Cancer Classification with a Novel Hybrid Feature Selection Technique

Dilwar Hussain Mazumder
Department of Computer Science and Engineering
National Institute of Technology Nagaland
Dimapur - 797103, Nagaland, India
Email: dilwar@nitnagaland.ac.in

Ramachandran Veilumuthu
Department of Information Science and Technology
Anna University
Chennai - 600025, Tamil Nadu, India
Email: rama5864@gmail.com

***Abstract* — The aim of this research is to select a small optimal set of relevant genes from microarray cancer datasets, which contributes to predict the type of cancer a patient is suffering from. It is proposed to perform the gene selection using a novel hybrid feature selection technique which consists of two stages. In the filter stage an enhanced normalized mutual information based filter algorithm called Joe's Normalized Mutual Information based Feature Selection Filter (JNMIF) is used to select top ranked 150 genes. These selected genes are fed to the wrapper feature subset selector stage to select the final optimal set of genes. The proposed technique is implemented and evaluated on seven benchmark microarray cancer datasets, viz. Central Nervous System, Leukemia (binary), Leukemia (3 class), Leukemia (4 class), Lymphoma, Mixed Lineage Leukemia and Small Round Blue Cell Tumor using the well known Instance Based (IB1) classifier. With less than ten genes selected by the proposed technique, 100 percent classification accuracies are observed for four datasets except 98.61 percent for Leukemia (binary) and Leukemia (3 class) and 93.33 percent for the Central Nervous System with eleven selected genes. These results confirm that the proposed method is capable of selecting a small but optimal set of relevant genes with sufficient discriminative capacity for accurate classification of cancer. A comparative study is also made to show the betterment of the proposed method over existing popular methods of feature selection.**

Keywords - Cancer Classification; Feature Selection; Filter; Wrapper; Microarray data.

I. INTRODUCTION

A microarray gene expression dataset consists of huge number of features (in terms of thousands) called genes. Such a dataset includes records (samples or instances) of only few patients (in terms of tens only). These characteristics of microarray gene expression data, lead to an observable reduction in classification accuracy and significant increase in training time for any classification tasks on such data. So, reduction of number of features before classification becomes essential. Feature selection, also called attribute selection or gene selection in context to microarray data classification aims to select a small subset of features, out of the huge feature space, which improves the classification accuracy and at the same time reduces the classification time by removing irrelevant and redundant features [1]. Feature selection can be broadly classified into four types – filter, wrapper, embedded and hybrid approach. In filter approach [2], individual features are ranked and a subset is selected without using a learning algorithm; whereas wrapper approach [3] uses a learner to evaluate the feature subset to be selected. Filter methods are faster while wrapper methods give higher classification accuracy, for particular classifiers with higher computational cost. Embedded approach [2], on the other hand performs the feature selection process during training phase of a specific classifier. In hybrid approach [4] both filter and wrapper methods are combined to get the best of both the methods.

The main contribution of this paper is twofold.

1. To propose an enhanced feature selection filter based on Joe's normalized mutual information.

2. To propose a novel hybrid feature selection technique by combining this filter with a wrapper method of Weka [11] which uses IB1 evaluator with five fold cross validation.

The organization of the paper is as follows. A review on relevant literature and the drawbacks of existing system are presented in literature review section.The information theoretic backgrounds of mutual information based feature selection filters are illustrated in the preliminaries section. In the next section, the proposed methodologies are presented, which describes the proposed model of cancer classification with the novel hybrid feature selection scheme including the experimental setup and descriptions of the used datasets. The detailed results and discussions of the empirical study are elaborated in the results and discussion section. Finally the conclusion section discusses the conclusion and possible extensions to the works done in this paper.

