pm$ sign corresponding to each percentage accuracy in Table 2. It is seen from the Table 2 that the proposed EnFRNN method performed better in terms all the validity measures over other methods namely, FKNN and FRNN for all the datasets experimented.

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 achieved by different methods viz., FKNN, FRNN, and proposed method EnFRNN 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
	EnFRNN	**96.85** ± 2.48	**0.9667**	**0.9661**	**0.9642**	**0.9662**	**0.9288**
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
	EnFRNN	**87.04** ± 3.59	**0.8691**	**0.8652**	**0.8323**	**0.8667**	**0.8296**
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
	EnFRNN	**89.15** ± 5.16	**0.9155**	**0.8552**	**0.8648**	**0.8841**	**0.8479**
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**
	EnFRNN	**97.40** ± 0.96	**0.9886**	0.9323	0.9566	0.9596	0.9368
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
	EnFRNN	**90.64** ± 3.84	**0.9069**	**0.9150**	**0.9058**	**0.9110**	**0.8130**
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
	EnFRNN	**95.26** ± 2.52	**0.9555**	**0.9486**	**0.9489**	**0.9519**	**0.8983**
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
	EnFRNN	**88.28** ± 7.04	**0.8933**	**0.8739**	**0.8741**	**0.8835**	**0.7533**
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
	EnFRNN	**73.66** ± 6.02	**0.8646**	0.6240	0.6572	**0.7244**	**0.5556**

5. CONCLUSIONS

Cancer subtype classes are usually overlapping and indiscernible in nature which can be handled by the fuzzy-rough set theory. Therefore, ensemble based fuzzy-rough nearest neighbour for cancer sample classification from gene expression data is proposed. Ensemble technique combines the predications of the individual classifier which improved the prediction accuracy compared to any individual classifier. Here, fuzzy and rough sets are able to handle vagueness, ambiguity, uncertainty, incompleteness and indiscernibility present in gene expression datasets.

The effectiveness of the proposed method is validated 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 EnFRNN 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 all the datasets investigated. Robustness of the EnFRNN method may further be tested on other kind of gene expression datasets such as microRNA in future.

REFERENCES

[1] D. Stekel, *"Microarray Bioinformatics",* 1st ed., Cambridge University Press, Cambridge, 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. I. Kuncheva. *"Combining Pattern Classifiers: Methods and Algorithms"*, John Wiley & Sons, 2nd ed., 2004.

[8] R. Polikar, "*Ensemble based systems in decision making*", IEEE Circuits and Systems Magazine, Vol. 6, Issue. 3, pp.21–45, 2006.

[9] P. Yang, Y. H. Yang, B. B. Zhou, and A. Y. Zomaya, "*A review of ensemble methods in bioinformatics*", Machine Learning, Vol. 5, Issue.4, pp.296–308, 2010.

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

[11] A. Osareh and B. Shadgar, *"An efficient ensemble learning method for gene microarray classification"*, BioMed Research International, Vol.2013, Issue.1, pp.1–10, 2013.

[12] G. Valentini, M. Muselli and F. Ruffino, *"Cancer recognition with bagged ensembles of support vector machines"*, Neurocomputing, Vol.56, pp.461–466, 2004.

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

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

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

[16] 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.

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

[18] Technology Agency for Science and Research, Kent ridge bio-medical dataset repository, <http:// datam.i2r.astar.edu.sg/datasets/krbd/ index.html>.

[19] 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.

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

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