Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 38 (2012) 1644 – 1649

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Modeling, Optimization and Computing (ICMOC-2012)

A Novel PSO-FLANN Framework of Feature Selection and Classification for Microarray Data

Pournamasi Parhi[*a], Debahuti Mishra[b], Sashikala Mishra[c] and Kailash Shaw[d]

[a,b,c]*Institute of Technical Education and Research, Siksha O Anusandhan Deemed to be University, Bhubaneswar, Odisha, India*
[d]*Gandhi Engineering College, Bhubaneswar, Odisha, India*

Abstract

Feature selection is a method of finding appropriate features from a given dataset. Last few years a number of feature selection methods have been proposed for handling the curse of dimensionality with microarray data set. Proposed framework has used two feature selection methods: Principal component analysis (PCA) and Factor analysis (FA). Typically microarray data contains number of genes with huge number of conditions. In such case there is a need of good classifier to classify the data. In this paper, particle swarm optimization (PSO) is used for classification because the parameters of PSO can be optimized for a given problem. In recent years PSO has been used increasingly as a novel technique for solving complex problems. To classify the microarray dataset, the functional link artificial neural network (FLANN) used the PSO to tune the parameters of FLANN. This PSO-FLANN classifier has been used to classify three different microarray data sets to achieve the accuracy. The proposed PSO-FLANN model has also been compared with discriminant Analysis (DA). Experiments were performed on the three microarray datasets and the simulation shows that PSO-FLANN gives more than 80% accuracy.

Keywords: Particle swarm optimization; Functional link artificial neural network; Discriminant analysis; Factor analysis; Principal component analysis

1. Introduction

Microarray data sets [1] are often limited to a small number of samples with a large number of gene expressions. So the importance of the dimensionality reduction technique is highlighted due to the cost of acquirement. Due to so many features in a microarray data set it hampers the classification performance. For which, feature selection process is highly important for classification. The purpose of that it successfully achieves high classification performance on microarray data set. Feature selection is used to choose a subset [2] of input variables by eliminating features with little or no predictive information. It is also known as variable subset selection .This technique is used to select a subset of relevant features for building robust learning model by removing most irrelevant and redundant features from the data. In pattern recognition and general classification problems, methods such as PCA, factor analysis and Fisher linear discriminate analysis have been extensively used. These methods find a mapping between the original feature spaces to a lower dimensional feature space. The proposed model used both PCA and FA for feature selection. PCA is a [2-3] mathematical procedure that uses an orthogonal transformation to

* Corresponding author. Tel.: +91-9040324334; fax: +91-674-2351880.
E-mail address: pournamasi.parhi@gmail.com

1877-7058 © 2012 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2012.06.200

convert a set of observation of possibly correlated variables called principal components. The number of principal components is less than equal to the number of original variables. FA is used to reduce the number of variables [4] and to detect the structures in the relationship between variables for classification. It is a statistical method used to describe variability among observed correlated variables in terms of potentially lower number of unobserved uncorrelated variables called factors. Classification is the process to find a model which describes the distinguished data classes. Here DA and PSO-FLANN are used for classification. DA is a multivariate statistical technique [5] which is commonly used to build a predictive/descriptive model of group. Discrimination based on observed predictor variables and to classify each observation into one of the group. In DA multiple quantitative attributes are used to discriminate single classification variable. PSO is an optimization technique [6] which works by maintaining a swarm of particles that move around in the search space influenced by the improvements discovered by other particles. The rest of this paper is organized as follows: section 2 deals with related work, section 3 represents the proposed PSO-FLANN with description of PSO-FLANN algorithm, section 4 presents the comparative experiments that have been carried out to test PSO-FLANN and analyzes the experimental results and finally, section 5 gives the conclusion and future work.