II. LITERATURE REVIEW

A filter based feature selector evaluates feature goodness either individually or through feature subsets. Individual feature ranking algorithm such as Relief [5], ranks the features based on their relevance to the target class. Feature redundancy is also considered in [1] to rank individual features. Feature subsets are evaluated based on certain evaluation measure. The authors in [6] used consistency measure and in [7] they used correlation measure to evaluate goodness of feature subsets. Lei Yu and Huan Liu [8] proposed a fast correlation based filter which performs feature selection identifying relevancy and redundancy among features without pair-wise correlation analysis. This method works in two phases, first it selects relevant features based on Symmetrical Uncertainty (SU) proposed in [9] using some predefined threshold. Then from this selected list it removes the redundant features keeping the predominant features. Symmetrical Uncertainty (SU) normalizes the mutual information values in the interval [0, 1]; the value 0 indicates two variables are independent while 1 indicates they are fully dependent. The main drawback of using Symmetrical Uncertainty as the measure for ranking features is that in the case of a perfect functional dependence between two variables, Symmetrical Uncertainty does not necessarily take the value 1. This was observed by Joe in his work [10] which motivated him to define a new version of mutual information. The new version also normalizes the values in the interval [0, 1] and ensures to take the value 1 if and only if there exists a perfect functional dependence between two variables. This normalized mutual information is used in this paper to propose a new enhanced feature selection filter. The selected genes by this filter are further passed through a wrapper which uses IB1 evaluator with five fold cross validation, to select the final set of relevant genes.

III. PRELIMINARIES

This section describes the information theoretic background of mutual information based feature selection filters which are relevant for understanding the proposed work.

The entropy of a random variable is a measure of uncertainty associated with it. Higher entropy reflects larger uncertainty as to the value of that random variable and vice versa.

Definition 1: The entropy H(X) of a random discrete variable X with a probability function p(x) is defined by:

$$H(X) = -\sum_{x} p(x) \log p(x) \tag{1}$$

Definition 2: The joint entropy H(X, Y) of a pair of discrete random variables X and Y with a joint distribution p(x, y) is defined as follows:

$$H(X,Y) = -\sum_{x}\sum_{y} p(x,y) \log p(x,y) \tag{2}$$

Definition 3: The conditional entropy H(Y|X) is defined as:

$$H(Y \mid X) = \sum_{x} p(x) H(Y \mid X = x) \tag{3}$$

The conditional entropy of Y conditioned on X refers to the average entropy of Y conditioned on the value of X averaged over all possible value of X.

Theorem 1: The chain rule of joint entropy.

$$H(X,Y) = H(X) + H(Y \mid X) \tag{4}$$

The chain rule for joint entropy states that the total uncertainty about the value of X and Y is equal to the uncertainty of X plus the average uncertainty about Y once X is known.

Definition 4: The mutual information I(X, Y) between two random variables X and Y, measures how much on average the realization of Y tells about the realization of X, is defined as follows:

$$I(X,Y) = \sum_{x}\sum_{y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{5}$$

Where p(x, y) is the joint probability mass function of X and Y; and p(x) and p(y) are marginal probability function of X and Y respectively.

Equation 5 can also be written in terms of entropies as follows, which states how much the entropy of X is reduced if we know the realization of Y:

$$I(X,Y) = H(X) - H(X \mid Y) \tag{6}$$

Mutual Information is symmetric; X tells us exactly as much about Y as Y tells about X.

Theorem 2: Symmetry of mutual information.

$$\begin{aligned} I(X,Y) &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) = I(Y,X) \end{aligned} \tag{7}$$

We can express mutual information applying the chain rule of Theorem 1 as follows:

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y \mid X) \\ &= H(Y) - (H(X,Y) - H(X)) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{8}$$

A normalized version of mutual information known as asymmetric uncertainty coefficient I(X, Y), indicates the relative decrease in uncertainty of X given Y, is expressed as follows:

$$U(X,Y) = \frac{I(X,Y)}{H(X)} \tag{9}$$

Symmetric Uncertainty (SU) coefficient [9] is a symmetric version of U(X, Y), is defined as follows:

$$SU(X,Y) = \frac{I(X,Y)}{\frac{1}{2}[H(X)+H(Y)]} \tag{10}$$

SU(X, Y) normalizes the mutual information values to the interval [0, 1]; value 0 indicates that X and Y are independent while value 1 indicates X and Y are fully dependent. But when there exists a perfect functional dependence between X and Y, it does not necessarily take the value 1. This observation made Joe to define another version of mutual information.

