Computational Biology and Chemistry 71 (2017) 161–169

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research article

A novel approach for dimension reduction of microarray

Rabia Aziz*, C.K. Verma, Namita Srivastava

Department of Mathematics & Computer Application, Maulana Azad National Institute of Technology, Bhopal, M.P., 462003, India

ARTICLE INFO

Article history:
Received 24 August 2016
Received in revised form 4 January 2017
Accepted 27 October 2017
Available online 28 October 2017

Keywords:
Feature selection (FS)
Artificial bee colony (ABC)
Independent component analysis (ICA)
Naïve bayes (NB)
Cancer classification

ABSTRACT

This paper proposes a new hybrid search technique for feature (gene) selection (FS) using Independent component analysis (ICA) and Artificial Bee Colony (ABC) called ICA + ABC, to select informative genes based on a Naïve Bayes (NB) algorithm. An important trait of this technique is the optimization of ICA feature vector using ABC. ICA + ABC is a hybrid search algorithm that combines the benefits of extraction approach, to reduce the size of data and wrapper approach, to optimize the reduced feature vectors. This hybrid search technique is facilitated by evaluating the performance of ICA + ABC on six standard gene expression datasets of classification. Extensive experiments were conducted to compare the performance of ICA + ABC with the results obtained from recently published Minimum Redundancy Maximum Relevance (mRMR) +ABC algorithm for NB classifier. Also to check the performance that how ICA + ABC works as feature selection with NB classifier, compared the combination of ICA with popular filter techniques and with other similar bio inspired algorithm such as Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). The result shows that ICA + ABC has a significant ability to generate small subsets of genes from the ICA feature vector, that significantly improve the classification accuracy of NB classifier compared to other previously suggested methods.

1. Introduction

Microarray technology has shifted to a new era in molecular classification (Chen et al., 2014; García and Sánchez 2015; Ng and Pei 2007). Classification is a type of technique in which a set of previously classified examples is given and the classifiers work is to find a rule that will agree to assign new samples to any one of the predefined classes. Some of the examples of such methods that have been used in molecular level cancer classification in recent years include the Decision Tree (Tan et al., 2006), Neural Networks (Khan et al., 2001), Support Vector Machine (Guyon et al., 2002) and the Naïve Bayesian classifier (Keller et al., 2000). Since the number of genes is very large compared with a very few number of samples in the microarray data, hence to get trained, is the main difficulty for most of the machine learning technique. This problem is known as 'curse of dimensionality'. The solution for this problem is dimension reduction, which plays an important role in the DNA microarray classification. In order to enhance accuracy and speedup model training time for the classifier, dimension reduction attempts to select a subset of the most significant features by rejecting inappropriate and redundant features (Lazar et al., 2012; Saeys et al., 2007). For dimension reduction, there are two important algorithm feature extraction and feature selection. Feature extraction algorithm transforms the feature into the lower dimension space by using the combinations of the original features. The feature selection method selects the most relevant features from the entire features to construct the model for classification. Feature selection algorithms can be arranged into three types, namely filter, wrapper and embedded methods. Filter methods are the ones that select features as a pre-processing step, they select features without considering the classification accuracy, while the wrapper method is an repetitive search process in which the results of the learning algorithm in each repetition are used to guide the search process (Kohavi and John 1997). Since wrapper methods continuously use the learning algorithm in the search process so their computational cost is more, especially for high dimensional datasets. The third type of feature selection algorithm is embedded method. The difference of embedded approach compares to other filter and wrapper approach is the search mechanism that is built into the classifier model. Recently a hybrid search technique has been used to take the advantages of the extraction/filter and the wrapper approach. In the hybrid search algorithm, the first subset of features is selected or extracted based on the filter/extraction method and after that the wrapper method is used to select the final feature set.

* Corresponding author.
E-mail address: rabia.aziz2010@gmail.com (R. Aziz).

https://doi.org/10.1016/j.compbiolchem.2017.10.009
1476-9271/

Therefore the computational cost of the wrapper method becomes acceptable due to the use of reduced size features. Information gain and a memetic algorithm (Zibakhsh and Abadeh 2013), Fishers core with a GA and PSO (Zhao et al., 2011), mRMR with ABC algorithm (Alshamlan et al., 2015a) and Indecomponent component analysis with fuzzy algorithm (Aziz et al., 2016),are recently used hybrid method to solve the problem of dimension reduction of microarray.

