RESEARCH ARTICLE OPEN ACCESS

SC-mRMR-FA Based Gene Selection Algorithm for Cancer Classification Using Microarray Gene Expression Profiling

P. Rajeshwari [1], N. Santhana Krishna [2]
Research Scholar [1], Assistant Professor (HOD) [2]
Department of Computer Science
AJK College of Arts and Science College, Coimbatore
Tamil Nadu - India

ABSTRACT

Micro-array gene expression data processing has become a major area of research due its applications in the medical field. The key uses of a microarray database are to store the measurement data, manage a searchable index, and make the data available to other applications for analysis and interpretation applications. However, the microarray dataset suffers from the curse of dimensionality, the limited number of samples, and the irrelevant and noise genes, all of which make the classification task for a given sample more challenging. In the proposed methodology, the gene selection process is improved further by utilizing the Spearman's rank correlation (SC) along with MI for feature selection. The ABC algorithm is replaced by a more efficient Firefly algorithm to form the proposed methodology SC-mRMR-FA. The proposed SC-mRMR-FA considers SC values between the genes in addition to the MI values for the predictive gene selection. The selection of genes is carried out using the Firefly algorithm with high accuracy. Finally the classification is carried out based on the FA selected genes using the SVM classifier. The evaluation results obtained on cancer dataset prove that the proposed SC-mRMR-FA Algorithm provides highly accurate cancer classification then the mRMR-ABC algorithm

Keywords: - Micro array, Gene Expression, Spearman co-relation, Firefly Algorithm.

I. INTRODUCTION

Biomedical engineering (BME) is the application of engineering principles and design concepts to medicine and biology for healthcare purposes (e.g. diagnostic or therapeutic). This field seeks to close the gap between engineering and medicine: It combines the design and problem solving skills of engineering with medical and biological sciences to advance healthcaretreatment, ncluding diagnosis, monitoring, and therapy [1].

BME has only recently emerged as its own study, compared to many other engineering fields. Such an evolution is common as a new field transitions from being interdisciplinary specialization among already-established fields, to being considered a field in itself. Much of the work in BME consists of research and development, spanning a broad array of subfields. Prominent biomedical engineering applications include the development of biocompatible prostheses, various diagnostic and therapeutic medical devices ranging from clinical equipment to micro-implants, common imaging equipment such as MRIs and EEGs, regenerative tissue growth, pharmaceutical drugs and therapeutic biological.

1. *Gene Expression Data*

Gene expression experiments provide the expression levels of thousands of genes. Computational analyses of such datasets present exciting opportunities for studying cancer and various genetic diseases. Gene expression datasets

have a complex structure owing to lesser number of samples in comparison to an extremely large attribute/feature space having high noise levels, irrelevant attributes, missing values, several outliers and sample variance. Computational analysis of such data can help derive informative genes which may be responsible for a certain disease. The problem of gene selection thus concerns itself with, identifying a subset of relevant genes (or input variables), that can help in building a robust classification model for the target disease [3].

1.2 Gene selection methods

Feature selection is a process commonly used in machine learning, wherein a subset of the features available from the data is selected for application of a learning algorithm [11]. The best subset contains the least number of dimensions that most contribute to accuracy. It discards

the remaining, unimportant dimensions. This is an important stage of pre-processing and is one of two ways of avoiding the curse of dimensionality. There are two approaches in feature selection namely forward selection and backward selection. Forward Selection starts with no variables and adds them one by one, at each step adding the one that decreases the error the most, until any further addition does not significantly decrease the error. Backward Selection starts with all the variables and removes them one by one, at each step removing the one that decreases the error the most, until any further removal increases the error significantly.

II. PROPOSED SYSTEM

Irrelevant, noisy and high dimensional data, containing large number of features, degrades the performance of data mining and machine learning tasks. One of the methods used in data mining to reduce the dimensionality of data is feature selection. Feature selection methods select a subset of features that represents original features in problem domain with high accuracy. Various methods have been proposed that utilize heuristic or nature inspired strategies to find these subsets. However these methods either consume more time to find subset or compromise with optimality.