IV. PROPOSED METHODOLOGIES

This section describes the proposed methodologies for cancer classification. The proposed model is shown in Figure 1, consists of two phases. Once the dataset is loaded, in Phase 1, gene selection is performed using the proposed hybrid feature selection scheme which consists of two stages. First stage is the feature selection filter. In this stage 150 top ranked genes are selected using the enhanced feature selection filter based on Joe's normalized mutual information. Second stage is the wrapper feature selection where the top 150 genes from the filter stage are used to select the final optimal set of genes which are utilized for classification.

In phase 2, the selected genes of phase 1 are used for classification to get the results. The two phases of the model are described below.

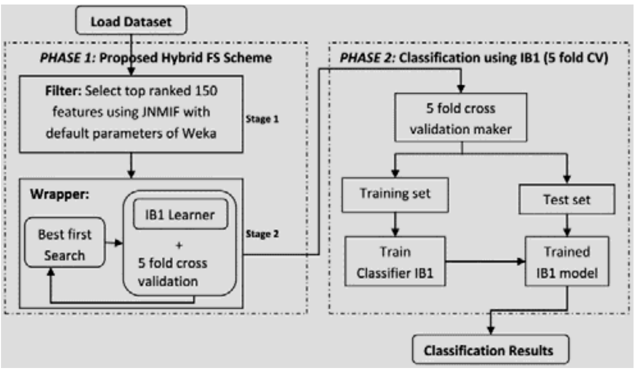

Load Dataset
PHASE 1: Proposed Hybrid FS Scheme
Filter: Select top ranked 150 features using JNMIF with default parameters of Weka
Stage 1
Wrapper:
Best first Search
IB1 Learner
+
5 fold cross validation
Stage 2
PHASE 2: Classification using IB1 (5 fold CV)
5 fold cross validation maker
Training set
Test set
Train Classifier IB1
Trained IB1 model
Classification Results

Figure 1. Architecture of the cancer classifiction system using the proposed hybrid feature selection scheme.

A. Phase 1: The proposed hybrid feature selection scheme

The two stages of the hybrid feature selection scheme are described below.

Stage 1: Filter Feature Selection

In this stage genes are ranked in descending order based on using the proposed JNMIF filter which uses normalized mutual information for ranking the genes.

The top ranked 150 genes are selected in this stage. The ranking genes are done using the version of normalized mutual information defined by Joe in his work [10] is as follows:

$$JNMI(X,Y) = \frac{I(X,Y)}{\min[H(X),H(Y)]} \tag{11}$$

JNMI(X, Y) also normalizes the values in the interval [0, 1]. Moreover JNMI(X, Y) is equal to 1 if and only if X and Y are functionally dependent [10]. This fact is exploited in the proposed filter method of feature selection in stage 1.

Based on the formulation of equation 11, an algorithm named Joe's Normalized Mutual Information based filter (JNMIF) is developed.

Algorithm JNMIF: Joe's Normalized Mutual Information based filter (JNMIF)

```
Input: Dataset D with n features F = {f1, f2, f3, …, fn} and class levels C
Output: FS the set of selected features
Steps:
Choose threshold k
FS = $\phi$
for i =1 to n do
        Calculate JNMI( $f_i$, C) using equation 11
end for
count = 1
while count <= k do
        Select the feature $f_i$ with maximum JNMI( $f_i$, C) value
        FS = FS U {$f_i$ }
        F = F - {$f_i$ }
        count = count + 1
end while
return FS
```

The algorithm finds JNMI(f_i, C) value between each feature f_i and class label C. Then, in each iteration, the algorithm picks up the feature with highest JNMI(f_i, C) value, includes it in the selected set and removes it from the original set, till top k features are selected.

