ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

Mean Weighted Artificial Bee Colony (MWABC) based Feature Selection for Gene Co-Expression using Microarray Data

M.Sofia[1], Dr. N.Tajunisha[2]

[1]Ph.D Research Scholar, Dept. Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore

[2]Associate Professor, Dept. Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore

mmahalakshmims@gmail.com, tajkani@gmail.com

Abstract: In microarray data analyses, three important issues are how to determine incomplete data, how to choose genes, which provide reliable and good prediction for disease status, and how to determine the final gene set that is best for classification. To deal with redundant information and improve classification, propose a Mean Weight Artificial Bee Colony (MWABC) gene selection which combines ABC and mean weight function. First select a small subset of genes based on fuzzy and mean value of the attribute by considering the preference-ordered domains of the gene expression data. Propose an MWABC analysis to select discriminative genes and to use these genes to classify tissue samples of microarray data. Experiments show that the proposed MWABC is able to reach high classification accuracies with a small number of selected genes and its performance is robust to noise.

Keywords: Gene selection, microarray, classification, supervised-learning, Mean Weight Artificial Bee Colony (MWABC)

I. INTRODUCTION

Microarray technology allows simultaneous measurement of the expression levels of thousands of genes within a biological tissue sample. An important application of gene expression is to classify samples according to their gene expression profiles, such as the diagnosis or the classification of different types or subtypes of cancer [1]. Different classification methods from statistical and machine learning have been applied to the classification of cancer or clustering, are executed for profiling gene expression patterns [2]. Various standards related to systems biology are discussed by Brazma et al. [3]. When sample sizes are substantially smaller than the number of features/genes, statistical modeling and inference issues are challenging, which is known as the "large p small n problem".

To address the "curse of dimensionality" problem, three strategies have been proposed: filtering, wrapper and embedded methods. Filter methods evaluate a subset of genes by looking at the intrinsic characteristics of data with respect to class labels, while wrapper methods evaluate the goodness of a gene subset by the accuracy of its learning or classification. Embedded methods are generally referred to as algorithms, where gene selection is embedded in the construction of the classifier .Filtering methods select subset features independently from the learning classifiers and do not incorporate learning. Yet, the combination of these features may have a combined effect that does not necessarily follow from the individual performance of features in that group [4]. One of the consequences of filtering methods is that may end up with many highly correlated features/genes; this highly redundant information will worsen classification and prediction performance. Furthermore, if there is a limit on the number of features to be chosen, may not be able to include all informative features. However the advantages of wrapper methods it becomes very suitable for all applications.

Genetic Algorithm-Support Vector Machine (GA-SVM) [5] creates a population of chromosomes as binary strings that represent the subset of features that are evaluated using SVMs. The problem with genetic algorithms is that the time complexity becomes O(nlog(n) + nmpg), where n is the number of samples, m is the dimension of the data sets, p represents the population size, and g is the number of generations. In addition like all wrappers, randomized algorithms take up more CPU time and more memory to run. Classification is carried out to correctly classify the testing samples according to the class. Therefore, performing gene selection antecedent to classification would severely improve the prediction accuracy of the microarray data. Random forest is an ensemble classifier which uses recursive partitioning to generate many trees and then combine the result. Using a bagging technique first proposed by [6], each tree is independently constructed using a bootstrap sample of the data. Classification is known as discrimination in the statistical literature and as supervised

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

learning in the machine learning literature, and it generates gene expression profiles which can discriminate between different known cell types or conditions. In practice, discretization is a common preprocess before rough set based mining on gene expression data, which transforms continuous gene expression levels to categorical item sets [7]. Previous research has shown that handling uncertainty in such applications by the representation as interval data leads to accurate learning algorithms [8].

In this study, propose a Mean Weight Artificial Bee Colony (MWABC) method to select discriminative genes, and to use these genes to classify tissue samples of microarray data. In the gene selection step, objective is to determine the mean value of the attribute that discern between objects belonging to different classes. Tissue sample classification is based on the decision table generated from Trapezoidal Fuzzy Membership Valued and selected gene features from the microarray dataset.

II. LITERATURE REVIEW

Selection of relevant genes for sample classification is a common task in most gene expression studies. Liu at al [9] introduced an ensemble gene selection method based on the conditional mutual information for cancer microarray classification. Propose a new ensemble method, called Ensemble Gene Selection by Grouping (EGSG), to select multiple gene subsets for the classification purpose. Multiple gene subsets serve to train classifiers and outputs are combined by a voting approach. As a result, they do not robustly capture the non-parameterized structure shared among genes.