2.1 Firefly Algorithm

In every iteration the brighter firefly applies its manipulated over additional fireflies and attracts them towards itself in maximization problems. In fact, in the standard FA, fireflies move despite of the global optima and it reduce the ability of the firefly algorithm to Discover global best. In this work, to remove performance shown on various expression datasets than by other criterion, such as the difference between relevancy and redundancy. After selecting the top-ranked genes, the subsequent genes are selected by forward selection, maximizing the criterion. Thus it selects features that have the highest relevance with the target class and are also minimally redundant, i.e. it selects features that are maximally dissimilar to each other. Both optimization criteria (maximum-relevance and minimum-redundancy) are based on mutual information.

2.2 Sc-FA Procedure

In this module, the Spearmen Correlation FA is used for finding important genes from the given datasets. Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data. Spearman rank correlation estimates for pairwise comparisons and it tests the correlation between two values.

The higher values indicate the Spearman rank correlation coefficient calculation for that test. Spearman rank correlation is able to detect similar and even time-shifted sub genes. The correlation coefficient is very simple to calculate and easy to use. When measurement errors exist in data, the Spearman rank correlation coefficient, utilizing the ranks of the data, is preferred. The Spearman rank correlation coefficient, between two profiles of gene i and gene j, is given by

$$S_{i,j} = 1 - \frac{x}{n(n^2-1)} \sum_{k=1}^{n} \{r_{x_i}(x_{i,tk}) - r_{xj}(x_{j,tk})\}^2 \quad (4.3)$$

Where $r_{xj}(x_{j,tk})$ is the rank of $x_{i,tk}$ in the gene $x_i = (x_{i,t1}, x_{i,t2}, \dots \dots x_{i,tn})$, $i = 1, 2, \dots p$. The Spearman correlation coefficient measures the strength of the curvilinear monotonic relationship between two genes. That is, it can measure the monotonic association between the profiles even when they do not show a linear relationship .

The same correlation coefficients, as well, can be obtained for similar and different pairs of profiles. A large coefficient does not necessarily indicate two similarly shaped profiles, nor does a small coefficient necessarily confirm differently shaped profiles. Therefore the correlation coefficient may not be a reliable measure of association when the experimental time points are few. In order to remedy the shortcomings of the correlation coefficient, it propose firefly algorithm to optimize the results more accurately weaknesses of FA and get better the collective movement of fireflies, propose a Firefly Algorithm (FA). Fireflies are unisex so that one firefly will be attracted to other fireflies regardless of their sex. In the proposed algorithm, use global optima in firefly's progress. Global optimum is associated to optimization difficulty and it is a firefly that has the greatest or smallest amount value. And the global optima will be modernized in any iteration of algorithm. In the proposed approach, when a firefly compare with another firefly as an alternative of the one firefly being permitted to authority and to attract its neighbors, global optima in every iteration is permitted to influence others and change in their association. In the MFA, when a firefly perform with correspond firefly, if the match firefly be brighter, the compared firefly will shift toward correspond firefly, measured by global optima.

FA is a nature inspired multi-modal metaheuristic algorithm based on the firefly's flashing behavior . Firefly uses the flashing as a signal to attract other fireflies. A FA assumes three basic rules [which are described as follows:

• Every firefly will be attracted to other fireflies irrespective to their gender because they are unisexual

• They attract each other proportionally to their illumination intensity and reversely proportional to their search spaces, the brighter flashing firefly will attract the other less bright ones, the more the distance the less attractiveness, if no brighter firefly nearby they will move randomly

• The brightest firefly cannot be attracted and it will travel randomly

The preprocessing helps us to improve data efficiency and remove the noisy data which helps to identify the survival of the fittest. The preprocessing is to clean the original data and extract the useful information from the data set. Various data sets have been taken for preprocessing such that MRMR filter the response of the data set has been tested under various classification algorithms. The classification algorithm which gives best accuracy for the data set is taken as input to the genetic algorithm for the next process of identifying the survival of the fittest. The authors employ data preprocessing step to apply piecewise regression as a predictive data mining technique that fits a data model which will be used for prediction. The MRMR method aims at selecting maximally relevant and minimally redundant set of genes for discriminating tissue classes. In MRMR method, gene ranking is performed by optimizing the ratio of the relevancy of a gene to the redundancy of