Stage 2: Wrapper Feature Subset Selection

In this stage the wrapper subset evaluator [12] is used to select the best subset of genes. Instance based classifier (IB1) is used as the induction algorithm. The IB1 classifier is discussed in next section. The best first search is employed on the 150 top genes from the filter in stage 1 as the strategy to search a good subset of genes. This subset of genes is 5-fold cross validated. IB1 classifier is repeatedly run with random permutation on 4 partitions as training set and the 5th partition as the test set. This process is repeated on all the subsets of genes generated by the search. Finally the gene subset with highest classification accuracy is selected as final set of optimal genes for classification in phase 2.

B. Phase 2: Classification

The Instance based classifier (IB1) [13] is one of the simplest and fast classifiers used for a wide range of machine learning activities. This classifier is used in our proposed model both as the induction algorithm for the wrapper in stage 2 of phase 1 and also for the final classification of cancer in phase 2. The best subset of genes from the proposed hybrid feature selection scheme is used as the input to the IB1 classifier with 5-fold cross validation in phase 2 to get the classification accuracies as the result of classification.

The IB1 Classifier

The IB1 algorithm [13] classifies instances based on a similarity score calculated based on equation 12.

$$Similarity(x,y) = -\sqrt{\sum_{i=1}^{n} f(\mathrm{x}_i, \mathrm{y}_i)} \qquad (12)$$

Where instances have n features and $f(x_i, y_i) = (x_i - y_i)^2$.

For a test instance *y* the similarity score with every training instance *x* is calculated. Then the maximum of the calculated similarity score of *y* is identified. If the class label of any one of the instances in the training set is same as that of the test instance with maximum similarity score, then the classification is correct; otherwise the classification is incorrect.

V. RESULTS AND DISCUSSIONS

A. Experimental setup and Microarray datasets used

The proposed hybrid feature selection scheme is implemented in WEKA framework (Witten and Frank, 2000). All experiments are performed in WEKA framework on a standalone PC with Intel i3 CPU (2 core with 2 threads each) and 3 GB of RAM. As suggested in [14], the top ranked 150 genes are selected by the filter in stage 1. These 150 genes are feed to the wrapper to obtain the final small set of optimal genes which are used for classification by the IB1 classifier in phase 2. A 5-fold cross validation mechanism is used to record the classification accuracy for all the seven datasets used in the experiments.

Seven benchmark microarray gene expression datasets studied in [15] are used in the experiments. A summary of these datasets is given in TABLE I.

TABLE I. DATASETS SUMMERY

Dataset name	No. of Genes	No. of Instances	No. of Classes
CNS	7129	60	2
Leukemia (binary)	7129	72	2
Leukemia (3 class)	7129	72	3
Leukemia (4 class)	7129	72	4
Lymphoma	4026	66	3
MLL	12582	72	3
SRBCT	2308	83	4

The Central Nervous System (CNS) dataset includes 60 samples. It has two classes, survivors' class represents the patients who are alive after treatment and the rest are represented by the failures class. The Leukemia (binary) dataset has gene expression profiles for two classes of leukemia, Acute Lymphoblastic Leukemia (ALL) and Acute Myeloblastic Leukemia (AML). The ALL part consists of two types, B-cell and T-cell whereas Bone Marrow samples (BM) and Peripheral Blood samples (PB) are two types in AML. Accordingly this dataset has three-class (B-cell, T-cell and AML) and four class (B-cell, T-cell, AML-BM and AML-PB) versions referred here as Leukemia (3 class) and Leukemia (4 class) respectively. All the three versions of Leukemia datasets have 72 samples. The Lymphoma dataset includes three classes of lymphoid malignancies with 66 samples. The Mixed Lineage Leukemia (MLL) dataset also consists of three classes and has 72 samples. The Small Round Blue Cell Tumor (SRBCT) dataset is of small, round blue cell tumors. It has four classes and includes 83 samples.