Likewise, Leung and Hung [10] initiated a Multiple-Filter-Multiple-Wrapper (MFMW) approach to gene selection to enhance the accuracy and robustness of the microarray data classification. Filters and wrappers have been combined in previous studies to maximize the classification accuracy for a chosen classifier with respect to a filtered set of genes. Some of MFMW-selected genes have been confirmed to be biomarkers or contribute to the development of particular cancers by other studies.

Bolón-Canedo et al [11] in another approach investigated a gene selection method encompassing an ensemble of filters and classifiers. A voting approach was employed to combine the outputs of classifiers that help reduce the variability of selected features in different classification domains. The outputs of these five classifiers are combined by simple voting. In this study three well-known classifiers were employed for the classification task: C4.5, naive-Bayes and IB1.

Recently, Du et al [12] suggested a forward gene selection algorithm to effectively select the most informative genes from microarray data. The algorithm combines the augmented data technique and L2-norm penalty to deal with the small samples' problem and group selection ability respectively. Finally, by defining a proper regression context, the proposed method can be fast implemented in the software, which significantly reduces computational burden Minimum Redundancy – Maximum Relevance (MRMR) feature selection framework is proposed by Ding & Peng [13] . Genes selected via MRMR provide a more balanced coverage of the space and capture broader characteristics of phenotypes. They lead to significantly improved class predictions in extensive experiments on 5 gene expression data sets: NCI, Lymphoma, Lung, Leukemia and Colon. Improvements are observed consistently among 4 classification methods: Naïve Bayes, Linear discriminant analysis, Logistic regression and Support vector machines. Wei et al [14] proposed a Model-Averaged Naive Bayes (MANB) method was applied to predict late onset Alzheimer's disease in 1411 individuals who each had 312,318 SNP measurements available as genome-wide predictive features. Its performance was compared to that of a Naive Bayes algorithm without feature selection (NB) and with Feature Selection (FSNB).

Performance of each algorithm was measured in terms of area under the ROC curve (AUC), calibration, and run time.

K nearest neighbor (KNN): KNN is a non-parametric classification method that predicts the sample of a test case as the majority vote among the k nearest neighbors of the test case [15]. To decide on "nearest" use the Euclidean distance. The number of neighbors used (k) is chosen by cross-validation for a given training set, the performance of the KNN for different values of k that produces the smallest error is used. Parry et al [15] focuses on the KNN modeling strategy and its clinical use. Although KNN is simple and clinically appealing, large performance variations were found among experienced data analysis teams in the MicroArray Quality Control Phase II (MAQC-II) project. For clinical end points and controls from breast cancer, neuroblastoma and multiple myeloma, we systematically generated 463,320 KNN models by varying feature ranking method, number of features, distance metric, number of neighbors, vote weighting and decision threshold.

Support Vector Machines (SVM): SVM are becoming increasingly popular classifiers in many areas, including microarrays . Mehenni and Moussaoui [16] propose a statistical method for selecting genes based on overlapping analysis of expression data across classes. This method results in a novel measure, called Proportional Overlapping Score (POS), of a feature's relevance to a classification task. Apply POS, along-with four widely used gene selection methods, to several benchmark gene expression datasets. The experimental results of classification error rates computed using the Random Forest, k Nearest Neighbor and

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

SVM classifiers show that POS achieves a better performance

III. METHODOLOGY

Microarray technology allows simultaneous measurement of the expression levels of thousands of genes within a biological tissue sample. However, high dimensionality and a small number of noisy samples pose great challenges to the existing methods. During high dimensionality problems, analysis of gene co expression is at the core of many types of genetic analysis. To solve these challenges in this research work a novel Mean Weight Artificial Bee Colony (MWABC) algorithm is proposed for solving high dimensionality problem. This research work not only handle the problem of feature selection, in addition these two problems also handled before performing co expression gene selection. The co expression between two genes can be calculated by using a traditional distance measures that is Euclidean distance measure. However, unobserved confounding effects may cause inflation of the Euclidean distance so that uncorrelated genes appear correlated. Since the data in real world application are often missing values. Missing values may generate bias and affect the quality of the supervised learning process or the performance of classification algorithm. But the quality of the data is major concern with machine learning and data mining. Missing value imputation is an efficient way to find or guess the missing values based on other information in the datasets. Many general methods have been suggested, which aim to solve missing data problem from microarray dataset.

Gene selection using Mean Weight Artificial Bee Colony (MWABC) algorithm

The Artificial Bee Colony (ABC) algorithm is used for real-time optimization is a recently introduced optimization algorithm which simulates the foraging behavior of a bee colony [17]. The minimal model of swarm-intelligent forage selection in a honey bee colony which the ABC algorithm simulates consists of three kinds of bees: employed bees, onlooker bees and scout bees.

