AN EFFICIENT TECHNIQUE TO CANCER CLASSIFICATION USING FAST IMPROVED BACTERIAL FORAGING OPTIMIZATION

Sindhu.V
Department of Computer Science, Karpagam University

Abstract—Microarray data play an essential role in the development of efficient cancer diagnoses and classification system with gene expression data. These expression data are usually unnecessary and noisy, and only a subset of them present discrete profiles for different classes of samples. Thus, selecting high discriminative genes from gene appearance data has become increasingly in the field of bioinformatics. In my work, a multi-objective biogeography based optimization technique is proposed to select the small subset of informative relevant to the classification. To choose multiple highly revealing gene subsets for cancer classification and diagnosis, a Fast improved Bacterial Foraging (FIBF) optimization is proposed to eliminate inappropriate, unnecessary, or noisy genes in changed granules at different stages and selects very informative genes with potentially different biological functions in balance. To show the efficiency of the expectations approach, we calculate the appearance of this technique with the sensitivity, specificity and accuracy. Using gene microarray datasets dataset is leukaemia (including both binary and multi-class classification problems), we reveal experimentally that our proposed scheme can recognize significant empirical success and is biologically related for cancer diagnosis and drug.

Keywords—Microarray data; leukaemia; Bacterial Foraging; swarm optimization; transductive; Sensitivity, Specificity and accuracy; Naïve bayes; Random forest, Decision stump.

I. INTRODUCTION
Cancer is also known as a malignant tumor or a malignant neoplasm, is a group of diseases concerning a typical cell growth with the potential to invade or spread to other parts of the body [1][2]. Causes of cancer are dangerous which including genetic factors like lifestyle of the humans such as tobacco use, diet, and physical action; certain types of infections; and environmental exposures to different types of chemicals and radiation [3]. Different types of cancer are bladder cancer, Brain tumor etc which can be have 60 and more types in cancer for both children’s and adults. Recent advances in microarray technology allow scientists to measure the expression levels of thousands of genes simultaneously in biological organisms and have made it possible to create databases of cancerous tissues. It ultimately produces gene expression information that contain useful information of genomic, diagnostic, and prognostic to select informative genes that contribute to a cancerous state. However, the gene selection process poses a major challenge because of the following characteristics of gene expression data [4]: the huge number of genes compared to the small number of samples (high-dimensional data), immaterial genes, and noisy data. To overcome this challenge, a gene selection technique is used to select a subset of informative genes that maximizes classifier’s ability to classify samples more accurately [5]. In computational intelligence domain, gene selection is called feature selection.

In classification of gene expression data, selecting a smaller subset of informative genes from thousands of genes is a critical step for accurate cancer classification. In the context of cancer classification, gene selection methods can be classified into two categories [6]. If a gene selection method is carried out independently from a classification procedure, it belongs to the filter
method. Otherwise, it is said to track a hybrid (wrapper) method. In the early era of microarrays analysis, most previous works have used the filter method to select genes since it is computationally more efficient than the hybrid method. Various filter methods are usually mentioned as individual gene-ranking methods such as $t$-test, signal-to-noise-ratio, information gain, etc. Cancer classification using gene expression data usually relies on traditional supervised learning techniques, in which only labeled data (i.e., data from a sample with clinical follow-up) can be exploited for learning, while unlabeled data (i.e., data from a sample without clinical follow-up) are disregarded[7]. Recent research in the area of cancer diagnosis suggests that unlabeled data, in addition to the small number of labeled data, can construct significant improvement in accuracy [8], a technique called semisupervised learning [9]. They evaluate a gene based on its discriminative power for the target classes without considering its correlations with previous genes. This means may possibly result in inclusion of unrelated and noisy genes in a gene subset for the cancer classification. The unrelated and noisy genes reduce the classification accuracy. Meanwhile, these genes also increase the dimensionality of the gene subset and, in turn, rise their computational time. At the moment, several hybrid methods, especially a combination between particle swarm optimization (PSO) and a classifier, have been implemented to choose enlightening genes. The hybrid methods usually provide greater accuracy than the filter methods since the genes are selected by considering and optimizing correlations among genes. Recently, several gene selection methods based on PSO have been proposed to select informative genes from gene expression data. PSO is a new population-based stochastic optimization technique. This approach produced 100% classification accuracy in many datasets, but it used a high number of selected genes (large gene subset) to achieve the high accuracy. It uses the high number because of the global best particle is reset to zero position when its fitness values do not change after three consecutive iterations. A hybrid of PSO and PCA has best classification for the same purpose. Unfortunately, the accuracy result is still not high and numerous genes are certain for cancer classification since there are no direct probability relations between PCA and PSO. Generally, the PSO-based methods are intractable to efficiently produce a small (near-optimal) subset of informative genes for high classification accuracy. This is generally because the total number of genes in gene expression data is too large (high-dimensional data). Present work of the cancer Diagnosis is carried out Optimization of the bacterial foraging which achieves the performance of the system to the good results in terms of Sensitivity and Specificity. The diagnostic goal is to develop a medical procedure based on the least number of possible genes that needed to detect diseases. The remainder of the Sections is organized as section 2 explains the related work and it is followed by proposed System in the section 3, finally the Section 4 proves the experimental part and it is concluded with section 5.