B. Experimental Results and Discussions

The proposed hybrid scheme of feature selection is able to select less than 10 genes in six of the seven datasets used in the experiments. Detailed results of every datasets used are described below. A list of final selected genes is presented in TABLE II.

With eleven genes selected a classification accuracy of 93.33 percent is observed for CNS dataset. Out of 60 samples, 56 are correctly classified and remaining 4 are misclassified. Leukemia (binary) and Leukemia (3 class) has received a classification accuracy of 98.61 percent with only 3 selected genes each. In both these datasets 71 out of 72 samples are correctly classified. In Leukemia (4 class), Lymphoma, MLL and SRBCT 100 percent classification accuracies are observed with only 7,3, 6 and 6 numbers of selected genes respectively.

It may be noted that the number of finally selected genes is less than 12 in all datasets. Moreover the classification accuracies achieved with such a small set of genes are excellent. These observations establish two facts.

1. The proposed hybrid scheme is capable enough to select a very small set of genes from the huge number of genes in the original dataset.

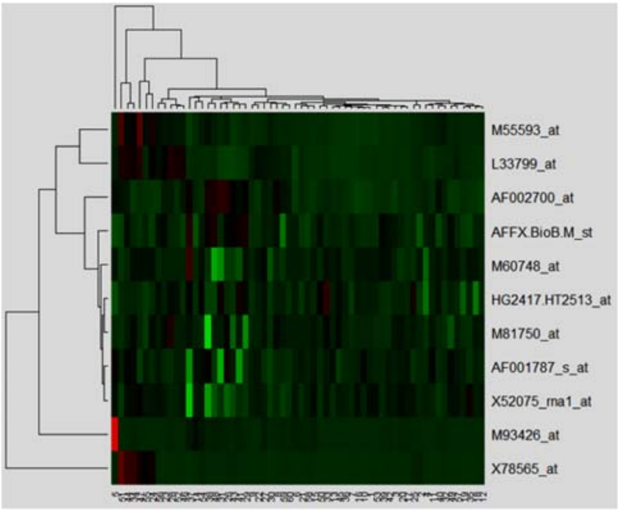
2. The selected small set of genes have sufficient discriminative capacity for accurate classification of cancer.

TABLE II. SUMMARY OF CLASSIFICATION RESULTS: THE NUMBER OF GENES SELECTED, SYMBOLS OF THE SELECTED GENES AND CLASSIFICATION ACCURACIES IN PERCENTAGE WITH THE SELECTED GENES USING IB1 CLASSIFIER FOR THE SEVEN DATASETS USED IN EXPERIMENTS

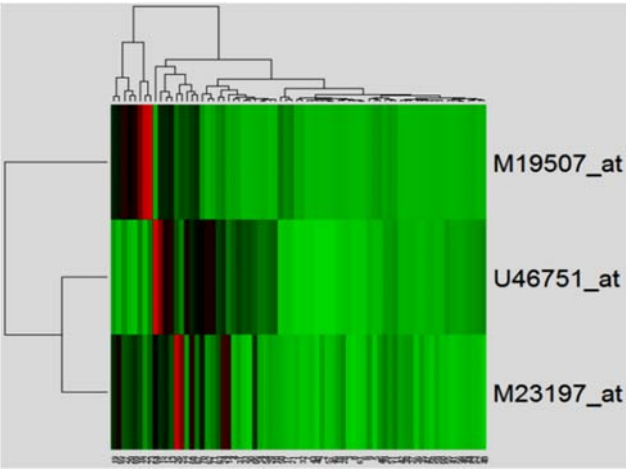
Dataset	No. of selected genes	Gene symbols of selected top genes	Accuracy in percentage by IB1 classifier
CNS	11	AF002700_at, AF001787_s_at, M55593_at, M60748_at, AFFX-BioB-M_st, HG2417-HT2513_at, L33799_at, X52075_rna1_at, X78565_at, M93426_at, M81750_at	93.33
Leukemia (binary)	3	M23197_at, M19507_at, U46751_at	98.61
Leukemia (3 Class)	3	D00749_s_at, M89957_at, L09209_s_at	98.61
Leukemia (4 Class)	7	D00749_s_at, M23197_at, M89957_at, L08895_at, M92287_at, HG2981-HT3127_s_at, M38690_at	100
Lymphoma	3	GENE524X, GENE1622X, GENE647X	100
MLL	6	1389_at, 36239_at, 1894_f_at, 35974_at, 33162_at, 37159_at	100
SRBCT	6	gene2198, gene1003, gene1389, gene174, gene236, gene910	100

