Journal of Theoretical and Applied Information Technology
31st January 2013. Vol. 47 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HYBRIDIZING RELIEF, MRMR FILTERS AND GA WRAPPER APPROACHES FOR GENE SELECTION

SALAM SALAMEH SHREEM, SALWANI ABDULLAH, MOHD ZAKREE AHMAD NAZRI, MALEK ALZAQEBAH
Data Mining and Optimisation Research Group,
Center for Artificial Intelligence Technology,
Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

E-mail: salam@ftsm.ukm.my, salwani@ftsm.ukm.my, mzan@ftsm.ukm.my, malek_zaqeba@ftsm.ukm.my

ABSTRACT

Gene expression data comprises a huge number of genes but have only few samples that can be used to address supervised classification problems. This paper is aimed at identifying a small set of genes, to efficiently distinguish various types of biological sample; hence we have proposed a three-stage of gene selection algorithm for genomic data. The proposed approach combines ReliefF, mRMR (Minimum Redundancy Maximum Relevance) and GA (Genetic Algorithm) coded as (R-m-GA). In the first stage, the candidate gene set is identified by applying the ReliefF. While, the second minimizes the redundancy with the help of mRMR method, which facilitates the selection of effectual gene subset from the candidate set. In the third stage, GA with classifier (used as a fitness function by the GA) is applied to choose the most discriminating genes. The proposed method is validated on the tumor datasets such as: CNS, DLBCL and Prostate cancer, using IB1 classifier. The comparative analysis of the R-m-GA against GA and ReliefF-GA has revealed that the proposed method is capable of finding the smallest gene subset that offers the highest classification accuracy.

Keywords: *Genetic Algorithm, Gene selection, Microarray datasets, mRMR, Relief*

1. INTRODUCTION

Gene selection for microarray classification is used to build an efficient model to discover the most important genes from a sample of gene expressions, i.e., to categorize tissue samples into various groups of diseases to help scientist to identify the underlying mechanism that relates gene expression of certain diseases. Many researchers have studied classification methods using the microarray data for various purposes, for example to distinguish cancerous and normal tissues [1-3]. However, the main challenge is that the microarray datasets have high dimensionality (more than 10000 gene expressions) but have small number of samples (hundred or less samples). Moreover, the microarrays datasets comprise a lot of genes that are unrelated or redundant to some specified disease. Therefore, before a classification method can be used on the microarray, researchers need to address challenges associated with high dimensional features known as "the curse of dimensionality". Hence, it is customary to use feature selection technique to solve the high dimensionality problem first before classifications by eliminating the redundant and irrelevant features through eliminating genes with little or unproductive information. It is critical to select highly discriminating genes for enhancing the accuracy of classification and prediction of diseases [4].

In feature selection problems, identifying a set of genes that best distinguishes the various types of biological samples is the biggest challenge. Feature selection entails in identifying a subset of features, in order to enhance the accuracy or minimise the size of the subset of genes, without drastically reducing the prediction accuracy of the classifier, which is built by using only the selected features [5].

Based on the combination of feature selection search with the construction of the classification model, Saeys et. al. [6] have categorized two feature selection techniques, such as: filter methods and wrapper methods. The former is initially employed to choose the subsets of features, without depending on the induction algorithm. They are computationally cheap. Conversely, the wrapper approach applies the induction algorithm as a fitness function to evaluate the features subset. They generally select more suitable subset of feature for the induction algorithm but it more computationally expensive than filter approach.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

It becomes unrealistic to search the best possible feature subset by investigating all potential subsets of genes due to the huge number of the subsets, where the search space is exponentially increased. Variety of heuristic approaches have been employed for feature subset selection, such as: Ant colony optimization [7], tabu search [8], memetic algorithm [2], simulated annealing [9], GRASP [3] and GA [10].

This paper has proposed a three-stage selection algorithm by hybridizing the ReliefF, mRMR filter (as filters method) and GA (as a wrapper method) for addressing gene selection problem (see section 3).