Half of the colony consists of employed bees, and the other half includes onlooker bees. In this paper work consider employed bees corresponding gene expression matrix can be represented as $Y = (y_{i,j})_{m \times n}$, where $y_{i,j}$ is the expression level of gene g_i in sample ds_i. Employed bees are responsible for selection of gene features the nectar sources explored before and giving information to the onlooker bees. Scouts bees randomly search the environment in order to find a new selected gene features depending on fitness function or accuracy value of the classifier can be summarized as follows:

1. At the initial phase of the foraging process, the bees start to search the micro array dataset samples randomly in order to find highest classification accuracy for gene features.

2. After finding a highest classification accuracy, the bee (gene features) becomes an employed forager and starts to exploit the discovered source. The employed bee returns best gene features with the nectar and unloads the nectar. After unloading the nectar, go back to discover selected gene features site directly by performing a dance on the dance area. If it reaches maximum iterations selected gene features is exhausted, it becomes a scout and starts to randomly search for a new gene features.

3. Onlooker bees waiting in the hive watch the dances advertising the profitable gene features and choose a gene features depending on the classification accuracy of a dance proportional to the quality of the dataset samples $Y = (y_{i,j})_{m \times n}$.

Initial gene expression matrix samples are produced randomly within the range of the boundaries of the parameters.

$$z_{ij} = z_j^{min} + rand(0,1)(z_j^{max} - z_j^{min}) \quad (1)$$

where $i = 1 \ldots SN, j = 1 \ldots D. SN$ is the number of gene expression matrix samples (food sources) and D is the number of optimization parameters. After initialization, the population of the gene expression matrix samples is subjected to repeat cycles of the search processes of the employed bees, the onlooker bees and the scout bees. Termination criteria for the ABC algorithm might be reaching a Maximum Cycle Number (MCN). As mentioned earlier, each employed bee is associated with only one gene expression feature matrix. Hence, the number of gene expression matrix samples is equal to the number of employed bees. An employed bee produces a modification on the position of the gene expression features in her memory depending on local classification accuracy and finds a neighboring gene features, and then evaluates its quality. In ABC, finding a neighboring gene feature is defined by

$$v_{ij} = z_{ij} + \phi_{ij}(z_{ij} - z_{kj}) \quad (2)$$

Within the neigbourhood of every gene features represented by z_i, a food source v_{ij} is determined by changing one parameter of z_i. In Eq. (2), j is a random integer in the range [1,D] and $k \in \{1, 2, \ldots SN\}$ is a randomly chosen index. ϕ_{ij} is a uniformly distributed real random number in the range [-1, 1]. As can be seen from Eq. (2), as the difference between the parameters of the $z_{i,j}$ and $z_{k,j}$ decreases. Thus, as the

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

search approaches to the optimal feature solution in the search space, the step length is adaptively reduced. After producing v_{ij} within the cycle (MCN), a fitness value for a gene feature selection problem can be assigned to the solution v_{ij} by (3).

$$fitness_i = \begin{cases} \frac{1}{(1+f_i.w_i)} & if\ f_i.w_i \geq 0 \\ \frac{1}{(abs(f_i.w_i)} & if\ f_i.w_i < 0 \end{cases} \quad (3)$$

where f_i is the classification accuracy. For maximization problems, the cost function can be directly used as a fitness function. According to the ABC declaration, assigned a weight $W(a_i)$ to each attribute a_i. The value of weight $W(a_i)$ for each a_i, which is set to zero initially, is calculated sequentially throughout the whole matrix using the mean value of the attribute and update using the following formula when a new entry a_i is met in the discernibility matrix:

$$w_i = w(a_i).\mu(a_i) \quad (4)$$

After all employed bees complete their searches, they share their information related to the nectar amounts and the positions of their sources with the onlooker bees on the dance area. An onlooker bee evaluates the nectar information taken from all employed bees and chooses a best gene features with a highest probability in the gene matrix dataset samples is employed (5):

$$p_i = \frac{fitness_i}{\sum_{i=1}^{SN} fitness_i} \quad (5)$$

In the ABC algorithm, a random real number within the range [0,1] is generated for each source. If the probability value (p_i in Eq. (5)) associated with that source is greater than this random number then the onlooker bee produces a modification on the position of this selected gene feature site by using Eq. (3) as in the case of the employed bee. After the highest classification accuracy is evaluated, greedy selection is applied and the onlooker bee either memorizes the new gene features position by forgetting the old one or keeps the old one. If the selected gene features z_i cannot be improved, its counter holding trials is incremented by 1, otherwise, the counter is reset to 0. This process is repeated until all onlookers are distributed onto food source sites. If the value of the counter is greater than the control parameter of the ABC algorithm, the current gene features is assumed to be exhausted and is abandoned. Assume that the abandoned source is z_i, then the scout randomly discovers a new gene feature selection food source to be replaced with z_i.

