Journal of mathematics and computer Science 13 (2014) 353 - 358

Contents list available at JMCS

Journal of Mathematics and Computer Science

Journal Homepage: www.tjmcs.com

Proposing a Neural-Genetic System for Optimized Feature Selection Applied in Medical Datasets

Saeed Ayat[1], Mohammad Reza Mohammadi Khoroushani[2,*]

[1]Associate Professor, Department of Computer Engineering and Information Technology, Payame Noor University, Iran
[2]M.Sc. student, Department of Computer Engineering and Information Technology, Payame Noor University, Esfahan, Iran

[1]dr.ayat@pnu.ac.ir, [2]mr_mohammadi@of.iut.ac.ir

Article history:
Received July 2014
Accepted October 2014
Available online October 2014

Abstract

This paper, presents a new system for selecting the best optimized features among a collection of features by combination of neural network and genetic algorithm. Feature selection is an important issue because it has a direct impact on the performance (Specificity, sensitivity) and system efficiency.

The proposed system uses neural network for selecting the best features based on Signal to Noise Ratio (SNR), and genetic algorithm for training the neural network by determining the optimum values of weighs and other parameters. This system is a combination of a Multi-Layer Perceptron (MLP) with 3 layers and decimal genetic algorithm

We evaluated our proposed system on 10 medical data sets and compared it with binary genetic algorithm that is used widely for feature selection. The results confirmed the superiority of the proposed system in Specificity, sensitivity and the number of selected optimized features.

Keywords: feature selection, optimized feature selection, neural network, genetic algorithm.

1. Introduction

Selecting the best optimized features is set among the NP-Hard issues which is it is concerned with optimization. The different methods regarding feature selection issues are of the filter and wrapper categories [1] which have under gone evaluations. Since the wrapper methods, are able to correspond to the learning machine algorithm, they provide better results in comparison with that of the filter methods-In these methods, the growing number of the features of training samples is needed to grow exponentially. In this case, the computational cost increases. So it is very important to find the appropriate method. Since the problem of optimized feature selection is part of optimization problems, evolutionary algorithms like genetic algorithm has mostly been studied and used [2, 3]. In some studies, a combined method is used. They aim to take advantage of algorithms for achieving an efficient method respectively [4, 5, 6]. A combined method of neural and genetic algorithm is recommended regarding the feature selection by [7] and [8]. In this study a combined neural-genetic approach is being introduced followed by evaluation of the performance and efficiency of these approaches by comparing them to the binary genetic algorithm regarding effective features selection. In a neural network learning is defined as determining the optimum values of weighs and other parameters like bias and the motion derivative gradient for more efficiency. In network learning the objective is to increase the network efficiency by reducing the errors between real output and network output for the training data and being tested against proper changes in weighs and other parameters of the network. The common approach in this method is the back propagation of the errors. Knowing that learning is a process regarding optimization, it is assumed that the evolutionary algorithms like genetic, Imperialist Competitive Algorithm (ICA), Particle Swarm Optimization (PSO), etc. might be able to increase the efficiency and performance in the neural networks regarding feature selection in its most suitable sense.

In this proposed combined approach the neural network is applied as the most suitable feature selector and the genetic algorithm is applied as a training algorithm. Here, the neural network advantages on the genetic algorithm are combined and the proper balance between efficiency (process speed) and performance (accuracy in pattern Recognition) is found and compared with the studies conducted by [7, 8]. The proposed combined approach is compared to the binary evolutionary genetic algorithm [9].

2. Proposed feature selection system

In this study the signal to Noise ratio, (SNR) criterion, [10] is applied. This criterion makes a feature's weight outstanding in a sense that the weights connected to features of low importance have small values or approaching zero, while the opposite is true which leads to (absolute big value). This approach is of the following two sections. The first section is responsible to select the candid chromosomes (the optimum value of weighs and the biases) to be used in section two. The Fitness function apply the recommended weight to the network, next the determined data set is applied to the network and the mean square of the errors are computed and returned to the genetic algorithm as the value of the Fitness. The error value here, is the result of the dataset outcome and the result if the network outcome, calculated through the mean square error (MSE) equation; hence, the final fitness value which is returned to the system.

The error is the result of the dataset outcome and the network outcome that calculated by the mean square error (MSE) equation. This error returned to the network as fitness feedback.

The second section is responsible to prune the architecture of the candidate chromosome from section one based on (SNR) and categorize the related training samples.