Where $I(l,i)$ is the mutual information between class labels l and gene i, where the summation is taken over the space of gene expression values. The redundancy of a gene subset is determined by the mutual information among the genes. The redundancy of gene i with the other genes in the subset S is given by

In the proposed algorithm the fitness function is used to indicate how good or bad a gene solution is. The way of selecting the fitness function is a very significant matter in designing the proposed feature selection and classification algorithm, since the solution optimization as well as the performance of the algorithm count mainly on this fitness function. Thus the solutions will be ordered in ascending way after measuring their fitness function based on their fitness value. In the proposed clustering algorithm (a firefly (brighter one) that have minimum fitness value) for each iteration will has the ability to affect and influence in the movements of the other fireflies. Therefore when comparing between two fireflies a and b, if b is brighter than firefly a, than firefly a will move toward firefly b. The proposed algorithm is designed to enhance the performance of the SVM in order to obtain more accurate classification process.

In order to update the solution with the newly generated a' = (a'1, a'2, a'3,....,a'N) the objective function is calculated for every new firefly solution f (a'), if the objective function value of the new solution is better than the current solution, then the worst solution will be replaced by the new solution, otherwise the new solution will not be used.

2.3 *Classification Process Based On Sc-mRMR-FA*

SVMs are supervised learners that construct a model from available training data with known classification. In order to obtain accurate class predictions SVMs provide a number of free parameters that have to be tuned to reflect the requirements of the given task. The SVM can be characterized as a supervised learning algorithm capable of solving linear and non-linear binary classification problems. Given a training set with m patterns $(x_i, y_i)\}_{i=1}^{m}$ where $x_i \in X \subseteq \square^n$ is an input vector and $y_i \in \{-1, +1\}$ its corresponding binary class label, the idea of support vector classification is to separate examples by means of a maximal margin hyperplane. Therefore, the algorithm strives to maximize the distance between genes that are closest to the decision surface.

The fitness function is an important factor for evaluation and evolution of SVMs providing satisfactory and stable results in real-world applications. The fitness function guides the superordinated evolutionary learning process determining the probability that an individual can hand down genetic information to the subsequent population. Therefore, it should express the user's objective and should favor SVMs with satisfactory generalization ability in order to select useful classifiers systematically instead of accidentally. Consequently, the fitness function effectively conducts model selection and it can incorporate arbitrary model selection criteria as fitness measure. Whereas the fitness function selects solutions for light intensity and absorption coefficient. The selection is implemented as tournament selection with a tournament size of two. Furthermore, an elitist mechanism is applied in order to ensure that the best SVM is member of the next generation.

The idea of SC-mRMR-FA based SVM model selection one chooses the individual with maximum overall fitness for future use on unseen data. To simulate this scenario, it assessed the performance by means of precision, recall, f-measure and accuracy metrics. The important gene selection data is used to increase the cancer classification accuracy using SC mRMR-FA algorithm. optimization algorithm is used to select the most significant genes from the given cancer datasets.

The overall architecture diagram of the proposed system is shown in figure 4.1. It describes about the proposed preprocessing step, feature selection and classification using effective approaches. In the proposed system, the mRMR approach is used to filter the unnecessary genes which are focused to improve the cancer classification accuracy results.

The firefly algorithm is finding and selecting the most informative genes using best objective function. Then apply the SVM classification algorithm to predict the more accurate results. The result proves that the proposed SC-mRMR-FA algorithm is better than the existing algorithm in terms of higher performance.

III. EXPERIMENTAL SETUP AND RESULTS

To implement the proposed system and generate various results the scenario use MATLAB in this environment. The scenario has been selected Leukemia 1, dataset.

