Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 4 (2016) pp. 164-174
© MIR Labs, www.mirlabs.net/jnic/index.html

A comparative study of hybrid feature selection methods using correlation coefficient for microarray data

C.Arunkumar[1], S.Ramakrishnan[2]

[1]Assistant Professor (Senior Grade), Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore-641112, Tamilnadu, India
arunkumar.chinnaswamy@gmail.com

[2]Professor and Head, Department of Information Technology, Dr.Mahalingam College of Engineering and Technology, Pollachi-642003, Tamilnadu, India
ram_f77@yahoo.com

***Abstract*: Feature selection is a key challenge before the process of classification could be performed. The classification accuracy would increase by using a good feature selection method and also at the same time reduces the cost and time involved in the computation. In this study, we applied hybrid methods by using Correlation Based Feature Selection combined with different search algorithms. The classification performance was evaluated using fuzzy rough neural network classifier on the selected gene subsets. The experimental results reveal that majority of the hybrid method selects very few gene subsets and produces much better classification accuracy. The results are validated using traditional approaches like Precision, Recall, F-Score and Region of Characteristic.**

Keywords: Feature selection, Fuzzy Rough Set, correlation, greedy stepwise, particle swarm optimization

I. Introduction

A new area of research has blossomed in the last two decades in the area of machine learning and bioinformatics. This area is powered by the concept of microarray gene expression data. Disease identification and treatment of a wide variety of tumors in the oncology research is possible by using the gene expression information obtained from microarray samples. Microarray cancer gene expression data is composed of very small samples (usually less than two hundred) and thousands of gene expression levels (ranging from 7000-20000). A typical classification task involves two different kinds of problems. The first one would be a binary problem to identify and classify the given sample as “normal” or “cancerous”. The second one would be a multi-class problem that involves the identification and classification of a variety of tumors. Researchers across the globe express their serious concern about the presence of binary and multi-class problems in microarray gene expression. The presence of a very few number of training and testing samples and large amount of gene information is the sole reason behind this concern. Hence a robust model is needed to perform feature selection and classification. Another essential component is the validation of the data. The presence of noise and outliers also make the concept of microarray gene expression data even more exciting for researchers worldwide [1]. Microarray technology has the capability to perform a single experiment to monitor and measure the gene expression activation levels. The analysis and diagnosis of a large number of diseases is possible by using this approach. Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning (ML) methods [2]. The current focus is to perform efficient clustering and also increase the classifier accuracy. The correlation and interaction pattern of the gene expression data could be obtained by performing an efficient clustering. The main aim of research in the area of classification accuracy involves prediction of the class membership of the data, production of the correct label for the training data and predicting the labels of unknown data with higher degree of accuracy [3]. The training and testing of the different classification methods has become difficult because of the two key aspects of microarray data namely the small sample size and high dimensionality. Also, it might be required to investigate thousands of gene expression data though only a very small number might show significant correlation with a particular phenotype [4]. So feature selection is a very crucial procedure to understand and analyze the gene expression profiles and hence aid in achieving higher classification accuracy. The prediction of classification accuracy of unknown samples in a medical diagnosis system plays a major role is clinical applications.

A subset of optimized features from the given dataset is selected using suitable search operations using statistical estimates. The main challenge in bioinformatics is feature subset selection. This is due to the fact that only a very small sample size is available for high dimensional data. This “large p, small n” problem is called the curse of dimensionality. Many dimensionality reduction algorithms have been developed to avoid this phenomenon. Filter and wrapper approaches are the two broad categories of feature selection approaches in data mining.

In the filter model approach, the process of classification is performed after filtering. The weight value for each feature is computed and higher values are chosen to represent the reduced feature subset. The statistical properties of the data contribute majorly in the relevant feature selection process using the filter model. The dimensionality of the dataset is greatly reduced by employing the filter model as it is independent of the learning algorithm. The interaction between features is not considered in the filter approach and this is one major disadvantage of this model.

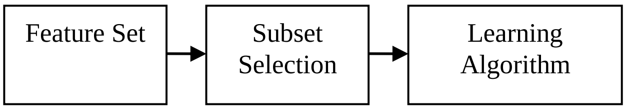


Figure 1. Working of a Filter

The working of a filter is depicted in Figure 1 as in. In the case of a filter approach, the filtering process is independent of the learning algorithm. This approach is suitable in most of the cases as it is independent of any particular algorithm

The wrapper model is applied on a subset of features obtained from the filter model. The subset features are estimated by using an evaluation function along with a learning algorithm. This model searches for an optimal solution in a given dimensional space by using an optimal algorithm [4]. The results of the wrapper model are validated using a suitable classification algorithm in a subset search space. The wrapper approach utilizes a given learning algorithm to evaluate the candidate feature subsets and hence is tied to the learning algorithm. Three main issues in a wrapper model make it challenging. They are search operation on a high dimensional space called the NP complete problem, uncertain assessments that make the choice of feature configuration difficult and the high dimensionality of a given problem that makes the selection of a feature subset complex.

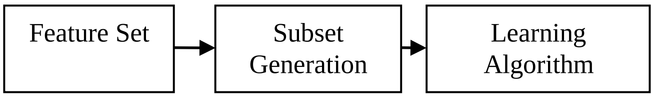


Figure. 2. Working of a Wrapper

The working of a wrapper is depicted in Figure 2 as in. In the case of a wrapper approach, the feature selection process is tied to the algorithm. This method searches through the feature subset space using the estimated accuracy from an induction algorithm as a measure of subset suitability. It involves the generation of a subset [6]. The commonly used gene selection & extraction approaches are t-test, Relief-F, information gain, SNRtest and principal component analysis (PCA), linear discriminant analysis, independent component analysis (ICA). These methods are capable of selecting a smaller subset of genes for sample classification [7].