2. Related Work

Arnaz Malhi *et al*. [8] used PCA as a feature selection scheme for machine defect classification. This effectiveness of the scheme was verified experimentally on a bearing test bed, using both supervised and unsupervised defect classification approach. P. Howland *et al*. [9] proposed discriminant analysis for solving the small sample size problem in face recognition. Alper Unler *et al*. [10] developed a modified discrete PSO algorithm for feature subset selection for binary classification which is competitive in terms of both classification accuracy and computation performance. Jun Sun *et al*. [11] proposed a new scheme for clustering gene expression datasets based on a modified version of Quantum-behaved PSO algorithm known as Multi-Elitist QPSO model. B. Tirimula Rao *et al*. [12] used FLANN for software cost estimation.

3. Schematic Representation of Proposed Model

Proposed model consists of feature reduction techniques for micro array data set which uses PCA [3][8] and FA [13] to reduce the dimension of data set. The reduced data set as given as the input to the two classifiers DA and PSO-FLANN then accuracy has been measured by individual classifier. Weight is updated of the classifier by using PSO. The total procedure is illustrated in fig. 1.

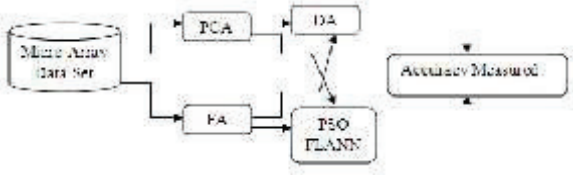

Fig. 1. The Proposed model

3.1 Working Procedure of PSO-FLANN Algorithm

FLANN is used for solving the classification problem [7]. FLANN is also used to reduce the computational complexity. FLANN architecture uses a single layer feed forward network by removing the concept of hidden layer. It is a computational model which is motivated by structural and/or functional characteristic of biological neural networks [12]. It is a single layer of artificial neural network (ANN) which is capable of forming arbitrary complex decision regions by generating nonlinear decision boundaries. The use of FLANN network is not only for functional approximation but also for decreasing

the computational complexity. The basic structure of FLANN model is shown in fig.2. The input signal *X (k)* is functionally expanded to a number of nonlinear values to feed to an adaptive linear combiner and its

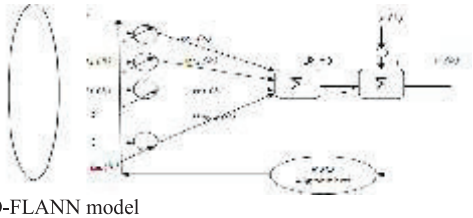

Fig.2. Schematic representation of PSO-FLANN model

weights are altered according to PSO algorithm. The obtained output is compared with desired output and the error. So the obtained output is used to alter the weights of the model. The obtained final weights represent the FLANN model for classification. The steps of PSO-FLANN are given in algorithm 1. PSO is a population based bio-inspired optimization. It is a stochastic based search algorithm commonly used to find optimum solution. PSO is an optimization technique which provides a population based search method in which individuals called changes their state with time. The velocity V_{id} and X_{id} position of the i^{th} particle are updated as follows:

$$V_{id} = V_{id} + c_1 * rand1_{id} * (pbest_{id} - X_{id}) + c_2 * rand2_{id} * (gbest_{id} - X_{id}) \quad (1)$$

$$X_{id} = X_{id} + V_{id} \quad (2)$$

Where, X_i is the position and V_i is the velocity of the particle. *pbest* is the best previous position yielding the best fitness value for the i^{th} particle and *gbest* is the best position discovered by the whole population. c_1 and c_2 are the acceleration constants reflecting the weighting of stochastic acceleration terms that pull each particle toward *pbest* and *gbest* positions respectively. $rand1_{id}$ and $rand2_{id}$ are two random numbers in range of (0,1).

Algorithm1 (PSO-FLANN)

Step 1: Read the data set *D* and class *S*, where *D* $\{D_{1, \ldots}, D_n\}$ is a vector and each vector is having m elements. *S* is an *n* * 1 matrix representing class label.

Step 2: Distribute the data set *D* into the two variables *DTR* and *DT*. *DTR* contains 80% of data for testing and 20 % of data is used for testing stored in *DT*. Similarly distribute *S* into *STR* and *ST* in same proposition.