$$v_{ij} = \begin{cases} z_{ij} + \mu_{ij}(z_{ij} - z_{kj}) & if\ R_{ij} < \mu_{ij} \\ z_{ij} & otherwise \end{cases} \quad (6)$$

However, the convergence rate of the ABC is poorer when working with constrained functions. In order to improve the convergence rate, ABC algorithm is modified by introducing a control parameter, mean value of the attribute(μ) number, ($0 < R_{ij} < 1$), is produced and if the random number is less than μ, then the parameter v_{ij}.

IV. RESULTS AND DISCUSSION

In this section, evaluate the discriminative performance of MWABC selected gene set on different classifiers. Also compare the performance of proposed IVC method to a wide range of standard classifiers: Support Vector Machine (SVM), K Nearest Neighbor (KNN), and Interval Value Classification (IVC). A set of experiments is conducted on the dataset by varying the number of genes selected to receive the highest classification accuracy. To evaluate the performance of the proposed method in practice, this research used the datasets containing gene expression profiles from patients with Acute Lymphoblastic Leukemia (ALL) and Acute Myeloblastic Leukemia (AML). The leukemia dataset is collected from the UCI Repository. This implementation of the various compared classifiers is based on the Weka (http://www.cs.waikato.ac.nz/ml/ weka/). For all the dataset, normalizations are performed so that every observed gene expression has a mean equal to 0 and a variance equal to 1. Table 2 provides a summary of the ALL/AML dataset.

TABLE II: SUMMARY OF THE ALL/AML DATASET

Dataset	Samples	Genes	Classes
ALL-AML-3	72	7129	3
ALL-AML-4	72	7129	4

Assessment Metrics In The Leukemia Datasets

Usually, the accuracy rate in Eq. (7) is the most frequently used measure in assessment metrics. But in the framework of the leukemia datasets, the accuracy is a proper measure, because it distinguishes between the numbers of correctly classified examples of different classes

Accuracy (Acc)=TP+TN/(TP+FN+FP+FN) (7)

TP rate=TP /(TP+FN) (8)

FP rate=FP /(FP+TN) (9)

Precision(P)=TP /(TP+FP) (10)

Recall(R)= TP rate (11)

Table III: Results of the ALL dataset

Methods	TP rate	FP rate	P	Fscore	Acc	ER
SVM	70.97	28.61	43.75	54.13	69.01	30.99
KNN	80.27	22.2	90	84.85	71.28	28.72
IVC	83.21	60.6	94.4	88.45	78.26	21.74
TFMVC	85.71	86.33	96.93	90.97	94.12	5.88

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

Table 3 shows the results for the leukemia dataset and Fig. 2 shows the performance comparison results of accuracy and error between existing and proposed algorithms for the leukemia dataset.

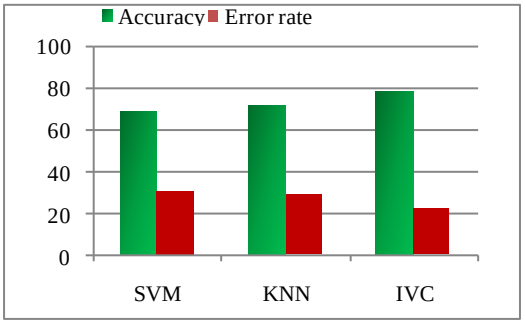

Fig. 1. Accuracy and error results comparison of methods for the leukemia dataset

From the experimental results it is inferred that for the leukemia dataset the proposed IVC algorithm performs 6.98 % better than the SVM algorithm, 9.25% better than the KNN algorithm is illustrated in Figure 1.

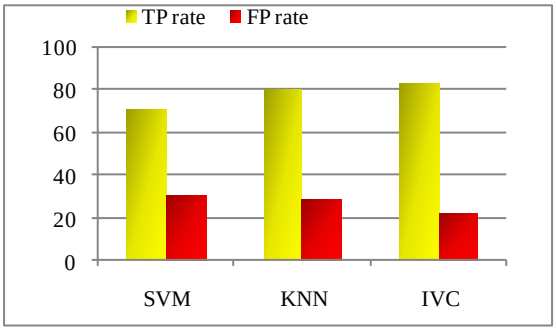

Fig. 2. TP rate and FP rate results comparison of methods for the leukemia dataset

From the experimental results it is conclude that TP rate for the leukemia dataset the proposed IVC algorithm performs 2.94% better than the SVM algorithm, 12.24% better than the KNN algorithm is illustrated in Figure 3. Similarly FP rate for the leukemia dataset the proposed IVC algorithm performs than the other classification methods is illustrated in Figure 2.