In this study, we compared the gene selection performance of the hybrid methods that makes use of correlation based feature selection with suitable search approaches. To evaluate the performance of the hybrid feature selection methods, we used fuzzy rough neural network classifier to determine their influence on classification. The results indicate that in terms of the number of genes that need to be selected and classification accuracy, several hybrid methods are superior to other methods in the literature. The sequence of steps followed is depicted in Fig 3

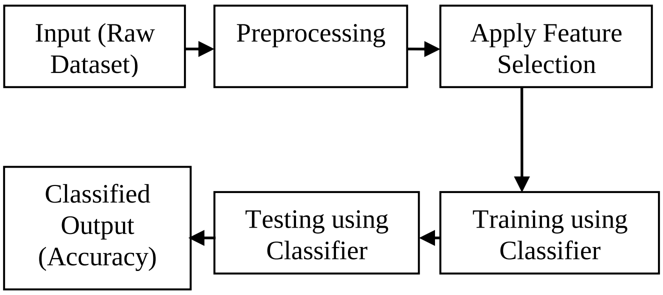


Figure. 3. Sequence of steps for Feature Selection and Classification

This paper is organized as follows: a brief overview introducing the methods is presented in Section II. The experimental framework and settings are described in Section III. Section IV summarizes the results obtained after feature selection and classification using different feature selection models. Finally, the conclusion and scope for further research is stated in Section V.

II. RELATED METHODS

A. Correlation-based Feature Selection

Correlation based heuristic evaluation function is used to rank the subset of genes in Correlation based feature selection by computing its coefficients. A subset of attributes is evaluated by considering the identification ability of each attribute. It overcomes the disadvantage of univariate filter approaches that does not take into account the interaction between features [8] [9]. The identification ability of each of the attributes is used to evaluate a subset of attributes. A multivariate approach is effective in identifying the correlation that exists among the different genes in the dataset [10]. Pearsons correlation coefficient is very sensitive to the presence of outliers and noise [10]. The relationship between variables (Genes) can be measured by the process of correlation [11]. The linear relationship between two variables is depicted by using the most common measure of correlation in statistics called the Pearson Product Moment Correlation. Formula for calculating Pearson correlation between features x_i and y_i is given in Eq 1

Correlation = ∑ (xi – mean (xi)*yi-mean (yi) / n*SD (xi)*SD (yi)) (1)

Pearson correlation coefficient between attributes is found out. Genes that possess low inter-correlation are selected [12]. The WEKA tool is used to implement CFS for selecting a subset of attribute gene information from a larger dataset. The selected genes were used to study the different types of cancer. The attributes exhibit high correlation if the value of correlation coefficient lies between 0.5 and 1 and is said to be less correlated if its value lies between 0.3 and 0.5. The common methods in CFS are best first, forward selection and backward elimination [11] [13] [14].

B. Greedy Stepwise Search

Greedy Stepwise Feature Selection starts with an empty "working" feature set and progressively adds features, one at a time, until a stopping criterion is reached. Greedy Stepwise operates in a very simple fashion [15]:

```
Step 1: At each step, consider all feature subsets which include the current "working" feature subset and exactly one feature not present in that set.
Step 2: Find the quality of each of these subsets, and then choose which of these gives the best performance to be the new "working" subset;
Step 3: Iterate this procedure until none of the new subsets improve performance.
Step 4: The final "working" subset (that is, the last subset which improved performance over its predecessor) is then given as the procedure's output.
```

C. Best First Search

Searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allowed controls the level of backtracking done. Best first may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions (by considering all possible single attribute additions and deletions at a given point)[16].

D. Combined Hill Climber

This Bayes Network learning algorithm uses a hill climbing algorithm adding, deleting and reversing arcs. The search is not restricted by an order on the variables (unlike K2). The difference with B and B2 is that this hill climber also considers arrows part of the naive Bayes structure for deletion [17] [18] [19].

E. Genetic Search

This Bayes Network learning algorithm uses genetic search for finding a well scoring Bayes network structure. Genetic search works by having a population of Bayes network structures and allow them to mutate and apply cross over to get offspring. The best network structure found during the process is returned [20].

F. Linear Forward Selection

Extension of BestFirst. Takes a restricted number of k attributes into account. Fixed-set selects a fixed number k of attributes, whereas k is increased in each step when fixed-width is selected. The search uses either the initial ordering to select the top k attributes, or performs a ranking (with the same evaluator the search uses later on). The search direction can be forward or floating forward selection (with optional backward search steps) [21] [22].

G. Particle Swarm Optimization Search

Performs a search using binary Particle Swarm Optimization. A number of particles are initialized at random locations (which correspond to feature subsets) and then swarm towards promising areas via the global best solution so far and each particle's local best. The smallest subset found overall with maximum quality is returned.

H. Subset Size Forward Selection

Extension of LinearForwardSelection. The search performs an interior cross-validation (seed and number of folds can be specified). A LinearForwardSelection is performed on each fold to determine the optimal subset-size (using the given SubsetSizeEvaluator). Finally, a LinearForwardSelection up to the optimal subset-size is performed on the whole data [21].

I. Linear Forward Fuzzy Rough Feature Selection

Linear Forward Fuzzy Rough Feature Selection selects only those features with gamma > 0. It performs a backward selection through the search space.

III. Experimental Framework

A. Hybrid filter and wrapper feature selection method

In this study, we used a hybrid of the filter and wrapper model methods to select feature genes in microarrays, and used four different feature selection algorithms to evaluate the performance of the proposed method. The filter model part correlation-based feature selection (CFS) is used to evaluate the ability of each feature which differentiates between different categories. The reasoning behind this method is that it can calculate the importance of each feature with respect to the class. Hybrid approaches are designed to eliminate the drawbacks in the filter and wrapper approaches. A combined filter-wrapper model makes up a hybrid model. The simplicity nature of the filter is combined with the optimized nature of the wrapper to build a hybrid model. The filter model aids in initial gene selection and the wrapper model helps to increase the classifier accuracy. The hybrid model is a two-staged model. The filter eliminates irrelevant and redundant genes from the original dataset in the first stage. The reduced gene information obtained in the first stage is given as the input to the second stage. In the second stage, the wrapper is applied on the filtered dataset and the training accuracy is optimized. This approach brings the hybrid model to an acceptable level of performance and satisfaction. The embedded approach that associates itself with a specific learning algorithm seeks to subsume feature selection as part of the model building process. The main goal of the hybrid model is to use the advantages of both the filter and wrapper models.