The steps of this approach consist of:

1. The SNR is computed by using the weights that are related to the obtained chromosomes from section1. Then the weights smaller than one will replace by zero. That means the features related to zero weights will remove from future computations.
2. The value of fitness function in this section is computed as the opposite errors categorization
3. The candidate chromosome that meets the fitness function requirements is selected as the most suitable chromosome with its related features

Table 1. Descriptions of neural networks applied in section 1 and 2

	Descriptions of neural networks applied in section 1	Descriptions of neural networks applied in section 2
Number of neurons at the output layer	Number of each dataset categories	Number of each dataset categories
Number of neurons at the input layer	To the number of features	To the number of non-zero chromosome genes
Intermediate layer	1	1
Number of neurons at the intermediate layer	10	10
First layer function	sigmoid	sigmoid
Second layer function	linear	linear
training rate	0.2	0.2
Repeating	500	500
The squares errors average	0.02	0.02

The number of intermediate layer neurons at both the sections is selected among 3, 5, 7, 10, 15 and 25 indicating that the best average efficiency is 10.

The decimal generic algorithm description regarding neural network training is presented in Table 2. These values are selected on experimented basis.

Table 2. The decimal generic algorithm description

Population	Next generation members	Generation	Cross Over rate	The mutation rate	Integration method	Selection function	The decimal chromosome bounds	Selection function
50	2	100	0.08	0.2	Tow point	Rank	[-10,10]	Stochastic uniform

The applied definitions and the implemented steps of this logarithm consist of:

a) Coding: beginning the algorithm with a random population of *n* chromosomes each with a length of 430 features
b) Evaluation: computation of the suitability function according to Table 1
c) Selection: here, two feature vector of (chromosome) are selected through the roulette function and the Fitness average are applied
d) Integration: the selected responses are subject to change. In this approach single point integration is applied where one point is selected along the length of the strip on random basis and the genes (features) are displaced after this point. The Cross Over rate in this study is 0.9.
e) Genetic mutation: this mutation is of 0.005 probability
f) Placing the newborn chromosomes in a collective as the new generation
g) Adding new generation members to the selected members of the initial population of 200 with the new generation

h) Repetition of the steps beginning from step b

The best response is selected after 100 executions.

3. Evaluation

The database applied here is from [11] which include 10 data profiles on types of cancer and tumors like brain, kidney, prostate etc. Here 80% of the samples of each dataset are used in training and 20% are used for testing.

Table 3. The medical dataset

Dataset title	Description	Sample number	Feature number	Category number
9 Tumors GEMS	Nine various human Tumor types	60	5726	9
11 Tumors GEMS	Eleven various human Tumor types	174	12533	11
14 Tumors GEMS	Fourteen various human Tumor types and 12 normal tissue types	308	15009	26
Brain Tumor1 GEMS	Five human brain tumor types	90	5920	5
Brain Tumor2 GEMS	Four malignant glioma types	50	10367	4
Leukemia1 GEMS	Acute myelogenous leukemia (AML), acute lymphoblastic leukemia (ALL) B-cell, and ALL T-cell	72	5327	3
Leukemia2 GEMS	AML, ALL, and mixed-lineage leukemia (MLL)	72	11225	3
Lung Cancer GEMS	Four lung cancer types and normal tissues	203	12600	5
SRBCT GEMS	Small, round blue cell tumors of children	83	2308	4
Prostate Tumor GEMS	Prostate tumor and normal tissue	102	10905	2

3.1. Evaluating the effectiveness of this proposed approach on feature dimension reduction

The dimension reduction made through different approaches is presented in Table 4.

Table 4. Comparison of dimension reduction on medical dataset s through different approaches

Dataset title	Number of features	Number of the most desirable features obtained through the neural network using genetic training	Number of the most desirable features obtained through binary Genetic algorithm
9 Tumors GEMS	5726	2582	2620
11 Tumors GEMS	12533	4910	5345
14 Tumors GEMS	15009	4845	4232
Brain Tumor1 GEMS	5920	635	1034
Brain Tumor2 GEMS	10367	995	1038
Leukemia1 GEMS	5327	1900	1872
Leukemia2 GEMS	11225	633	969
Lung Cancer GEMS	12600	3199	4211
SRBCT GEMS	2308	264	313
Prostate Tumor GEMS	10905	585	594

3.2. Correct classification rate

The correct classification rate based on two sensitivity and specificities, applied in medical evaluations, are tabulated in Table 5 [12].

Table 5. Correct classification rate of the dataset based on the two sensitivity and specificities

Approach	Neural network through genetic training algorithm		Binary Genetic algorithm	
	Sensitivity	Specificities	Sensitivity	Specificities
9 Tumors GEMS	61	66	56	58
11 Tumors GEMS	62	73	62	61
14 Tumors GEMS	54	61	50	52
Brain Tumor1 GEMS	66	74	68	73
Brain Tumor2 GEMS	66	76	67	69
Leukemia1 GEMS	73	66	62	63
Leukemia2 GEMS	72	76	69	72
Lung Cancer GEMS	84	91	85	87
SRBCT GEMS	58	61	53	55
Prostate Tumor GEMS	62	64	61	64

The numbers in this table are rounded and presented as percentages. As the performance (Sensitivity, Specificities) evaluation criterion, result of accuracy multiplied by the detecting sensitivity is computed in accordance with [7], since when both these values indicate a rise it means the whole diagnostic accuracy rate has increased. This phenomenon is presented through the following Eqn. :

$$\boldsymbol{Performance\ Evaluation} = \sum^{\boldsymbol{i} \in \{\boldsymbol{Datasets}\}} \boldsymbol{Sen}(\boldsymbol{i}) \times \boldsymbol{Spec}(\boldsymbol{i}) \qquad (1)$$

Where, *i* is one of the 10 dataset s and sen(i) and spec(i) are the diagnostic sensitivity and specificities of the i^{th} dataset.