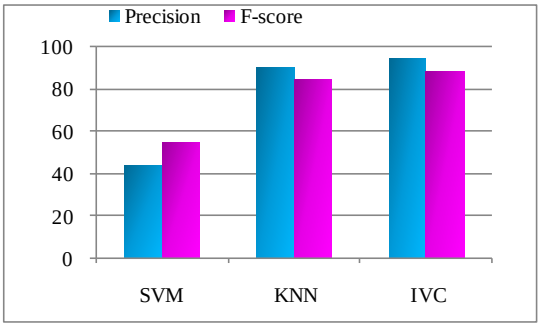

Fig. 3. F-measure results comparison of methods for the leukemia dataset

From the experimental results it is conclude that F-Measure for the leukemia dataset the proposed IVC algorithm performs 4.4% better than the KNN algorithm, 50.65% better than the SVM algorithm is illustrated in Figure 3.

V CONCLUSION AND FUTURE WORK

In this paper, propose a combination method of Mean Weight Artificial Bee Colony (MWABC) based gene selection and tissue classification of microarray data. During high dimensionality problems, analysis of gene coexpression is at the core of many types of genetic analysis. To solve these challenges in this research work a novel MWABC algorithm is proposed for solving high dimensionality problem. The results demonstrated that this approach reduces the number of genes selected and increases the classification accuracy rate. The performed various studies to compare the performance between different types of classifiers including Naive Bayes, k-NN, Decision Tree and SVM. The performances of all the methods were improved by the MWABC gene selection method. Though the experimental datasets are related to gene expression data, the method can be applied to other large datasets that require feature selection.

REFERENCES

[1] Wang X., R. Simon, Microarray-based cancer prediction using single genes, BMC Bioinformatics 12 (1) (2011) 391.
[2] Chen Z, McGee M, Liu Q, Scheuermann RH. A distribution free summarization method for Affymetrix GeneChip Arrays. Bioinformatics . 2007;23(3):321–327.
[3] Brazma A, Krestyaninova M, Sarkans U. Standards for system biology. Nat Rev Genet. 2006; 7:593–605.
[4] Yu J, Chen X-W. Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data. Bioinformatics . 2005;21(suppl 1):i487–i494
[5] Perez M., Marwala T. Microarray data feature selection using hybrid genetic algorithm simulated annealing. Proceedings of the IEEE 27th Convention of Electrical and Electronics Engineers in Israel, 2012; pp. 1–5.
[6] Revathy N., Balasubramanian R. GA-SVM wrapper approach for gene ranking and classification using expressions of very few genes. Journal of Theoretical and Applied Information Technology. 2012;40(2):113–119.
[7] Lee, J.W., Lee, J.B., Park, M., Song, S.H.: An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis 48, 869–885 (2004)
[8] Li, T., Zhang, C., Ogihara, M.: A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 20, 2429–2437 (2004)
[9] Liu H, Liu L, Zhang H (2010) Ensemble gene selection by grouping for microarray data classification. Journal of Biomedical Informatics, 43(1), 81–87.
[10] Leung Y, Hung Y (2010) A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 7(1), 108–117.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

[11] Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2012) An ensemble of filters and classifiers for microarray data classification. Pattern Recognition, 45(1), 531–539.
[12] Du D, Li K, Li X, Fei M (2014) A novel forward gene selection algorithm for microarray data. Neurocomputing. 133, 446–458.
[13] Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02), 185-205.
[14] Wei W., S. Visweswaran, and G. F. Cooper, "The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data," Journal of the American Medical Informatics Association, vol. 18, no. 4, pp. 370–375, 2011
[15] R. M. Parry, W. Jones, T. H. Stokes et al., "K-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction," Pharmacogenomics Journal,10(4), pp. 292–309, 2010.
[16] T. Mehenni and A. Moussaoui, "Data mining from multiple heterogeneous relational databases using decision tree classification," Pattern Recognition Letters, vol. 33, no. 13, pp. 1768–1775, 2012.
[17] Karaboga D., An Idea Based On Honey Bee Swarm for Numerical Optimization, Technical Report TR06, Erciyes University, Engineering Faculty, Computer Engineering Department, 2005.