B. Correlation Based Feature Selection in different search spaces

As mentioned previously, filtering methods are amongst the most common methods for gene selection. These methods have low computational complexity and so can be used easily in large, high dimensional datasets such as microarrays; but these methods evaluate the discriminative power of each gene separately and the interaction of genes are ignored. Also these methods do not take into account the correlation among genes and so the selected gene set may have redundancy [23].

In this study, we created a hybrid approach of correlation based feature selection combined with several search strategies to select feature genes in microarrays. The different parameters used to perform the Feature Selection is tabulated as under in Table 1

Name of the Search Strategy	**Parameters used in Feature Selection**
Best First	Direction=Forward
	Loop Cache Size=1(default)
	Search termination=5(Number of Backtracking)
Combined Hill Climber	generateRanking=false
	numtoSelect=-1(default)-Retain all attributes
	searchBackwards=false(means do forward search)
	alpha=1.0
	threshold=1.0
Genetic	Cross over probability=0.6
	Max generations=20
	Mutation probability=0.033
	Population size=20
Greedy	generateRanking=false
	numtoSelect=-1(default)-Retain all attributes
	searchBackwards=false(means do forward search)
	threshold=-1.8(default)
Linear Forward Selection	performRanking=true(To select top ranked attributes)
	Loop Cache Size=1(default)
	Search termination=5(Number of Backtracking)
	Type=fixed-set
Particle Swarm Optimization	maxGenerations=50
	numParticles=100
	prune=false
Subset Size Forward Selection	performRanking=true(To select top ranked attributes)
	Loop Cache Size=1(default)
	numSubsetSizeCVFolds=5(cross validation)
	numUsedAttributes=50
	Type=fixed-set
Linear Forward Fuzzy Rough Feature Selection	alpha=0.2
	prune=false
	numtoSelect=-1(default)-Retain all attributes

Table 1. Parameters used for feature selection

IV. Results and Discussion

A. Preprocessing

Microarray gene expression data suffers from the problems of missing values due to several experimental reasons. The lymphoma dataset used for our study suffers from this problem. In order to solve this issue, preprocessing is performed on the raw dataset using the impute method. In this case, the missing values are treated using the ‘mode’ statistical operation wherein the missing values are filled with the value that occurs more often in the dataset. This imputed data is then subjected to feature selection and classification to achieve better classifier accuracy.

B. Dataset Description

We used three multi-class cancer-related human gene expression datasets, which were downloaded from [33] to evaluate the performance of the proposed method. The data format is shown in Table 2; it includes the data set name and the number of genes

Name of the dataset	Number of Genes in the raw dataset	Number of Classes
SRBCT	2308	4
Lymphoma	4026	3
MLL	12582	3

Table 2. Dataset and Number of Genes

The small round blue cell tumors (SRBCTs) are 4 different childhood tumors. They appear similar on routine histology. This makes the diagnosis of the disease an extremely challenging task. But this disease requires accurate diagnosis for deciding on the treatment options, evaluating the responses after the treatment and prognosis of the disease. They include Ewing's family of tumors (EWS), neuroblastoma (NB), non-Hodgkin lymphoma and rhabdomyosarcoma (RMS) [24].

The malignant cells in T-cell acute lymphoblastic leukemia (T-ALL) and T-cell lymphoblastic lymphoma (T-LL) are morphologically indistinguishable, and they share the expression of common cell surface antigens and cytogenetic characteristics. However, despite these similarities, differences in the clinical behavior of T-ALL and T-LL are observed [25].

Mixed-lineage leukemia (MLL) is a subset of human acute lymphoblastic leukemia’s with a chromosomal translocation involving the mixed-lineage leukemia gene. MLL translocations are typically found in infant leukemia and in chemotherapy-induced leukemia and have a particularly poor prognosis. The original research on this dataset suggested, that MLL have a highly uniform and distinct pattern that clearly distinguishes them from conventional acute lymphoblastic (ALL) or acute myeloid leukemia (AML) [27].

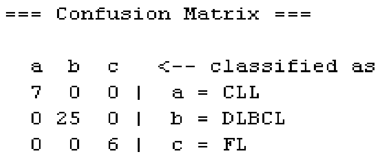
C. Classifier performance

The performance of the proposed method was evaluated by using selected feature gene subsets from microarray cancer gene expression data using fuzzy rough neural network classifier. The entire dataset was used for the purpose of training and testing by using 10-fold cross validation strategy.

In this study, we tested and compared the hybrid feature selection method’s performance on the classification of three multi-class cancer microarray expression data sets. This performance was evaluated on eight different hybrid approaches that used correlation based coefficient as the base technique. After feature selection, the selected feature subsets were evaluated using fuzzy rough neural network classifier using a 10-fold cross validation technique. In order to evaluate the performance of the classifier, the following parameters were used namely Accuracy, Precision, Recall, F-Measure and Region of Characteristic (ROC) Area [17]. In order to compute the above parameters, it is essential to define certain terminologies namely:

True Positive (t_p) – equivalent with hit
True Negative (t_n) – Correct rejection
False Positive (f_p) – False Alarm
False Negative (f_n) – Miss

The true positive, true negative, false positive and false negative could be computed easily by observing the confusion matrix. The sample confusion matrix is shown in the figure 4 below:



```
=== Confusion Matrix ===

  a  b  c   <-- classified as
  7  0  0 |  a = CLL
  0 25  0 |  b = DLBCL
  0  0  6 |  c = FL
```

Figure. 4. Sample Confusion Matrix

The formulae used to compute the Accuracy of the classifier is given in Eq 2:

$$\text{Accuracy} = (t_p+t_n) / (t_p+t_n+f_p+f_n) \quad (2)$$

The denominator value in Eq 2 is called the total population size

Precision and Recall are the two basic parameters used for evaluation in search strategies and based on understanding and measure of relevance. Precision also called the positive predictive value is the fraction of the retrieved instances that are relevant. Recall also called as sensitivity is the fraction of relevant instances that are retrieved [28] [29].