Step 3: Normalize the *DTR*

Xc_{max}=max (*DTR*);
Xc_{min}=min (*DTR*);
for *i*=1: *ac*
for *j* =1: *bc*
$DTR(i,j)=(DTR(i,j)-Xc_{min}(1,j)) / (Xc_{max}(1,j)-Xc_{min}(1,j))$;
end
end

Step 4: Initialize the expected output

d1= [1 0 0]; *d2*= [0 1 0]; *d3*= [0 0 1];
each d_i is having 3 value representing three class label. If you have *k* class then go up to d_k.

Step 5: Initialize the parameters of PSO

c_1=0.09; c_2=0.9; *mu*=0.0001; *nump*=10;

Step 6: Randomly initialize the *wt* for the FLANN

w=2*(rand (3* *nump*, *bc* * 3)-0.5);
v=2*(rand (3* *nump*, *bc* * 3)-0.5);
Step 7: Initialize *pbest* = *w*;
Step 8: While (*maxItr)*
Step 9: for *i* =1: *n*
Step 10: *X*= [1 sin (*pi** DTR_i) cos (*pi** DTR_i)];
Step 11: *outputNeuron* = *X***w*;
Step 12: *e(i)*= d_k – *outputNeuron*; where *k*=1…number of class
Step 13: Calculate mean square error
E_k=[E_k; *e* (*i*,1).^2]
Step 14: Update velocity of particle
V_i = *c1***V*$_i$ + *c2**rand*(*gbest-pbest*)
Step 15: If rand ( ) > sigmoid (V_i)
gbest$_i$ = 1;
else
gbest$_i$ = 0;
Step 16: Update *w*
w= *w*+ μE_k * *gbest*
Step 17: *pbest=w;*
Step 18: End of for loop
Step 19: End of while loop
Step 20: Plot the error graph

3.2 Discriminant Analysis: DA [9] is one of the data mining techniques used to discriminate a single classification variable using multiple attributes. It also assigns observation to one of the predefined groups based on the knowledge of multi-attributes. In discriminant analysis the prior knowledge of the classes is required in the form of sample from each class.

4. Experimental Evaluation and Result Analysis

This section describes the experiments carried to explore the properties as well as to evaluate the performance of our proposed model. Three different data sets such as leukemia, lung cancer and breast cancer [15] have been used for experimental evaluation. The leukemia data set has contains total 72 * 7129 features, with two class level. 1to 47 belongs to claas1and 48 to 72 belongs to class 2. Lung cancer data contain 197*581 data which will have four class levels; 1 to 139 belongs to class1, 140 to 156 belongs to class 2, 157 to 177 class 3 and 178 to 197 belongs to class level 4. The breast cancer data set contains 98 * 1213 features and three class levels, where 1-11 features used for class level 1, 12-62 for class level 2 and rest belongs to class level 3. After implementing PCA the original data set has been reduced significantly as given in table 1. The input vector for our PSO-FLANN classifier is taken as the individual dataset also according to the class level. We have implemented the PSO -FLANN algorithm and feature reduction technique using both PCA and FA in MATLAB. Pseudo code for PSO-FLANN is shown in algorithm 1. We have successfully achieved nearly more than 80 % accuracy shown in table 1 and fig. 3for all the different data sets.

Table.1 Classification results obtained using PSO-FLANN

Data sets	Actual Size of data sets	Reduced size after application of PCA	Reduced size after application of FA	% of accuracy achieved using DA	% of accuracy achieved using PSO-FLANN
Leukaemia	72*7129	72*2	72*65	63 %	78 %
Lung Cancer	197*581	197*8	197*120	55 %	85.3 %

Breast cancer	98*97	98*3	98*25	60.1 %	86.3 %

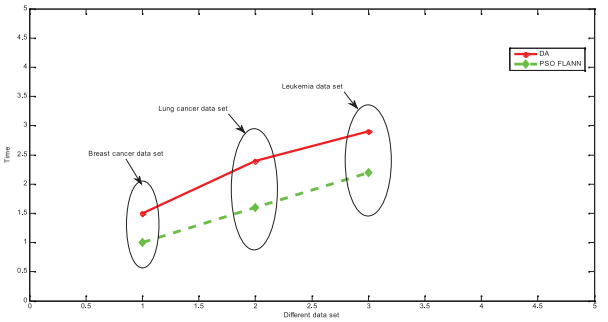

Fig.3. Performance comparison of PSO-FLANN model with DA for leukaemia, lung cancer and breast cancer data sets.