C. Biological Significance

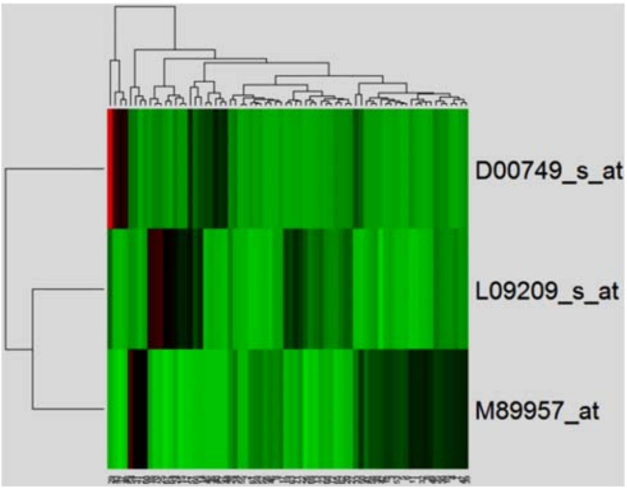
The expression levels of the selected genes across the number of samples in the datasets are represented in the form of heat maps. Heat map graphically represents expression levels of genes across the samples in the microarray as matrix of colors. Darker color represents larger values while lighter color represents smaller values. Figure 2, below and on the next page, shows the heat maps for the seven datasets with reduced gene subset of the samples.



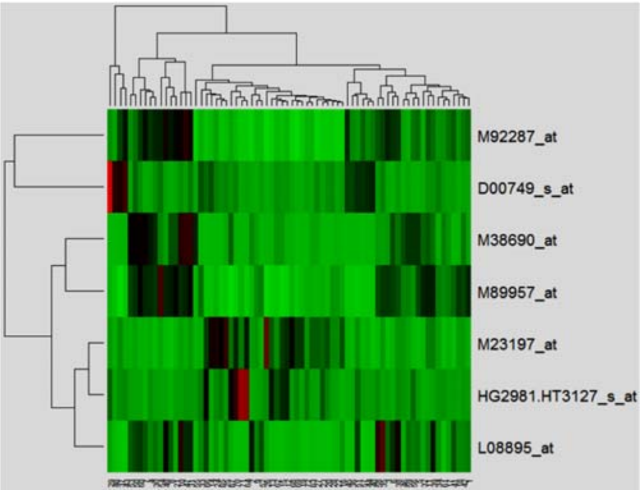
(A)



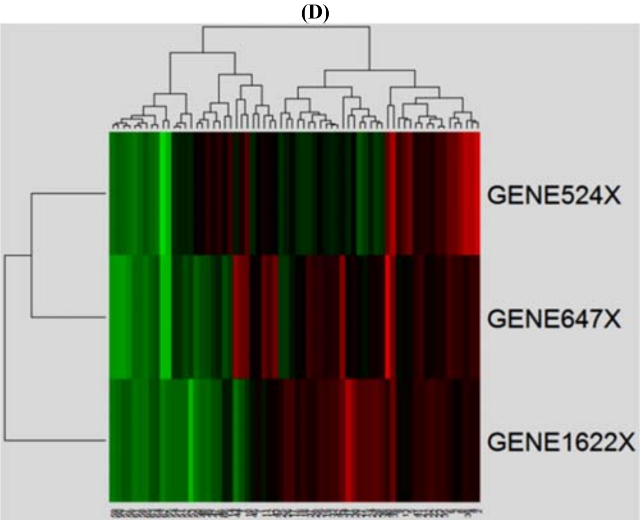
(B)