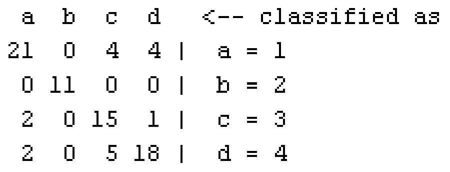
The formulae used to compute the Precision is given in Eq 3:

$$\text{Precision} = t_p/ (t_p+f_p) \quad (3)$$

The formulae used to compute the Recall also called Sensitivity is given in Eq 4:

Recall = t_p/ (t_p+f_n) (4)

The Precision and Recall could be easily computed from the confusion matrix. Consider a resultant sample confusion matrix for the SRBCT dataset obtained by applying the Fuzzy Rough Set Theory and Particle Swarm Optimization hybrid approach as given below in Figure 5:



```
 a  b  c  d   <-- classified as
21  0  4  4 |  a = 1
 0 11  0  0 |  b = 2
 2  0 15  1 |  c = 3
 2  0  5 18 |  d = 4
```

Figure. 5. Sample Confusion Matrix for computing Precision and Recall

The row total of the above confusion matrix in Figure 5 is R1 – 29, R2 – 11, R3 – 18 and R4 – 25. Similarly, the column total of the above confusion matrix is C1 – 25, C2 – 11, C3 – 24 and C4 – 23.

The Precision for Label a is computed using the formula in Eq 5

Precision (for label A) = TP_a/ (TP_a+FP_a)
(5)

where TP stands for True Positive and FP stands for True Negative [28].
Precision (for label A) = 21/R1 (29) = 0.724
Precision (for label B) = 11/R2 (11) = 1.0
Precision (for label C) = 15/R3 (18) = 0.833
Precision (for label D) = 18/R4 (23) = 0.783

After the Precision values are computed for each label, the average value is computed and is found to be 0.84 as tabulated in Table 4.

The Recall for Label a is computed using the formula in Eq 6

Recall (for label A) = TP_a/ (TP_a+FN_a) (6)

where TP stands for True Positive and FN stands for False Negative.
Recall (for label A) = 21/C1 (25) = 0.84
Recall (for label B) = 11/C2 (11) = 1.0
Recall (for label C) = 15/C3 (24) = 0.625
Recall (for label D) = 18/C4 (23) = 0.783

After the Recall values are computed for each label, the average value is computed and is found to be 0.78 as tabulated in Table 4. The F-Score is the harmonic mean of Precision and Sensitivity. In other words, F-Score or F-Measure in statistics is a measure of test's accuracy [31] [32]. It is computed using the formula

F-score = 2*(Precision*Recall)/(Precision+Recall) (7)
Table 3 shows the results of the various parameters computed for the Lymphoma Dataset using Fuzzy Rough Neural Network Classifier

FS Method	Raw Data Gene count	FS Gene Count	Accuracy (%)	Precision	Recall	F-Score	ROC Area
CFS+CHC		3	92.1	0.92	0.92	0.92	0.98
CFS+GREEDY		141	100	1	1	1	0.99
CFS+BEST FIRST		146	100	1	1	1	0.995
CFS+GENETIC	4026	1361	100	1	1	1	0.967
CFS+LFFRFS		1600	100	1	1	1	0.973

Table 3. Results for Lymphoma Dataset

Table 4 shows the results of the various parameters computed for the MLL Dataset using Fuzzy Rough Neural Network Classifier

FS Method	Raw Data Gene count	FS Gene Count	Accuracy (%)	Precision	Recall	F-Score	ROC Area
CFS+CHC		4	87.5	0.88	0.88	0.88	0.91
CFS+GREEDY		142	100	1	1	1	1
CFS+BEST FIRST		149	100	1	1	1	1
CFS+GENETIC		193	79.17	0.831	0.792	0.797	0.907
CFS+LFFRFS	12582	3438	93.06	0.932	0.931	0.93	0.97
CFS+LFS		91	100	1	1	1	1

Table 4. Results for MLL Dataset

Table 5 shows the results of the various parameters computed for the SRBCT Dataset using Fuzzy Rough Neural Network Classifier

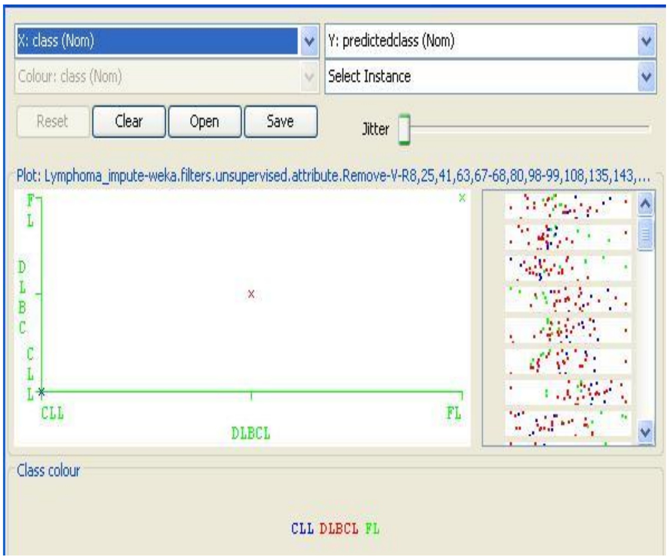
FS Method	Raw Data Gene count	FS Gene Count	Accuracy (%)	Precision	Recall	F-Score	ROC Area
CFS+CHC		83	100	1	1	1	0.996
CFS+GREEDY		112	100	1	1	1	0.998
CFS+BEST FIRST		111	100	1	1	1	0.998
CFS+GENETIC		124	84.34	0.852	0.843	0.844	0.83
CFS+LFFRFS	2308	366	98.8	0.989	0.98	0.988	0.989
CFS+LFS		77	100	1	1	1	1

Table 5. Results for SRBCT Dataset

In Table 3, 4 and 5, FS stands for Feature Selection, ROC stands for Region of Characteristic, CFS stands for Correlation Based Feature Selection, CHC for Combined Hill Climber, PSO for Particle Swarm Optimization, LFS for Linear Forward Selection, LFFRFS for Linear Forward Fuzzy Rough Feature Selection and SSFS for Subset Size Forward Selection.