Leukemia 1 dataset

Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}}$$

An accuracy of 100% means that the measured values are exactly the same as the given values.

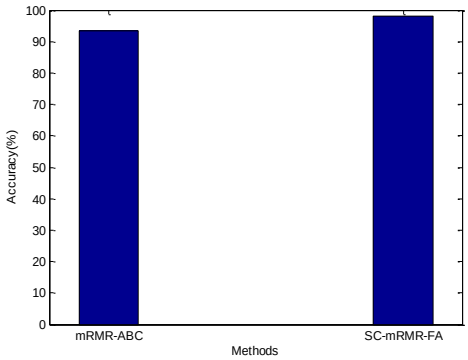

Fig 3.1 Accuracy comparison

From the above fig 5.1 it can observe that the comparison of existing and proposed system in terms of accuracy metric. In x axis plot the types and in y axis plot the accuracy values. In existing scenario, the accuracy values are lower by using mRMR-ABC algorithms. The accuracy value of existing scenario is 93.3507% for given Leukemia 1 dataset using mRMR-ABC algorithm. In proposed system, the accuracy value is higher by using the SC-mRMR-FA algorithm. The accuracy value of proposed scenario is 98.1746% for given Leukemia 1 dataset. From the result, it concludes that proposed system is superior in performance.

precision

The precision is calculated as follows:

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

Precision can be seen as a measure of exactness

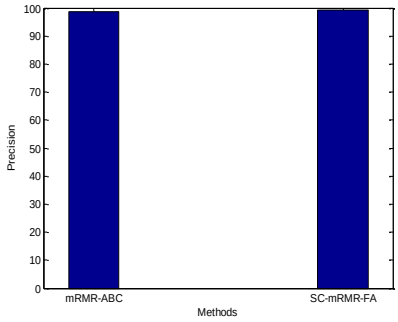

Fig 3.2 Precision comparison

REFERENCES

[1] S. Bandyopadhyay, U. Maulik, and D. Roy, "Gene identification: Classical and computational intelligence approaches," IEEE Trans. Syst., Man, Cybern. C, vol. 38, no. 1, pp. 55–68, Jan. 2008.

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

[3] C. Ambroise, G. Mclachlan, Selection bias in gene extraction on the basis of microarray geneexpression data, Proc. Natl. Acad. Sci. 99 (10) (2002) 6562– 6566.

[4] R. Breitling, P. Armengaud, A. Amtmann, P. Herzyk, Rank products: a simple, yet powerful, new method to detect differentially regulated

genes in replicated microarray experiments, FEBS Lett. 573 (1–3) (2004) 83– 92.

[5] Carlos J. Alonso-González, Q. Isaac Moro-Sancho," Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods", an international journal Expert Systems with Applications 39 (2012) 7270–7280

[6] A. Abderrahim, E. Talbi, and M. Khaled, "Hybridization of genetic and quantum algorithm for gene selection and classification of microarray data," in Proc. IEEE International Symposium In Parallel Distributed Processing, , pp. 1–8, 2009.

[7] E. Alba,J. Garcia-Nieto et al., "Gene selection in cancer classification using pso/svm and ga/svm hybrid algorithms," Evolutionary Computation, pp. 284–290, 2007.

[8] C. Alonso, I. Moro-Sancho, A. Simon-Hurtado, and R. Varela-Arrabal, "Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods," Expert Systems with Applications, vol. 39, no. 8, pp. 7270 – 7280, 2012.

[9] H.-L. Huang and F.-L. Chang, "Esvm: Evolutionary support vector machine for automatic feature selection and classification of microarray data," Biosystems, vol. 90, no. 2, pp. 516–528, 2007.

[10] Guyon and A. Elissee, "An introduction to variable and feature selection," Journal of Machine Learning Research, pages: 1157-1182, 2003.

[11] C.Deisy, B.Subbulakshmi, S.Baskar, and R.Ramaraj , " Efficient dimensionality reduction approaches for feature selection ," Proceedings of the International Conference on Computational Intelligence and Multimedia Applications, pages: 121-
127, 2007.