Swarm and Evolutionary Computation 28 (2016) 144–160

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

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

Regular Paper

Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system

P. Mohapatra [a], S. Chakravarty [b], P.K. Dash [c,*]

[a] IIIT, Bhubaneswar, India
[b] Orissa Engineering College, Bhubaneswar, India
[c] Siksha 'O' Anusandhan University, Bhubaneswar, India

ARTICLE INFO

Article history:
Received 20 September 2015
Received in revised form
21 January 2016
Accepted 19 February 2016
Available online 2 March 2016

Keywords:
Microarray medical data
Pattern classification
Modified cat swarm optimization
RR
KRR and its variants
Support vector machine and random forest

ABSTRACT

Microarray gene expression based medical data classification has remained as one of the most challenging research areas in the field of bioinformatics, machine learning and pattern classification. This paper proposes two variations of kernel ridge regression (KRR), namely wavelet kernel ridge regression (WKRR) and radial basis kernel ridge regression (RKRR) for classification of microarray medical datasets. Microarray medical datasets contain irrelevant and redundant genes which cause high number of gene expression i.e. dimensionality and small sample sizes. To overcome the curse of dimensionality of the microarray datasets, modified cat swarm optimization (MCSO), a naturally inspired evolutionary algorithm, is used to select the most relevant features from the datasets. The adequacies of the classifiers are demonstrated by employing four from each binary and multi-class microarray medical datasets. Breast cancer, prostate cancer, colon tumor, leukemia datasets belong to the former and leukemia1, leukemia2, SRBCT, brain tumor1 to the latter. A number of useful performance evaluation measures including *accuracy*, *sensitivity*, *specificity*, *confusion matrix*, *Gmean*, *F-score* and the area under the receiver operating characteristic (ROC) curve are considered to examine the efficacy of the model. Other models like simple ridge regression (RR), online sequential ridge regression (OSRR), support vector machine radial basis function (SVMRBF), support vector machine polynomial (SVMPoly) and random forest are studied and analyzed for comparison. The experimental results demonstrate that KRR outperforms other models irrespective of the datasets and WKRR produces better results as compared to RKRR. Finally, when the results are compared on the basis of binary and multi-class datasets, it is found that binary class yields a little bit better result as compared to the multiclass irrespective of models.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Microarray analysis and classification are very much essential for early diagnosis and treatment of the most dreaded disease like cancer. It shows the highest rate of morbidity and mortality in economically developed countries and stands second in developing countries [1]. Mostly, the human beings suffer from 200 types of cancer and the microarray technology is adopted to keep records of them [2]. The GLOBOCAN database, World Health Organization, Global health observatory and United Nations World population prospectus report that the four most common cancers occurring worldwide are lung, female breast, bowel and prostate cancer [3]. It causes abnormal and uncontrolled cell growth. It is related to genome and caused by oncogenes. Molecular analysis reveals that different cancer types will have different gene expression profiles [4,5] and these may then be utilized to diagnose different cancers. High-density DNA microarray measures the activities of several thousand genes in a parallel way. This new approach helps in giving better therapeutic measurements to cancer patients by diagnosing cancer types with improved accuracy [5]. Early detection of any type of cancer increases the chance of survival for the victim. This detection is often formulated as a classification problem [6].

Microarray technology produces big datasets with gene expression values for thousands of genes (6000–60,000) in a cell mixture [7]. Hence, it becomes economically prohibitive to have a large sample size. This phenomena is called as a curse of dimensionality where samples (n) ≪ the number of features (p) [8]. To overcome this problem, microarray medical datasets need dimension reduction [8]. Dimensionality reduction methods are broadly classified into two types i.e. feature extraction [6,7] and feature selection [8–11]. During feature extraction, the features are projected into a new feature space with low dimensionality where the new features are generated as the combinations of original

* Corresponding author. Tel.: +91 674 2727336.
E-mail address: pkdash.india@gmail.com (P.K. Dash).

http://dx.doi.org/10.1016/j.swevo.2016.02.002
2210-6502/© 2016 Elsevier B.V. All rights reserved.

features. Widely-used feature extraction techniques are principal component analysis (PCA) [12–14], kernel principal component analysis (KPCA) [14], linear discriminate analysis (LDA) [12,13] and canonical correlation analysis (CCA) [15]. On the other hand, feature selection method helps in selecting a subset of highly discriminating features from the original feature set without any transformation. Hence,feature selection is superior to feature extraction in terms of better readability and interpretability [11].

Feature selection algorithms are classified into supervised, unsupervised and semi-supervised depending on the presence or absence of class [9]. Supervised feature selection method includes filter, wrapper and embedded models. Filter models do not use any classifier [9]. This technique evaluates the significance of features by looking at the intrinsic properties of the data. In this approach, all the features are scored and ranked based on certain statistical criteria. Accordingly, features with highest ranking values are selected and the low scoring features are removed. As compared to other feature selection methods, filter methods are faster but they have three major limitations: (1) they ignore the interaction with the classifier; (2) each feature is considered independently thus ignoring feature dependencies; and (3) it is very difficult to determine the threshold point for ranking the features.