We have compared the R-m-GA with GA and ReliefF-GA in isolation. Further comparison has been carried out with other available methods on three datasets. The experimental results have revealed that the R-m-GA effectively selects the most relevant genes.

The remainder of this paper is organized as follows: Section II introduces the ReliefF, mRMR and GA. Section III presents our proposed method. Sections IV and V illustrate the experimental results and conclusion, respectively.

2. RELIEFF, MRMR FILTERS AND GENETIC ALGORITHM WRAPPER APPROACHES FOR GENE SELECTION

In gene expression microarray data, the capability of selecting few numbers of predictive and important genes, not only makes the data analysis efficient but also helps their biological interpretation and understanding of the data. In this section, we have described three popular methods for gene selection and classification for microarray data. Initially a short overview of the relief filter is provided, followed by the short introduction of the mRMR filter. Finally the genetic algorithm wrapper search strategy has been discussed.

a. First Stage: Relief Filter

ReliefF is an uncomplicated and efficient method used as a pre-processing feature subset selection method, to assess the quality of the features that have very high dependencies between the features [11]. ReliefF is capable of dealing with multiclass datasets and is an efficient method to deal with noisy and incomplete datasets. It can be used to estimate the quality and identify the existence of conditional dependencies between attributes effectively.

The main concept of theReliefF , is to assess the quality of genes based on their values to differentiate among instances that are close with each other. Given a randomly chosen instance Ins_m from class L, the ReliefF searches for K of its nearest neighbours, from the same class known as nearest hits H, and also K nearest neighbours from each of the different classes, called nearest misses M. Later it updates the quality estimation W_i for gene i depending on their values for Ins_m, H, M. If the instance Ins_m and the others in H have dissimilar values on gene i, then the quality estimation W_i is reduced. In contrast, if instance Ins_m and those in M have dissimilar values on the gene i, then W_i is increased. The entire process is iterated n times, which is set by users. To update W_i the Equation1is used as follows:

$$W_i = W_i - \frac{\sum_{k=1}^{k} D_H}{n.K} + \sum_{c=1}^{C-1} P_c . \frac{\sum_{k=1}^{K} D_{M_c}}{n.k} \quad (1)$$

where n_c is the number of instances in class c, D_H (or D_{mc} is the sum of distance between the selected instance and each) H (or M_c), M_c is the probability of class c.

Comprehensive information on ReliefF can be found in [15] and of late, it was proved that ReliefF is an on-line solution to a conplex optimization problem, increasing a margin-based algorithm [12].

b. Second Stage: mRMR filter

The mRMR filter method selects genes with the highest relevance and minimally redundant with the target class [13]. In mRMR, the Maximum Relevance and Minimum Redundancy of genes are based on mutual information. Given g_i, which represents gene i and c represents the class label, the mutual information of g_i and c is defined in terms of their probability frequencies of appearances $P(g_i)$, $P(c)$, and $P(g_i, c)$ as follows:

$$I(g_i, c) = \iint p(g_i, c) \ln \frac{p(g_i, c)}{p(g_i) p(c)} d_{g_i} d_c \quad (2)$$

The Maximum Relevance method selects the highest top m genes, which have the highest relevance correlated to the class labels from the descent arranged set of $I(g_i, c)$.

$$\max \frac{1}{|S|} \sum_{g_{i \in S}} I(g_i; c) \quad (3)$$

However, it is well known to the researchers that "the m best features are not the best m features"

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

,because the correlations among those top genes might also be high [14]. Therefore, Minimum-Redundancy criterion is introduced by [14] in order to remove the redundancy features. The following is the Minimum Redundancy criterion:

$$\min_{S} \frac{1}{|S|^2} \sum_{g_i, g_j \in S} I(g_i; g_i) \quad (4)$$

Equation (4) shows that the mutual information between each pair of genes is taken into consideration. The (MRMR) filter combines both optimization criteria of Eqs. (3 and 4).