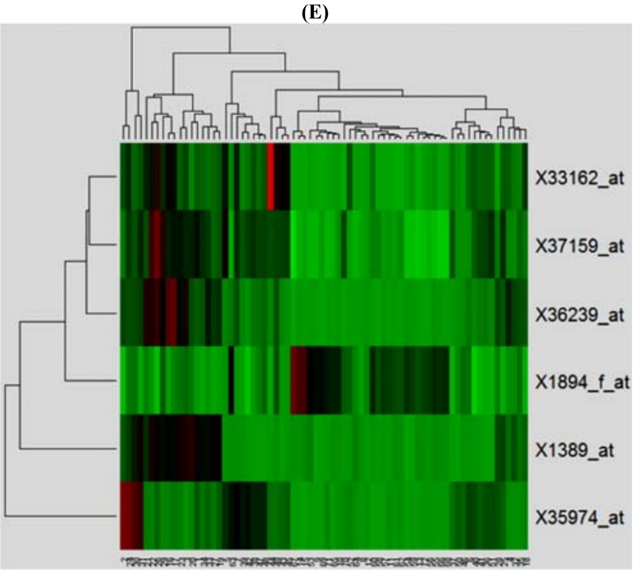
(C)



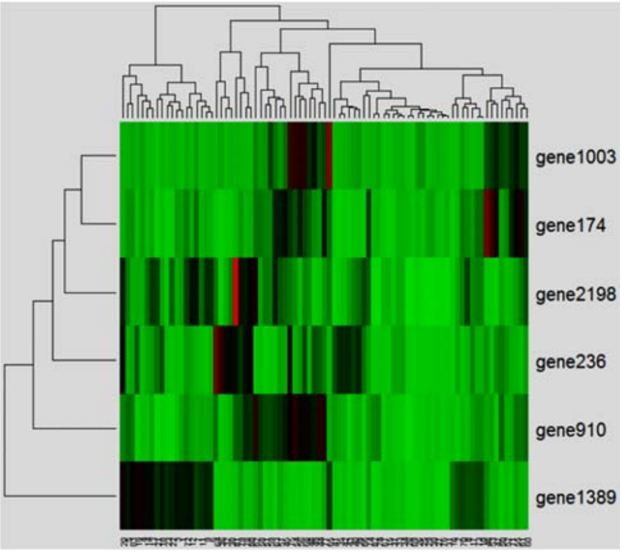
(D)



(E)



(F)



(G)

Figure 2. Heat maps of the seven datasets with reduced features (A) heat map on CNS data with 11 genes (B) heat map on Leukemia (binary) data with 3 genes (C) heat map on Leukemia (3 class) data with 3 genes (D) heat map on Leukemia (4 class) data with 7 genes (E) heat map on Lymphoma data with 3 genes (F) heat map on MLL data with 6 genes (G) heat map on SRBCT data with 6 genes.

D. Comparisons

The classification accuracies with IB1 classifier achieved by the proposed method on the seven datasets are compared with that of three other popular implementations of feature selection techniques available in Weka viz. Information Gain (IG), Gain Ratio (GR) and Symmetric Uncertainty (SU). TABLE III lists these comparisons. The comparison of accuracies is also shown graphically in Figure 3. From the comparisons it is evident that the proposed feature selection performs better than existing popular feature selection techniques in all the seven datasets used in the experiments.

TABLE III. COMPARISON OF CLASSIFICATION ACCURACIES (%) OF THE PROPOSED METHOD WITH THE OTHER THREE POPULAR METHODS VIZ. INFORMATION GAIN (IG), GAIN RATIO (GR) AND SYMMETRIC UNCERTAINTY (SU).