With reference to Table 3 above, in the case of the Lymphoma dataset with 4026 genes in the raw dataset, our majority of the hybrid methods selects maximum of 0.07-7.6% (3 – 306 features) of the 4026 genes in the raw dataset and produces an accuracy of 100%. With reference to Table 4 above, in the case of MLL dataset, our majority of the hybrid methods selects 0.03-8.4% (4-1058 features) from the raw dataset with 12582 genes and produces a classifier accuracy of 100%. With reference to Table 5 above, in the case of SRBCT dataset, our majority of the hybrid methods selects 2.7-5.4% (63-124 features) from the entire raw data with 2308 genes and produces the highest classifier accuracy of 100%. Since the dataset involves the multi-class data, some feature selection methods selects about 25% of the total genes in order to produce an acceptable level of classifier accuracy.

The Classifier Errors could be visualized by plotting suitable graphs as depicted in Fig 6, 7 and 8 for Lymphoma, MLL and SRBCT datasets respectively.



X: class (Nom)

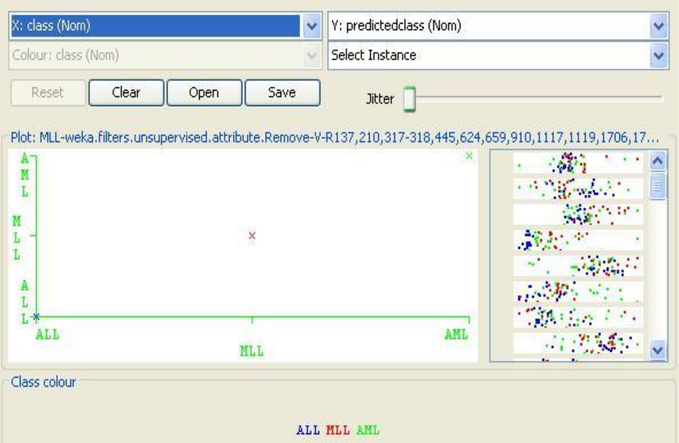
Y: predictedclass (Nom)

Colour: class (Nom)

Select Instance

Plot: Lymphoma_impute-weka.filters.unsupervised.attribute.Remove-V-R8,25,41,63,67-68,80,98-99,108,135,143,...

Figure. 6. Visualizing Classifier Accuracy for Lymphoma Dataset



X: class (Nom)

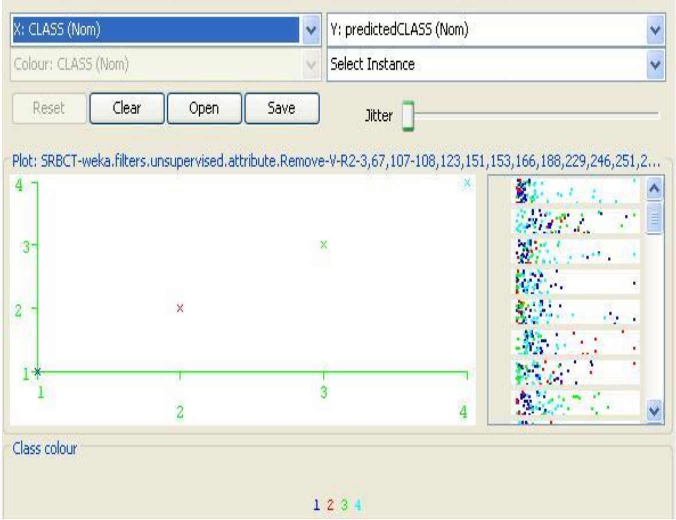
Y: predictedclass (Nom)

Colour: class (Nom)

Select Instance

Plot: MLL-weka.filters.unsupervised.attribute.Remove-V-R137,210,317-318,445,624,659,910,1117,1119,1706,17...

Figure. 7. Visualizing Classifier Accuracy for MLL Dataset



X: CLASS (Nom)

Y: predictedCLASS (Nom)

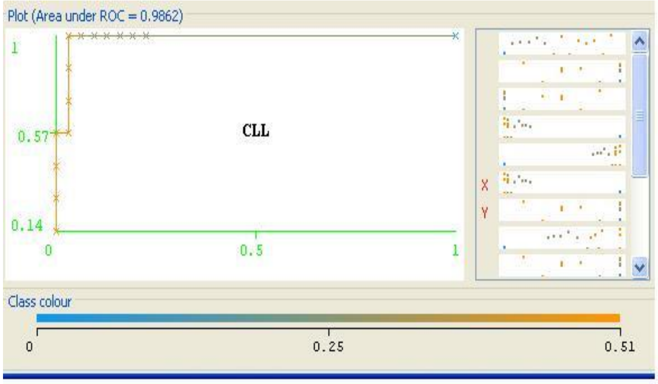
Colour: CLASS (Nom)

Select Instance

Plot: SRBCT-weka.filters.unsupervised.attribute.Remove-V-R2-3,67,107-108,123,151,153,166,188,229,246,251,2...

Figure. 8. Visualizing Classifier Accuracy for SRBCT Dataset

The Receiver Operating Characteristic (ROC) curve can be plotted for each of the datasets considering the False Positive Rate (FPR) along the X-Axis and True Positive Rate (TPR) along the Y-axis of the graph. The ROC plots for the three datasets namely Lymphoma, MLL and SRBCT is depicted in the below Figures 9 - 18



Plot (Area under ROC = 0.9862)

Figure. 9. ROC Plot for CLL (Lymphoma Dataset)

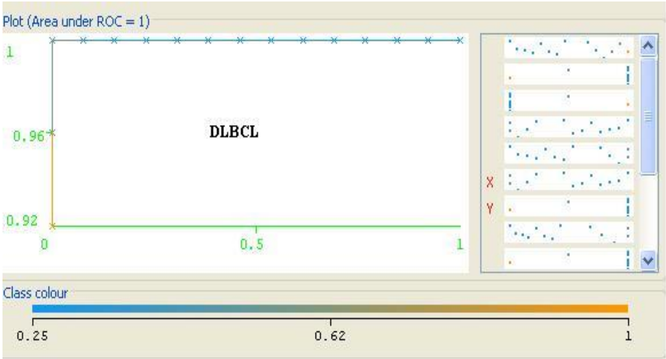


Figure. 10. ROC Plot for DLBCL (Lymphoma Dataset)