A sequential incremental algorithm to solve the simultaneous optimizations of optimization criteria of Eq. (3 and 4) is explained as follows: Assume, we have G which represent a set of genes and also have $S_m - 1$, the gene set with $m-1$ genes, and then the task is to choose the $m-$th gene from the set $\{G - S_m - 1\}$. This feature is chosen by increasing the single-variable relevance minus redundancy function.

$$\max_{g_i \in G - S_{m-1}} [I(g_i; c) - \frac{1}{1-m} \sum_{g_i \in S_{m-1}} I(g_j, g_i)] \quad (5)$$

The $m-th$ features also can be chosen by maximizing the single variable relevance divided by redundancy function

$$\max_{g_i \in G - S_{m-1}} [I(g_i; c) / \frac{1}{1-m} \sum_{g_i \in S_{m-1}} I(g_j, g_i)] \quad (6)$$

c. Third Stage GA Wrapper

A genetic algorithm (GA) is a stochastic search algorithm developed by Holland in 1970. GA is a very successful algorithm, especially when applied to address machine learning, search and optimization problems [15]. GA iteratively generates new population of solutions from the older solutions, by intelligent exploitation of a random search. In the first step, some chromosomes are randomly created from the whole solution space. The number of chromosomes is known as the population size. Then, two parent chromosomes C1 and C2 are selected to generate new chromosome (i.e., offspring C') by using crossover operator followed by the mutation operation to slightly perturb offspring C'. Once C' is obtained, the evaluation function computes and compares it with the fitness of the parent chromosomes. The chromosome with the highest fitness value is kept, whereas the ones with less fitness are discarded. This process will be repeated until a predefined maximum number of generations are met. More details of GA can be found in Holland [15].

In the feature selection problem, each chromosome is represented by binary values (1 or 0), which identify the selected or excluded feature in the corresponding feature subset, respectively. This process is called as chromosome or solution encoding. For instance, in the chromosome of ten genes, e.g. 1100111000, the genes at position 1,2,5,6 and 7 are selected, while the genes at position 3,4,8,9 and 10 are excluded in the corresponding feature subset. Then the genetic operator (selection, crossover and mutation) is applied on the initial random population to generate new chromosomes.

3. THE PROPOSED METHOD

As mentioned earlier, the proposed three-stage method hybridizes mRMR, ReliefF and GA algorithms. These three algorithms are executed one after the other.

In the first stage, the ReliefF is employed on the dataset to acquire a candidate gene set. The genes are assessed and arranged according to the ReliefF criterion. Then, the ReliefF searches and selects the top P genes (150 genes in this case), which will become the candidate gene set. This process filters the insignificant genes and therefore minimizes the computational load for the next stage.

In the second stage, the mRMR method is applied on the top genes. Take note that the P genes (i.e., 150 genes) are gained from the first stage (ReliefF filter). In this stage, the mRMR reduces the number of redundancy and insignificant genes in order to choose a compact and effectual gene (select the top 50 genes). The main objective here is to reduce the computational load for GA wrapper. From the reduced set of genes obtained in the previous stage, the third stage uses a wrapper approach that combines GA and IB1 classifier, to accomplish the gene subset selection process. The main reason of using GA in this proposed method is to discover good subsets of genes, since GA is well known for its exploitation capabilities and efficiency in exploring the search space. The scheme for our proposed model is illustrated in Fig. 1.

Journal of Theoretical and Applied Information Technology
31st January 2013. Vol. 47 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