II. RELATED WORKS

2.1 Fuzzy Preference for feature classification:
In practice, fuzzy rough set models [1] address three key issues: Inducing a granular structure on the universe based on an attribute or a criterion; ii) aggregating the granular structures obtained from different attributes or criteria, and iii) determining the lower and upper approximations of decisions. Two kinds of fuzzy preference relations are used in practice for a variety of decision-making models [2]: i) multiplicative preference relation and ii) fuzzy preference relation three techniques for fuzzy preference modeling are discussed in the paper. The first technique is based on the construction of membership functions of subsets of no dominated alternatives with simultaneous considering of all criteria (fuzzy preference relations). The second method is of a lexicographic character and consists of step-by-step introducing of fuzzy preference relations [12]. The third technique is based on aggregating membership functions of subsets of no dominated alternatives corresponding to each preference relation [13]. These techniques have served for developing a corresponding system for multiobjective decision making (MDMS).
2.2 Independent Component Analysis

Independent component analysis (ICA) is a statistical method for transforming an observed multidimensional random vector into components that are statistically as independent from each other as possible. We use a combination of two different approaches for linear ICA: Common’s information theoretic approach and the projection detection approach. Using maximum entropy approximations of differential entropy, we establish a family of new contrast functions for ICA. The statistical properties of the estimators based on such contrast functions are analyzed under the assumption of the linear combination model, and it is shown how to choose contrast functions that are robust and/or of minimum variance.

III. PROPOSED MODEL

3.1 Extracting the Feature using the Fast improved Bacterial Foraging

Fast improved bacterial foraging is used for finding more relevant gene markers from microarray gene expression data. The feature selection reflects the degree of preference quantitatively making it more powerful in extracting information from microarray data than equivalence or dominance relations. Features is obtained by the equivalence relation between the two set obtained the component extraction. In classification analysis for description, the concepts obtained by condition attributes are used to approximately describe the decision.

This approach involves a modification of traditional mathematical programming methods and is associated with formulating and solving one and the same problem within the framework of mutually interconnected models. The use of the approach allows one to maximally cut off dominated alternatives. The following contraction of the decision uncertainty region is based on reducing the problem to models of multiobjective choosing alternatives in a cancer environment with the use of SVM techniques for analyzing these models.

3.2 Knowledge based SVM classification through features

SVM is learning machine based on two key elements: a general purpose learning algorithm and a problem specific kernel that computes the inner product of input data points in a feature space. SVM was initially developed as two class pattern recognition problem [10] which has been extended to the multi-class problem. To ease the difficulty of small-size training set, transductive support vector machine (TSVM) was projected. TSVMs seek largest partition in presence of both labelled and unlabeled data through regularization. At the early iteration, the standard SVMs are used to obtain an initial discriminating hyper plane based on the labeled data alone. The pseudo labels are then assigned to the unlabeled samples.

Figure 3.1. Architecture Diagram of the Bacterial foraging technique for Cancer Classification
These are called semi labeled samples. Subsequently, transductive samples are particular from the semilabeled samples according to a given criterion. Fig3.1 explains the overall process involved in the cancer classification from the micro array data.

Cancer classification is initially classified for features using SVM and results of the SVM is passed to Principle component analysis for the classification and clustering of the data based on dimensionality reduction, the possible maximum variance is estimated by the multivariate analysis.

A hybrid training set is thus obtained consisting of the original labeled set and transductive set. The resulting hybrid training set is used at the following iterations to find a more reliable separating hyperplane [11]. Training the TSVM algorithm can be almost outlined as the following steps:

**Input:** An early training set and an unlabeled set.

**Output:** Transductive SVM classifier with initial training set and a transductive set [14].

**Step 1:** Specify and; execute an initial inductive learning using all labeled samples, and obtain an initial SVM classifier.

**Step 2:** Compute the decision function values of all the unlabeled samples with the initial classifier. Attain label vector of all the unlabeled examples. Choose all the positive and negative semilabeled points within the margin band as transductive samples and add them to the initial training set to obtain a hybrid training set.

**Step 3:** Retrain the SVM using this hybrid training set. Compute the decision function values of all the unlabeled samples [16]. Obtain the label vector of the unlabeled samples. Select all the positive and negative semilabeled points within the margin band as transductive samples.

In cancer gene expression datasets, it is common that only some of the samples have sufficient clinical follow-up data and others are unlabeled with regard to the clinical question of interest [15]. Therefore, we investigated whether integrating unlabeled data from the same dataset could improve the prediction performance.