Dataset	Accuracy of Feature Selection			
	IG	GR	SU	Proposed
CNS	73.33	73.33	73.33	93.33
Leukemia (binary)	97.22	97.22	97.22	98.61
Leukemia (3 class)	94.44	97.22	94.44	98.61
Leukemia (4 class)	97.22	91.67	95.83	100
Lymphoma	92.42	98.48	98.48	100
MLL	95.83	94.44	97.22	100
SRBCT	100	100	100	100

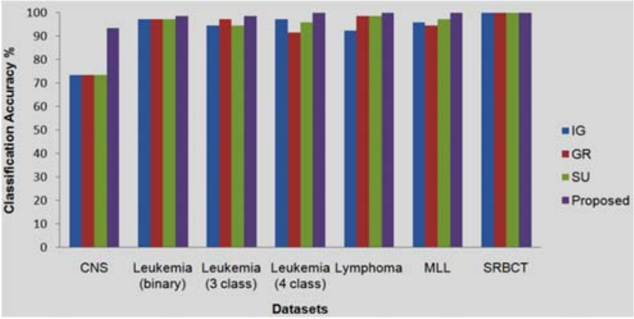


Figure 3. Comparison of classification accuracies (%) of the proposed method with the other three popular methods viz. Information Gain (IG), Gain Ratio (GR) and Symmetric Uncertainty (SU).

VI. CONCLUSION

In this paper a cancer classification system with a novel hybrid feature selection scheme is presented.

The main contribution of this paper is twofold.

1. An enhanced feature selection filter based on Joe’s normalized mutual information is proposed.

2. A novel hybrid feature selection scheme is proposed by combining this filter with a wrapper method of Weka (Witten and Frank, 2000) which uses IB1 evaluator with five-fold cross validation.

The proposed hybrid scheme of feature selection is capable enough to select a small but optimal set of relevant genes. Also the selected small set of genes possesses sufficient discriminative capacity for accurate classification of cancer. Moreover the proposed feature selection method performs better than existing popular feature selection techniques.

REFERENCES

[1] Dash, M., & Liu, H. (1997). Feature selection for classifications. Intelligent Data Analysis: An International Journal, 1, 131–156.

[2] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3, 1157–1182.

[3] Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97(1), 245–271.

[4] Hsu, H.-H., Hsieh, C.-W., & Lu, M.-D. (2011). Hybrid feature selection by combining filters and wrappers. Expert Systems with Applications, 38(7), 8144–8150.

[5] Kira, K., & Rendell, L. (1992). The feature selection problem: Traditional methods and a new algorithm. Proceedings of the Tenth National Conference on Artificial Intelligence (pp. 129–134). Menlo Park: AAAI Press/The MIT Press.

[6] Dash, M., Liu, H., & Motoda, H. (2000). Consistency based feature selection. Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining (pp. 98–109). Springer-Verlag.

[7] Hall, M. (1999). Correlation based feature selection for machine learning. Doctoral dissertation, University of Waikato, Dept. of Computer Science.

[8] Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings of the Twentieth International Conference on Machine Learning. ICML 2003.

[9] Sarndal, C., E. (1974). A comparative study of association measures. Psychometrika, 39, 165–187.

[10] Joe, H. (1989). Relative entropy measures of multivariate dependence. Journal of the American Statistical Association, 84(405), 157–164.

[11] Witten, I., H., & Frank, E. (2000). Data Mining: Practical Machine Learning Tools with Java Implementations. Morgan Kaufmann, San Francisco, CA.

[12] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. 1997; 97(1-2):273-324.

[13] Aha DW, Kibler D, Albert MK. Instance-based learning algorithms. Mach Learn. 1991; 6(1):37-66.

[14] Li, T., Zhang, C., & Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20(15), 2429–2437.

[15] Zhu, Z., Ong, Y., S., & Dash, M. (2007). Markov Blanket-Embedded Genetic Algorithm for Gene Selection. Pattern Recognition, 49(11), 3236-3248.