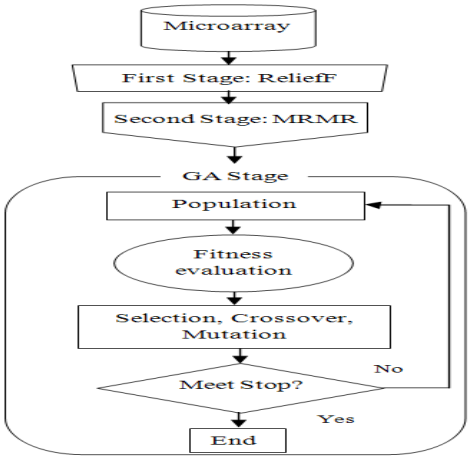

Fig. 1 The General Scheme Of The R-M-GA Algorithm

4. EXPERIMENTAL RESULTS AND DISCUSSION

All the experiments in this study were performed on system with Intel Core i5, 2.4 Ghz processor with 4 GB RAM. Our proposed algorithms implemented using Java. 10 folds cross validation (10-CV) has been performed using IB1 classifier to assess the fitness function and the classification accuracy. We have performed two set of experiments. In the first experiment we have compared the performance of the R-m-GA against GA and ReliefF-GA. In the second experiment we have compared the performance of R-m-GA against the state-of-the-art approaches. In both these experimental, the results are reported over 10 independent runs.

a. Datasets

To assess the worth of the R-m-GA approach, we have conducted experiments on three microarray datasets of gene expression profiles. The datasets are tabulated in Table 1, which can be downloaded from "http://datam.i2r.a-star.edu.sg/datasets/krbd/"

Table 1 Description Of The Datasets

Data Set	Genes	Samples
CNS	7129	60
DLBCL	4026	47
PROSTATE	12600	109

b. Parameter settings

Table 2 illustrates the GA parameters that have been used in our proposed model. In our proposed algorithm, the first stage will stop, after choosing the top 150 genes. However, the mRMR stops after choosing the top 50 genes.

Table 2: Parameters Settings

Parameter	Value
Population size	50
Number of generation	30
Crossover probability	0.6
Mutation rate	0.5

c. Comparison of R-m-GA against GA and ReliefF-GA

In order to evaluate the merits of incorporating Relief, mRMR with GA, we have first compared *R-m-GA* against results obtained by GA and Relief-GA. This comparison allows us to analyse the impact of combining the ReliefF with mRMR filter and GA in a single process. Table 3 shows the comparison of results between GA, ReliefF-GA and R-m-GA on three datasets.

TABLE 3 Results Comparison

Datasets		GA	ReliefF-GA	R-m-GA
CNS	|#G|	32	31	**21.8**
	ACC	76.6	78	**90.2**
DLBCL	|#G|	31.2	29	**2.8**
	ACC	98.9	91.6	**100**
Prostate	|#G|	25	30	**2.2**
	ACC	85.2	86.4	**100**

Note: ACC: Average accuracy rate (%), |#G|: Average number of genes.

Each cell in Table 3 shows the average of classification accuracy and number of genes over ten independent runs. The best results in each row are shown in bold.

Based on the table 3, it is noteworthy that the R-m-GA has outperformed GA and Relief-GA, on all datasets in terms of accuracy and the number of selected genes. Hence, our method is capable of effectively reducing the size of the dimensionality and accurately eliminating noise and irrelevant features.

d. Comparison with other studies

Table 4 tabulates the performance of other methods obtained from the literature and the proposed *R-m-GA* algorithm.

Table 4. Comparison Of R-M-GA Results Against Other Methods On Cancer Classification

Datasets		R-m-GA	[16]	[17]	[18]	[19]	[1]
CNS	|#G|	21.8	20	30	65	---	**3**
	ACC	90.2	68.5	80	46	---	**99.3**
DLBCL	|#G|	**2.8**	20	30	---	30	3
	ACC	**100**	93	92.2	---	98	99.5
Prostate	|#G|	**2.2**	20	30	---	30	3
	ACC	**100**	91.7	95.5	---	97	99.5

Note: ACC: Average accuracy rate (%), |#G|: Average number of genes, --- : not available.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

From Table 4, one can easily observe that, in terms of number of generated genes and accuracy, the R-m-GA perform better on 2 datasets (DLBCL and Prostate) compared to others methods. It is believed that the better performance is the result of the combination of the two filters (Relief and mRMR) with the wrapper approach (GA) in a single process. These positive results reveal that the R-m-GA is effective when dealing with huge numbers of genes.