### IV. EXPERIMENTAL RESULTS

In this sector, we analyses the performance of the system using the microarray data, since classification is a typical and fundamental issue in diagnostic and prognostic prediction of cancer, we applied three feature selection methods to identify genes that have the discrimination capability to be used as gene markers. Our results demonstrated significant potential of semisupervised learning in the domain of clinical problems with prevalent techniques like Particle swarm optimization for the optimization of the SVM for the cancer diagnosis.

Feature selection is a useful technique in dealing with dimensionality reduction. In classification, it is used to find an optimal subset of relevant features so that the overall accuracy is increased while the data size is reduced. When a classification problem is defined by features, the number of description can be quite large, many of which can be irrelevant. A relevant feature can enhance the performance of a classifier while an irrelevant feature can deteriorate it. Therefore, in order to select the relevant features, it is necessary to measure the goodness of selected features using a feature selection criterion. The class separability is often used as one of the basic selection criteria. In this study, reliability measure is exploited as a selection criterion that does not attempt to maximize the class separability but aims to retain the discriminatory power of the original features. Top ranked genes are extracted from each dataset that are mostly responsible for distinguishing a particular tumor class from the remaining ones. In addition to the small-size training samples, unlabeled data
can be also exploited to increase the classification performance. In the experimental set up, we measured the labelled datasets which are roughly divided into training and test/validation sets.

![Figure 4.1 Comparison of the Cancer Classification technique using Fast Improved Bacterial forging](image)

In Figure 4.1, we explain the different technique outcomes in the bar chart to depict its accuracy against the Sensitivity, Specificity and accuracy.

### Table 4.1 Accuracy and testing for existing and proposed algorithm

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correctly Classified Instances</th>
<th>Testing Percentage split 70%</th>
<th>Testing Percentage split 80%</th>
<th>Testing Percentage split 90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NaiveBayes</td>
<td>82.34</td>
<td>82.26</td>
<td>80.85</td>
<td>76.59</td>
</tr>
<tr>
<td>Random Forest</td>
<td>91.91</td>
<td>76.59</td>
<td>71.27</td>
<td>72.34</td>
</tr>
<tr>
<td>DecisionStump</td>
<td>85.10</td>
<td>84.39</td>
<td>82.97</td>
<td>76.59</td>
</tr>
<tr>
<td>SVM with Fuzzy Preference</td>
<td>94.8</td>
<td>82.97</td>
<td>81.91</td>
<td>76.59</td>
</tr>
<tr>
<td>Fast Improved Bacterial Foraging</td>
<td>98.2</td>
<td>85.97</td>
<td>83.53</td>
<td>77.21</td>
</tr>
</tbody>
</table>

### Table 5.2 Performance of the Cancer Classification Technique

<table>
<thead>
<tr>
<th>Attributes</th>
<th>NaiveBayes</th>
<th>Random Forest</th>
<th>DecisionStump</th>
<th>SVM with Fuzzy Preference</th>
<th>Fast Improved Bacterial Foraging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>82.34</td>
<td>95.65</td>
<td>85.1</td>
<td>94.8</td>
<td>98.2</td>
</tr>
<tr>
<td>Precision</td>
<td>0.763</td>
<td>0.81</td>
<td>0.712</td>
<td>0.76</td>
<td>0.78</td>
</tr>
<tr>
<td>Recall</td>
<td>0.823</td>
<td>0.855</td>
<td>0.844</td>
<td>0.65</td>
<td>0.88</td>
</tr>
<tr>
<td>FMeasure</td>
<td>0.782</td>
<td>0.813</td>
<td>0.773</td>
<td>0.75</td>
<td>0.80</td>
</tr>
</tbody>
</table>

It is evident that the proposed method is significantly better than the homogenous Methods.

Due to the fact that support vectors contain the richest information among the informative samples (i.e., the ones in the margin band), the unlabeled patterns closest to the margin bounds have the highest probability to be exactly classified. Therefore, in the proposed approach, we plan a
selection procedure (i.e., filtering process) to increase the acceptability of the samples with the expected correct labeling.

V. CONCLUSION

We have analyzed a combination of fuzzy preference based feature selection and semi supervised SVM to address the problem of cancer classification and we designed and implemented a fast improved bacterial foraging optimization to address the problem of the cancer classification through the feature selection and feature extraction through SVM. In addition, we have compared the cross-validation accuracies among the various classification algorithm combinations of dimensionality reduction techniques in weka tool. When using the weka tool, it shows the correctly classified percentage values for all varieties. SVM with fuzzy and Naïve Bayes classification gives more accuracy than a Naïve Bayes classification algorithm. The rule induction generates only correct rules based on the accuracy. Random Forest algorithm shows the different models and each model gives different results. Random Forest outperforms than other classification algorithms instead of selecting all the attributes for classification. In this analysis, finally we had found that the Fast improved Bacterial foraging is best for the efficient cancer classification algorithms.

REFERENCES