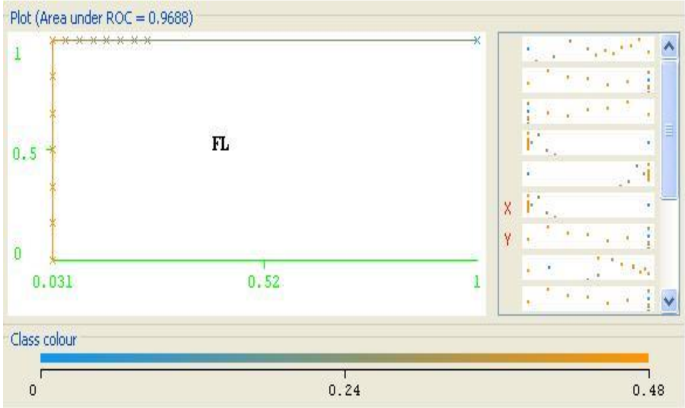


Figure. 11. ROC Plot for FL (Lymphoma Dataset)

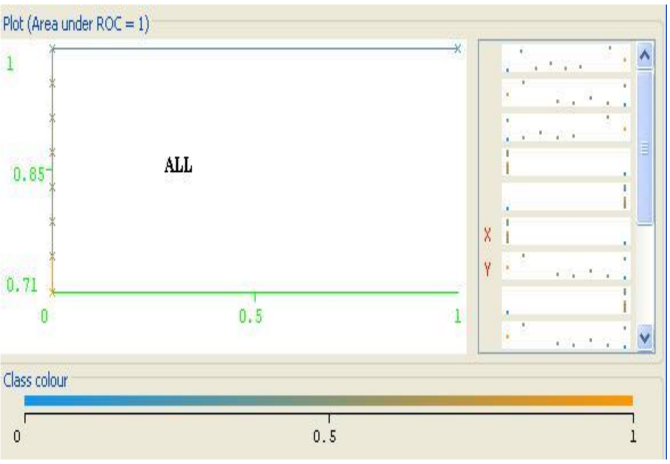


Figure. 12. ROC Plot for ALL (MLL Dataset)

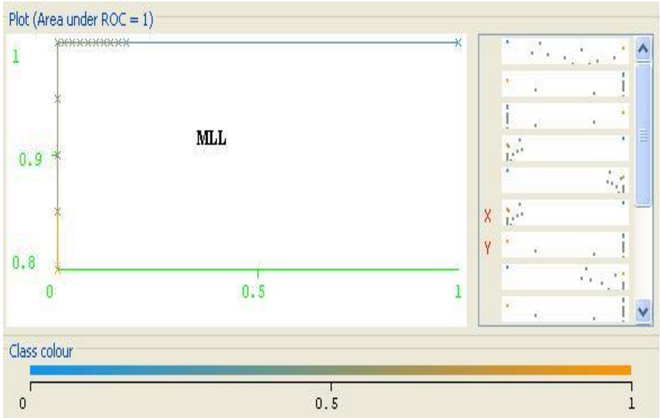


Figure. 13. ROC Plot for MLL (MLL Dataset)

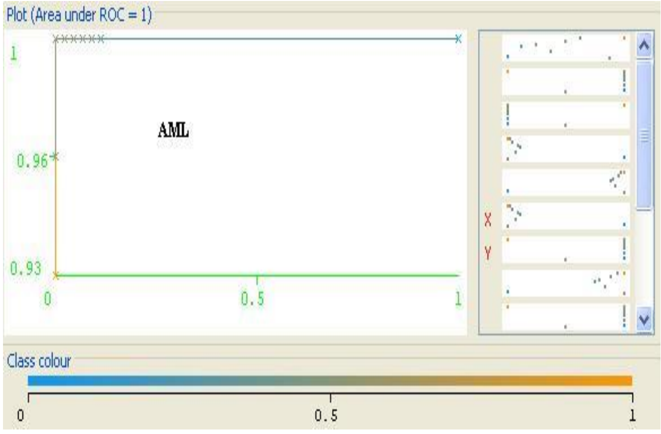


Figure. 14. ROC Plot for AML (MLL Dataset)

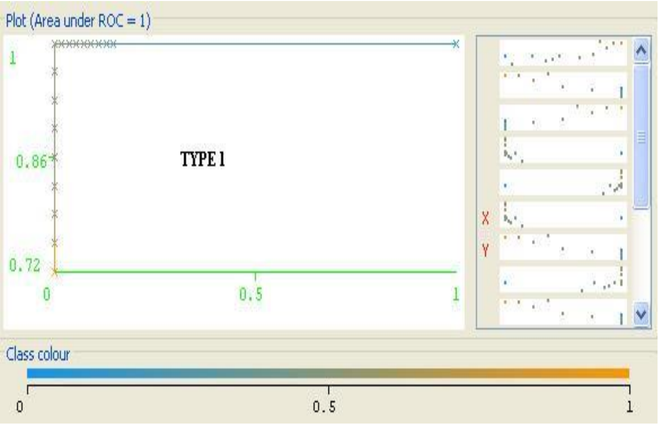


Figure. 15. ROC Plot for TYPE 1(SRBCT Dataset)

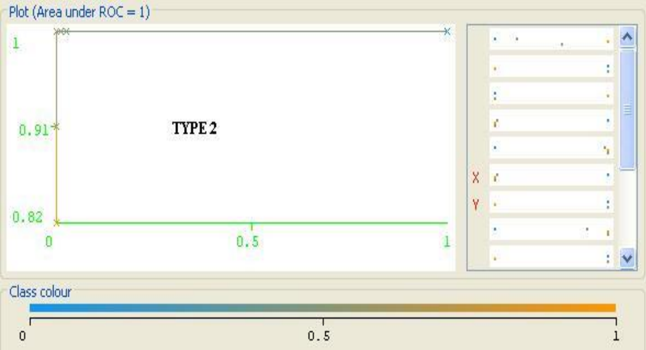


Figure. 16. ROC Plot for TYPE 2(SRBCT Dataset)