5. CONCLUSION

In this paper, we have proposed a scheme for gene selection by combining the ReliefF, mRMR filter and GA wrapper approaches in a single process. The new scheme constitutes three-stage processes, each with different role. In the first stage, the RelieF filter was used to identify a candidate gene set. This process filters insignificant genes and therefore had minimized the computational load for mRMR. The mRMR had been applied in the second stage of the proposed algorithm. The mRMR is efficient in directly minimizing redundancy and selecting effective gene subset, from the candidate gene set collected from the first stage. In the final stage, the GA wrapper approach was applied. The approach is comprised of the GA search strategy and a learning algorithm as a fitness function that will assess and achieve the gene subset selection. The experiments were carried out with three diverse datasets for cancer classification. The results had illustrated that the proposed method (m-R-GA) is very effective and has great potential for gene selection.

REFERENCES

[1] E. Bonilla-Huerta, *et al.*, "Hybrid Filter-Wrapper with a Specialized Random Multi-Parent Crossover Operator for Gene Selection and Classification Problems," *Bio-Inspired Computing and Applications*, pp. 453-461, 2012.

[2] Z. Zhu, Y. S. Ong, and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection," *Pattern Recognition*, vol. 40, pp. 3236-3248, 2007.

[3] P. Bermejo, J. A. Gámez, and J. M. Puerta, "A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets," *Pattern Recognition Letters*, vol. 32, pp. 701-711, 2011.

[4] T. R. Golub, *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, pp. 531-537, 1999.

[5] A. El Akadi, *et al.*, "A two-stage gene selection scheme utilizing MRMR filter and GA wrapper," *Knowledge and Information Systems*, vol. 26, pp. 487-500, 2011.

[6] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507-2517, 2007.

[7] C. L. Huang, "ACO-based hybrid classification system with feature subset selection and model parameters optimization," *Neurocomputing*, vol. 73, pp. 438-448, 2009.

[8] M. A. Tahir, A. Bouridane, and F. Kurugollu, "Simultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier," *Pattern Recognition Letters*, vol. 28, pp. 438-446, 2007.

[9] S. W. Lin, *et al.*, "A simulated-annealing-based approach for simultaneous parameter optimization and feature selection of back-propagation networks," *Expert Systems with Applications*, vol. 34, pp. 1491-1499, 2008.

[10] R. Agrawal and R. Bala, "A hybrid approach for selection of relevant features for microarray datasets," *Intl. J. Computer and Information Science and Engineering*, vol. 1, pp. 196-202, 2007.

[11] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine learning*, vol. 53, pp. 23-69, 2003.

[12] Y. Zhang, C. Ding, and T. Li, "A two-stage gene selection algorithm by combining reliefF and mRMR," pp. 164-171, 2007.

[13] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," 2003, pp. 523-528.

[14] T. M. Cover, "The best two independent measurements are not the two best," *Systems, Man and Cybernetics, IEEE Transactions on*, pp. 116-117, 1974.

[15] Z. W. Geem, J. H. Kim, and G. Loganathan, "A New Heuristic Optimization Algorithm: Harmony Search," *SIMULATION*, vol. 76, pp. 60-68, 2001.

[16] G. Z. Li, *et al.*, "Partial least squares based dimension reduction with gene selection for tumor classification," pp. 1439-1444, 2007.

[17] L. J. Zhang, Z. J. Li, and H. W. Chen, "An effective gene selection method based on relevance analysis and discernibility matrix," *Advances in Knowledge Discovery and Data Mining*, pp. 1088-1095, 2007.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[18] S. Pang, *et al.*, "Classification consistency analysis for bootstrapping gene selection," *Neural Computing & Applications*, vol. 16, pp. 527-539, 2007.

[19] B. Liu, *et al.*, "A combinational feature selection and ensemble neural network method for classification of gene expression data," *BMC bioinformatics*, vol. 5, p. 136, 2004.