5. Conclusion and Future Work

It is a challenge to the classifiers to identify the data properly caused by high dimensional input space. Table 1 shows the competitive classification results for three benchmark dataset used in PSO-FLANN classifier. The paper also investigates the accuracy of statistics method like DA. To achieve the accuracy for multi-class is a really a challenging problem; for which, we have used the 4 class problem in our proposed classifier and also it is giving more than 80% of accuracy. In future more data set can be taken for the analysis; also the model can be used for other engineering application areas and this work can be extended to enhance the model performance using other bio inspired algorithms.

References

[1] R. Juana Canual, O.Hall Lawrence, G.Dmity, A.Eschrich Steven. Feature Selection for Microarray Data by AUC Analysis. *International Conference on Systems, Man and Cybernetics (SMC) IEEE* 2008; 768 – 773.
[2] H. Jiawei,K. Micheline. Data Mining Concepts and Techniques. *Morgan Kaufmann Publishers,* 2006.
[3] S. Fengxi, G. Zhongwei, M. Dayong. Feature Selection using Principal Component Analysis. *International Conference on System Science, Engineering Design and Manufacturing Informatization (ICSEM) IEEE* 2010; 27-30.
[4] R. Piriyakul, P. Piamsa-nga. Features Selection Analysis for Pattern Classification. *Asia-Pacific Conference on communications IEEE*. 2007; 37-40.
[5] S. Ariyo Oludare, O. Arogundade, M. Abdul-Rafiu. Discriminant and Classification Analysis of Health Status of Bell Pepper. *Research Journal of Mathematics and Statistics* 2011; 3: 2: 77-81.
[6] P. Magnus Erik Hvass. Good Parameters for Particle Swarm Optimization. *Hvass Laboratories Technical Report no. HL1001* 2010.
[7] D. Satchidananda, M. Bijan Bihari, C. Sung-Bae. Genetic Feature Selection for Optimal Functional Link Artificial Neural Network in Classification. *Springer-Verlag Berlin Heidelberg* 2008; 156-163.
[8] M. Arnaz, X. Robert Gao. PCA-Based Feature Selection Scheme for Machine Defect Classification. *IEEE Transactions on Instrumentation and Measurement.* 2004; 53: 1517–1525.
[9] P. Howland, J. Wang, H. Park. Solving the small sample size problem in face recognition using generalized discriminant analysis. *Pattern Recognition* 2006; 39: 2: 277-287.
[10] U. Alper, M. Alper. A discrete particle swarm optimization method for feature selection in binary classification problems. *Europian Journal of Operational Research* 2010; 206, 6, 528-539.
[11] S. Jun, C. Wei, F. Wei, W. Xiaojun, X. Wenbo. Gene expression data analysis with the clustering method based on an improved quantum behaved particle swarm optimization. *Engineering Applications of Artificial Intelligence* 2012; 25, 2, 376-391.

[12] B.Tirimula Rao, B.Sameet, G. Kiran swathi, K.Vikram Gupta, Ch. RaviTeja, S. Sumana. A Novel Neural Network Approach For Software Cost Estimation Using Functional Link Artificial Neural Network.*International Journal of Computer Science and Network Security* 2009; 9: 6: 126- 131.

[13] David A. Walker. A Confirmatory Factor Analysis of the Attitudes towards Research Scale. *Multiple Linear Regression View Points.* 2010; 36:1: 18-27.

[14] R. Kumar, M.S.B. Saithij, S. Vaddadi, S.V.K.K. Anoop. An intelligent functional link artificial neural network for channel equalization. *Proc. of Int. Conf. on Signal Processing Robotics and Automation* 2009; 240-245.

[15] www.uci repository.com