The wrapper model uses a predictive accuracy of a predetermined learning algorithm to determine the quality of selected features. This method is computationally expensive to run big datasets with large number of features. The embedded model bridges the gap between these two models by taking the advantages from both the techniques [9]. Feature selection methods proposed in the literature are fast correlation based filter (FCBF) [16], relief algorithm [17], support vector machine recursive feature elimination [18], sequential forward selection (SFS) [19] and sequential backward elimination (SBE) [19]. Amongst all the methods, SFS and SBE are extensively used due to their simplicity and low computational overhead. But they also have their own limitations. The major drawback of the sequential search method is the nesting effect i.e. in backward search when a feature is deleted it cannot be reselected and in forward search when a feature is selected, it cannot be deleted [20]. That is why the stochastic search strategy is adopted where some randomness is introduced in the search process and the feature selection process becomes less sensitive to the particular dataset. The most popular stochastic methods of feature selection are genetic algorithm [21], simulated annealing [22], ant colony optimization [23], particle swarm optimization [24–26], differential evolution [27,28], bacterial foraging optimization [29], harmony search [30], cuckoo search [31], firefly [32], bat algorithm [33] and cat swarm optimization [34]. So,major advantages of feature selection method are selection without transformation, better readability and decrease in computational overhead [6].

Dimensionality reduction helps in the classification of microarray medical datasets by improving its accuracy. The important role of medical data classifier is to provide the explanation and justification for the accurate prediction of the disease [6]. Many traditional classifiers like KNN [35], naïve-bayes (NB) [36], decision tree [37], random forest [38], ID3 [39], C4.5 [40] and various neural network based classifiers like multilayer perceptron (MLP) [41], RBFNN [42], FLANN [43], SVM [44–47] are found in the literature. Amongst all the classifiers, ANN and its variant are extensively used by researchers to classify medical datasets [48]. The success of the ANN based classifier is mostly dependent on the number of hidden layers, number of nodes in each hidden layer, values of the weights between input to hidden layers, hidden to output layer and the learning algorithms. In the literature, it is generally seen that when ANN is associated with gradient descent learning algorithm the performance of the model becomes time consuming. It also increases the computational overhead [49]. Beside this, due to the initial random choice of parameters, the convergence rate of the gradient descent learning algorithm becomes very slow and most often it gets trapped in the local minima. To avoid the above said limitations, pseudo-inverse based neural network [50–55] has been proposed by many researchers like Schmidt [54], Pao [50], Broomhead and David Lowe [51]. Pseudo-inverse based neural network is recently re-named as extreme learning machine (ELM) [56] with the bias in ELM set to zero. However, this paper explores the possibility of using kernel ridge regression (KRR) [57,58] that is recently renamed as kernel ELM [59] for microarray data classification. The architecture of ridge regression has some similarity with RVFL [52] and pseudo-inverse based neural network as it uses randomly assigned input weights between the input layer and hidden layer and the weights between the output layer and hidden layer are learnt using a pseudo-inverse formulation. However, ridge regression produces a large variation in the classification accuracy in different trials with the same number of hidden nodes. But kernel function addresses this problem by replacing the hidden layer of the ridge regression. The main advantage of kernel ridge regression is that the kernel function does not need to satisfy Mercer's theorem and there is no need of any randomness in assigning the connection weights between input and hidden layers. Literature suggests that kernel ridge regression is very much similar to kernel pseudo-inverse based neural network (KPINN) [58]. It exploits the concept of quadratic program algorithms for convex optimization from mathematical programming. It also borrows the idea of kernel representations from mathematical analysis and adopts the objective of finding maximum margin classifier from machine learning theory [60].

This paper proposes a modified cat swarm optimization (MCSO)technique to select the most optimal features from microarray medical datasets and kernel ridge regression (WKRR and RKRR) to classify the features obtained from MCSO algorithm. Literature in this domain also shows that CSO performs better than PSO though its computational complexity is higher than PSO [61]. In addition to it, both PSO and DE [62] sometime get influenced by parameter convergence and stagnation problem [63] which is not there in CSO. Further, modified cat swarm optimization based feature selection method (MCSO) that is capable of improving search efficiency within the entire problem space has been used to get best optimal candidate features from the high dimensional microarray medical dataset. The proposed feature selection methods employ k-nearest neighbor algorithm as the classifier and use five-fold cross validation technique to determine the classification accuracy.

The paper is organized as follows; Sections 2 and 3 describe the process model and benchmark microarray medical datasets respectively. Section 4 deals with modified cat swarm optimization based feature selection method (MCSO). All the classifiers used in this study i.e. RR, OSRR, KRR, SVM and random forest, etc. are discussed in Section 5. Performance evaluation measures are presented in Section 6. Simulation results and analysis appear in Sections 7 and 8. Finally, conclusion is drawn in Section 9.

2. The process model for the classification of microarray datasets

All the microarray medical datasets are normalized using max–min normalization method as shown in Eq. (1). The modified cat swarm optimization algorithm (MCSO) is used to select the optimal feature subsets from these normalized datasets. For each dataset, MCSO is used to derive 10 subsets consisting of 10–100 genes in the interval of 10. To get the most optimal candidate features, k-nearest neighbor (KNN) classifier is considered to find out the classification accuracy. The subset with lower number of

Download English Version:

https://daneshyari.com/en/article/493995

Download Persian Version:

https://daneshyari.com/article/493995

Daneshyari.com