4. Conclusion

Using neural network in selecting the most desirable features optimized with respect to combined training process, the feature extraction, feature selection and rational classification are of major concern. In this article the signal to noise criterion ratio technique is introduced in selecting the most desirable process through pruning architecture. Training is a process towards optimization and here the meta-heuristic genetic algorithm is applied to select weights with proper and optimized biases in neural network learning algorithm. The results here indicate the establishment of appropriate balance between efficiency and performance. The proposed system here is compared with Genetic Algorithm introduced in other studies like [9] that claimed to be an optimized approach, the evaluation results confirmed the superiority of the proposed system.

Acknowledgement

Appreciations are extended to the following esteemed individuals:

- Assistant Prof., Mohammad Reza AkhawanSaraf, FAVA Research Center, Informatics Dept. head and faculty member at Isfahan Industrial University
- MD, SiminHemati, Radiation Treatment Dept. head at Isfahan Medical Sciences University
- MD, Mina Tajvidi, specialist in Dermatology and Radiotherapy Enchology at Seid al Shohada Hospital
- MD, Rezaii specialist in Dermatology and RadioteraphicEnchology at Seid al Shohada Hospital

References

[1]N. Sanchez-Marono, A. Alonso-Betanzos, M. Tombilla-Sanroman, "Filter methods for feature selection: a comparative study", Proceedings of the 8th international conference on Intelligent data engineering and automated learning, Birmingham, UK, Springer-Verlag: 178-187, 2007.

[2] M. Saberi, D. Safaai, "Feature Selection Method Using Genetic Algorithm For The Classification Of Small and High Dimension Data" in International Journal on IEEE Transaction On Pattern Analysis And Machine Intelligence, VOL. 23, NO. 11, 2005.

[3] M. T. Miller, A. K. Jerebko, J. D. Malley, R. M. Summers, "Feature Selection for Computer-Aided Polyp Detection using Genetic Algorithms", Proceedings of SPIE, Vol. 5031 ,2003.

[4] H. Chouaib, O.R. Terrades, S. Tabbone, F. Cloppet, N. Vincent, "Feature selection combining genetic algorithm and Adaboost classifiers" in 19th International Conference on Pattern Recognition(IEEE), ICPR 2008, Tampa, FL,2008.

[5] E. P. Ephzibah, "Cost Effective Approach on Feature selection Using Genetic Algorithms And Fuzzy Logic for Diabetes Diagnosis" , in International Journal on Soft Computing(IJSC), Value 2,No1 , February 2011.

[6] I. S. Oh, J. S. Lee, B. R. Moon, "Hybrid Genetic Algorithms for Feature Selection", IEEE Transaction On Pattern Analysis And Machine Intelligence, VOL. 26, NO. 11, November 2004.

[7]M. Nirooee1, P. Abdolmaleki2, M. Gity, "Designing a Hybrid Model to Differentiate between Malignant and Benign Patterns in Breast Cancer from Mammographic Findings (Text in Persian)", in Iranian Journal of Medical Physics, Pages 67-80, May 2008.

[8] S.Zanganeh, R.Javanmard, M. Ebadzadeh, "A Hybrid Approach for Features Dimension Reduction of Datasets using Hybrid Algorithm Artificial Neural Network and Genetic Algorithm-in Medical Diagnosis" in 4rd Iran Data Mining Conference (IDMC), 2010.

[9] L. Ballerini, X. Li, R. B. Fisher, and J. Rees, "A query-by-example content-based image retrieval system of non-melanoma skin lesions" presented at the Proceedings of the First MICCAI international conference on Medical Content-Based Retrieval for Clinical Decision Support, London, UK, 2010.

[10] A. Verikas. M. Bacauskiene, "Feature selection with neural networks", in Journal Pattern Recognition Letters, Value 23 Issue 11, September 2002 Pages 1323-1335.

[11] gems. (2014, 2014/22/07).Gene Expression Model Selector. Available: http://www.gems-system.org.

[12] M. R. MohammadiKhoroushani, S. Mahzounieh, "Intelligent Skin Cancer Detection Software System based on the principles of telemedicine", in Journal of Hospital, Value 5, Pages 55-60, April 2014.