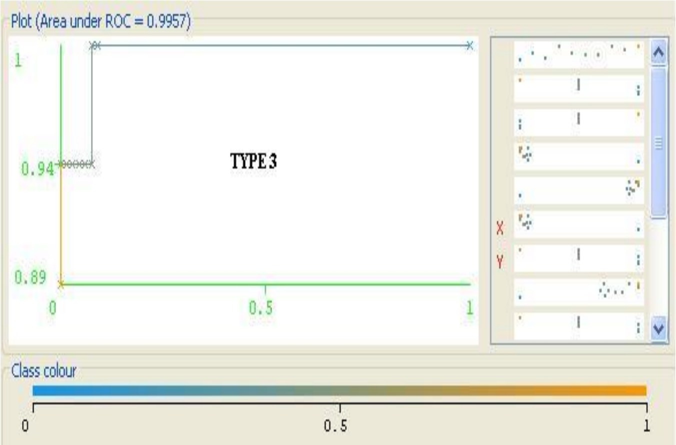


Figure. 17. ROC Plot for TYPE 3(SRBCT Dataset)

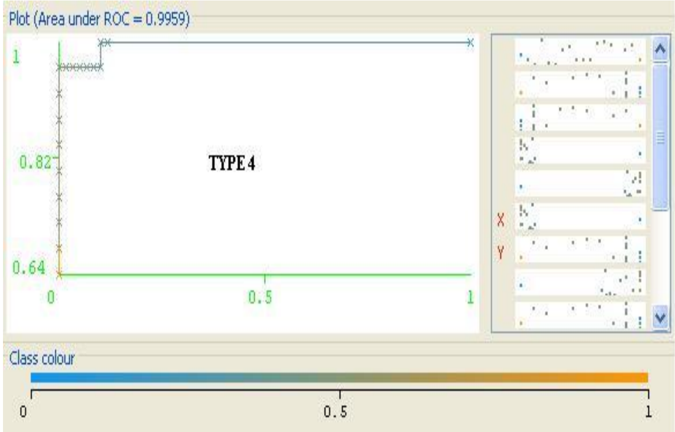


Figure. 18. ROC Plot for TYPE 4(SRBCT Dataset)

It is clearly evident from the above tables that the classifier accuracy of the hybrid methods that combines the Correlation Based Feature Selection with suitable search spaces produces higher accuracy as close to 100%. The Filter approach does not reduce the number of features beyond a certain level. Hence another approach becomes essential to reduce the number of features. The wrapper approach reduces the number of features produced by the filter approach. The combination of filter and wrapper is useful in selecting an efficient subset of features for classification purpose. The combination of feature subsets could be evaluated by using the wrapper approach that depends on the chosen classifier. The interaction among different features could be identified simultaneously using the wrapper model. The main area of research is to identify the number of features that would be required for effective cancer classification. For the entire feature selection methods, the average accuracy of the hybrid model that combines correlation based feature selection with suitable search space was better. Also the number of selected feature was also comparatively lesser for the certain models compared to the Linear Forward Fuzzy Rough Feature Selection model.

V. CONCLUSION

In this paper, we have adopted the hybrid feature selection combining correlation based filter with Best First, Combined Hill Climber, Genetic search, Greedy Stepwise method, Linear Forward Selection, Linear Forward Fuzzy Rough Feature Selection, Particle Swarm Optimization and Subset size forward selection. Later fuzzy rough neural network classifier was used to evaluate the classification performance (percentage of accuracy and other related parameters). The majority of hybrid methods have higher potential in aiding further research in the area of feature selection simplified the process of gene selection which is evident from the experimental results. The majority of the hybrid methods significantly reduces the number of genes needed for classification and has also contributed to the improvement in classifier accuracy. These hybrid methods have greater scope of application to problems in other domains in future.

VI. REFERENCES

[1] V. Bolon-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, “A review of microarray datasets and applied feature selection methods”, Information Sciences– An International Journal, Elsevier, Vol 282, pp 111-135,2014

[2] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos,Michalis V. Karamouzis, Dimitrios I. Fotiadis," Machine learning applications in cancer prognosis and prediction", Computational and Structural Biotechnology Journal, Vol 13, pp 8–17, 2015

[3] Li-Yeh Chuang, Cheng-San Yang, Kuo-Chuan Wu, Cheng-Hong Yang, “Gene selection and classification using Taguchi chaotic binary particle swarm optimization”, Expert Systems with Applications – An International Journal, Elsevier, Vol 38, No 10, pp 13367-13377,2011

[4] Li-Yeh Chuang, Chao-Hsuan Ke, Cheng-Hong Yang, “A Hybrid Both Filter and Wrapper Feature Selection Method for Microarray Classification”, Proceedings of the International Multi Conference of Engineers and Computer Scientists 2008, Vol I, pp 146-150, 2008

[5] Santanu Ghorai, Anirban Mukherjee, Sanghamitra Sengupta, Pranab K. Dutta,”Cancer Classification from Gene Expression Data by NPPC Ensemble”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 8, No 3, 2011

[6] Shahla Nemati, Mohammad Ehsan Basiri, Nasser Ghasem-Aghaee, Mehdi Hosseinzadeh Aghdam, “A novel ACO–GA hybrid algorithm for feature selection in protein function prediction”, Expert Systems with Applications – An

International Journal, Elsevier, Vol 36, No 10, pp 12086–12094,2009

[7] Rabia Aziz, C.K. Verma, Namita Srivastava,"A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data",Genomics Data, Vol 8, pp 4–15, 2016

[8] Cheng-San Yang,Li-Yeh Chuang, Chao-Hsuan Ke, Cheng-Hong Yang, "A hybrid method of feature selection for microarray gene expression data", IAENG International Journal of Computer Science, Vol 35, No 3, pp 219-225, 2008

[9] Sujata Dash,Bichitrananda Patra,B.K. Tripathy, "Study of Classification Accuracy of Microarray Data for Cancer Classification using Multivariate and Hybrid Feature Selection Method", IOSR Journal of Engineering (IOSRJEN),Vol 2, No 8, pp 112-119, 2012

[10] J.R.Anarki, M.Eftekhari, "Rough Set Based Feature Selection – A Review", Proceedings of the 5th conference on Information and Knowledge Technology, pp 301-306, 2013

[11] Gianluca Bontempi, "A Blocking Strategy to Improve Gene Selection for Classification of Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 4, No. 2, pp 293-300, 2007

[12] Mark.A.Hall, "Correlation-based Feature Selection for Machine Learning", University of Waikato, April 1999