ICA is a multi-dimensional statistical method for finding the hidden information that situated under a set of random variables (Hyvarinen and Karhunen 2001). Now a days ICA technique have received growing attention as effective dimension reduction algorithm for NB classification of high dimensional data (Kong et al., 2008). The reason for that is the conditional independence hypothesis rooted in the algorithm of NB classifier could be successfully resolved as the components extracted by the ICA are statistically independent. There still exists an unsolved problem that, how one can chooses a subset of Independent Component (IC) that improves the performance of base classifier. To solve this problem, different authors used different wrapper methods to choose best subset of the IC. For example, the Sequential Floating Forward Selection (SFFS) method is used to choose the ICA feature vector for SVM classification of microarray data (Zheng et al., 2006). Zheng et al. classifies gene expression data with consensus independent component analysis as a dimension reduction technique (Zheng et al., 2008). A sequential feature extraction method was used to choose best genes set of independent component vector for NB classifier (Fan et al., 2009). Some of the authors used different filter method for ranking of ICA feature vector to increase the classification accuracy of SVM and NB classifier (Rabia et al., 2015a; Rabia et al., 2015b). On the other hand Bio inspired evolutionary techniques based wrapper methods such as Ant Colony Optimization (ACO) (Tabakhi et al., 2014), Genetic Algorithm (GA) (Huang and Wang, 2006), Particle Swarm Optimization (PSO) (Lin et al., 2008), Bacterial Foraging Algorithm (BFA) and Fish School Search (FSS), are more relevant and provide a more exact solution than the other filter based wrapper techniques because they have the ability for searching and fining the optimum or near-optimum solutions on high-dimensional solution spaces. These bio inspired algorithm have been effectively applied for resolving the problem of dimension reduction in various applications such as financial domains, face recognition and text classification. While on the other hand, the result of these bio-inspired techniques depends on the complexity of the search space, fitness function, the parameters used for the algorithm, convergence, etc. All of these methods are classifier based algorithms and has obtained satisfactory performance for dimension reduction in different types of fields, they have not been frequently used for features selection of DNA microarray array data due to high computational cost. But the computational cost of the hybrid technique is lesser than only wrapper technique because of the use of the reduced number of features in its second step.

Optimization has been expanding in all directions at an astonishing rate during the last few decades. Since the ICA aims to address the issues arising from NB classification of microarray data, therefore ABC based wrapper approach with NB classifier is applied to optimize the ICA feature vectors for finding the smallest number of features that improved the classification accuracy of NB. In the research paper, we focus on the impact of the proposed algorithm ICA + ABC with NB classifier and how the performance of NB classifier improves using this combination. The proposed ICA + ABC hybrid algorithm is an iterative upgrading computational process where the population of agents, chooses a different subset of features in each iteration. After that, the performance of the different subsets of features is estimated using the classification accuracy of NB.

2. Proposed approach

2.1. Feature extraction by ICA

Indipendent Component Analysis (ICA) is a feature extraction technique, which was proposed by Hyvarinen to solve the typical problem of the non-Gaussian processes and has been applied successfully in different fields (Hyvarinen and Karhunen 2001). The extraction process of ICA is very similar to the algorithm of Principal Component Analysis (PCA). PCA map the data into an another space with the help of principal component. In place of Principal component, the ICA algorithm finds the linear representation of non-Gaussian data so that the extracted components are statistically independent. Theory of ICA algorithm can be found elsewhere (Aziz et al., 2016; Hsu et al., 2010).

2.2. Feature selection by ABC

Artificial Bee Colony (ABC) is an evolutionary feature selection algorithm that is used to select a best feature subset. The ABC algorithm reproduces the appearance of the intelligent foraging behavior of honey bee swarms for the optimization of feature subset, introduced by Dervis Karaboga in 2005 (Karaboga 2005). Garro and Beatriz A. applied ABC for selecting best genes set of microarray data with distance classifier for Artificial Neural Network (ANN) classification. ABC algorithm is a combination of local search and global search method managed by three classes of bees (employed, onlooker and scouts bees) (Garro et al., 2016). These three classes of bees with different work in the search space (colony) find the convergence of the problem near to the optimal solution.

Employed bees: These bees search new neighborhood food around their hive. After that, they compare the new food source with the old food source by using Eq. (7).

$$v_i^j = x_i^j + \varphi_i^j\left(x_i^j - x_k^j\right). \tag{7}$$

Where $v_i{}^j$ is new generated solution and $k \neq i$. φ_i^j is a random number between [−1,1]. If the fitness value $v_i{}^j$ is better than $x_i{}^j$ then changed $x_i{}^j$ with $v_i{}^j$ otherwise $x_i{}^j$ unchanged.

Onlooker bees: Employed bees share this information of solution with onlooker bees. Then by using the information of employed bee's onlooker bee find a food source with the probabilities related to its amount of nectar. This probability of finding the food source is calculated by Eq. (8)

$$p_i = \frac{fit_i}{\sum_{k=1}^{NB} fit_k} \tag{8}$$

Scout bees: If the fitness value of the i^{th} solution (fit_i) cannot be improved longer during a predefined number called "limit" then this criteria is called "abandonment criteria". For this type of criteria the scout bee creates new solutions to replace i^{th} solution by using Eq. (9).

$$x_i^j = x_{min}^j + rand(0,1)\left(x_{max}^j - x_{min}^j\right). \tag{9}$$

3. Classifier (NB)

Naïve Bayes is a simple supervised learning algorithm for machine learning classification. NB classifier used Bayes's rule with strong independence assumption for mining of different type of data (Friedman et al., 1997; Hall 2007). Due to its simplicity, Naïve Bayes is an attractive classifier amongst the researcher for solving different classification problems, including microarray (Chen et al., 2009; Sandberg et al., 2001). It is found that its performance is