[13] Cosmin Lazar, Jonatan Taminau, Stijn Meganck, David Steenhoff, Alain Coletta, Colin Molter, Virginie de Schaetzen, Robin Duque, Hugues Bersini, and Ann Nowe, "A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 9, No. 4, 2012

[14] C.Arunkumar,S.Ramakrishnan,"Hybrid Feature Selection using correlation coefficient and particle swarm optimization on microarray gene expression data",Innovations in Bioinspired computing and applications, Proceedings of the 6th International Conference in Bioinspired computing and Applications,Advances in Intelligent Systems and Computing, Springer,pp 229-239, 2015

[15] Huanjing Wang, Taghi M. Khoshgoftaar, Amri Napolitano, "An Empirical Investigation on Wrapper-Based Feature Selection for predicting software quality", International Journal of Software Engineering and Knowledge Engineering, World Scientific, Vol 25, No 1, 2015

[16] Li S, Yao X, Liu H, Li J, Fan B, "Prediction of T-cell epitopes based on least squares support vector machines and amino acid properties", Analytica Chimica Acta, PubMed, Vol 584, No 1, pp 37-42, 2007

[17] Paradee Namwongse,Yachai Limpiyakorn, "Learning Bayesian Network to Explore Connectivity of Risk Factors in Enterprise k Management", IJCSI International Journal of Computer Science Issues, Vol. 9, No 2, 2012

[18] M. Julia Flores ,José A. Gámez ,Ana M. Martínez,José M. Puerta, "Handling numeric attributes when comparing Bayesian network classifiers: does the discretization method matter?", Applied Intelligence, Vol 34, No 3, pp 372–385, 2011

[19] Mahbod Tavallaee, Wei Lu, Ali A. Ghorbani, "Online Classification of Network Flows", Proceedings of 2009 Seventh Annual Communications Networks and Services Research Conference, pp 78-85, 2009

[20] Pijush Barthakur, Manoj Dahal, Mrinal Kanti Ghose, "An Efficient Machine Learning Based Classification Scheme for Detecting Distributed Command & Control Traffic of P2P Botnets", International Journal of Modern Education and Computer Science, Vol 10, pp 9-18, 2013

[21] Antonio J. Rivera,Pedro Pérez-Recuerda,María Dolores Pérez-Godoy,María Jose del Jesus,María Pilar Frías,Manuel Parras, "A study on the medium-term forecasting using exogenous variable selection of the extra-virgin olive oil with soft computing methods", Applied Intelligence, Vol 34, No 3, pp 331–346, 2011

[22] Randa Oqab Mujalli, Juan de Oña, "A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks", Journal of Safety Research, Vol 42, No 5, pp 317–326, 2011

[23] Azadeh Mohammadi, Mohammad H Saraee, Mansoor Salehi, "Identification of disease-causing genes using microarray data mining and Gene Ontology", BMC Medical Genomics, Vol 4, No 12, pp 1-9, 2011

[24] Kong J, Wang S, Wahba G, "Using distance covariance for improved variable selection with application to learning genetic risk models", Statistics in Medicine,PubMed, Vol 34, No 10, pp 1708-1720, 2015

[25] Raetz EA, Perkins SL, Bhojwani D, Smock K, Philip M, Carroll WL, Min DJ, "Gene expression profiling reveals intrinsic differences between T-cell acute lymphoblastic leukemia and T-cell lymphoblastic lymphoma", Paediatric Blood and Cancer,PubMed, Vol 47, No 2, pp 130-140, 2006

[26] Cauwelier B, Dastugue N, Cools J, Poppe B, Herens C, De Paepe A, Hagemeijer A, Speleman F, "Molecular cytogenetic study of 126 unselected T-ALL cases reveals high incidence of TCRbeta locus rearrangements and putative new T-cell oncogenes", Leukemia, PubMed, Vol 20, No 7, pp 1238-1244, 2006

[27] Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia", Nature Genetics, PubMed, Vol 30, pp 41-47, 2002

[28] Qiang Xu, George Ibrahim, Rong Zheng, Norm Archer, "Toward automated
categorization of mobile health and fitness applications", Proceedings of the 4th ACM MobiHoc workshop on Pervasive wireless healthcare – MobileHealth, pp 49-54, 2014

[29] Papagelis, Athanasios, Christos Zaroliagis, "A Collaborative Decentralized
Approach to Web Search", IEEE Transactions on Systems, Man and Cybernetics - Part A, Systems and Humans, Vol. 42, No. 5, 2012

[30] C. Rubio-Escuder, Coral del Val,O. Cordón, I. Zwir, "Decision Making Association Rules for Recognition of Differential Gene Expression Profiles", Intelligent Data Engineering and Automated Learning – IDEAL 2006, Lecture Notes in Computer Science, Vol 4224, pp 1137-1149, 2006

[31] Lochandaka Ranathunga, "Performance evaluation of the combination of Compacted Dither Pattern Codes with Bhattacharyya classifier in video visual concept depiction", Multimedia Tools and Applications, Vol 54, No 2, pp 263-289, 2011

[32] Siyu Zhu, Lei Hu, Richard Zanibbi, Bertrand Coüasnon. "Rotation-robust math symbol recognition and retrieval using outer contours and image subsampling", Document Recognition and Retrieval XX, Proceedings of the SPIE, Vol 8658, 2013

[33] http://www.biolab.si/en/

Author Biographies

C.Arunkumar - Coimbatore, Tamilnadu, India.
M.Tech in Computer Science and Engineering from Vellore Institute of Technology University, Vellore, Tamilnadu in 2006, Bachelor of Engineering in Computer Science and Engineering from Bharathiar University, Tamilnadu in 2004, Area of Interest is bioinformatics.

S.Ramakrishnan - Coimbatore, Tamilnadu, India.
PhD in Information and Communication Engineering from Anna University, Chennai, Tamilnadu in 2007, M.E in communication systems from Madurai Kamaraj university, Madurai, Tamilnadu in 2000, Bachelor of Engineering in Electronics and Communication Engineering from Bharathidasan university, Trichy, Tamilnadu in 1998. He is a reviewer of 25 International Journals, in the editorial board of 7 international journals and has published 116 papers in national and international journals and conferences. His areas of research include digital image processing, soft computing, human-computer interaction, wireless sensor network and cognitive radio.