Journal of Digital Contents Society Vol. 17 No. 6 Dec. 2016(pp. 509-521)
https://doi.org/10.9728/dcs.2016.17.6.509

Bio-marker Detector and Parkinson's disease diagnosis Approach based on Samples Balanced Genetic Algorithm and Extreme Learning Machine

Vasily Sachnev*, Sundaram Suresh, YongSoo Choi*****

Abstract

A novel Samples Balanced Genetic Algorithm combined with Extreme Learning Machine (SBGA-ELM) for Parkinson's Disease diagnosis and detecting bio-markers is presented in this paper. Proposed approach uses genes' expression data of 22,283 genes from open source ParkDB data base for accurate PD diagnosis and detecting bio-markers. Proposed SBGA-ELM includes two major steps: feature (genes) selection and classification. Feature selection procedure is based on proposed Samples Balanced Genetic Algorithm designed specifically for genes expression data from ParkDB. Proposed SBGA searches a robust subset of genes among 22,283 genes available in ParkDB for further analysis. In the "classification" step chosen set of genes is used to train an Extreme Learning Machine (ELM) classifier for an accurate PD diagnosis. Discovered robust subset of genes creates ELM classifier with stable generalization performance for PD diagnosis. In this research the robust subset of genes is also used to discover 24 bio-markers probably responsible for Parkinson's Disease. Discovered robust subset of genes was verified by using existing PD diagnosis approaches such as SVM and PBL-McRBFN. Both tested methods caused maximum generalization performance.

Keywords : Bio-marker, Parkingson Disease, Genetic Algorithm, Machine Learning

균형 표본 유전 알고리즘과 극한 기계학습에 기반한 바이오표지자 검출기와 파킨슨 병 진단 접근법

Vasily Sachnev*, Sundaram Suresh, 최용수*****

요 약

본 논문에서는 파킨슨 병 진단 및 바이오 표지자 검출을 위한 극한 기계학습을 결합하는 새로운 균형 표본 유전 알고리즘(SBGA-ELM)을 제안하였다. 접근법은 정확한 파킨슨 병 진단 및 바이오 표지자 검출을 위해 공개 파킨슨 병 데이터베이스로부터 22,283개의 유전자의 발현 데이터를 사용하며 다음의 두 가지 주요 단계를 포함하였다 : 1. 특징(유전자) 선택과 2. 분류단계이다. 특징 선택 단계에서는 제안된 균형 표본 유전 알고리즘에 기반하고 파킨스병 데이터베이스(ParkDB)의 유전자 발현 데이터를 위해 고안되었다. 제안된 제안 된 SBGA는 추가적 분석을 위해 ParkDB에서 활용 가능한 22,283개의 유전자 중에서 강인한 서브셋을 찾는다. 특징분류 단계에서는 정확한 파킨슨 병 진단을 위해 선택된 유전자 세트가 극한 기계학습의 훈련에 사용된다. 발견 된 강인한 유전자 서브세트는 안정된 일반화 성능으로 파킨슨 병 진단을 할 수 있는 ELM 분류기를 생성하게 된다. 제안된 연구에서 강인한 유전자 서브셋은 파킨슨병을 관장할 것으로 예측되는 24개의 바이오 표지자를 발견하는 데도 사용된다. 논문을 통해 발견된 강인 유전자 하위 집합은 SVM이나 PBL-McRBFN과 같은 기존의 파킨슨 병 진단 방법들을 통해 검증되었다. 실시된 두 가지 방법(SVM과 PBL-McRBFN)에 대해 모두 최대 일반화 성능을 나타내었다.

키워드 : 바이오표지자, 파킨스 병, 유전 알고리즘, 기계 학습

※ Corresponding Author : Yong Soo Choi

Received : December 10, 2016
Revised : December 27, 2016

1. Introduction

Parkinson's disease (PD) is a very dangerous disease spread worldwide. Parkinson's disease may cause by genetic, environmental conditions, or both. PD diagnosis is hard and mostly inefficient in early stages. Recent PD diagnosis tools include: NMS - PD Non-motor symptoms questionnaire, ADL - Schwab and England Activities on Daily Living, UPDRS - unified Parkinson's disease Rating Scale.

Few automatic approaches based on machine learning techniques were recently developed for PD diagnosis. Mostly those techniques analyze special vocal and gait features.

Little et. al. [1] collected a vocal data from 23 PD patients and 31 normal persons. Authors utilized a kernel support vector machine (SVM) for PD recognition. Caglar et al. [2] implemented a concept of linguistic hedges for feature selection and fuzzy interface to build a PD classifier. Sakar and Kursun [3] tried to select features using maximum relevance minimum redundancy (mRMR) criteria and optimize SVM parameters for better PD diagnosis. Das [4] examined several machine learning tools for PD diagnosis. Authors reported that neural network tuned by Levenberg - Marquardt algorithm is the best choice for PD diagnosis. Sateesh Badu et. al [5] build a PD classifier based on meta-cognitive radial basis function network.

Motion or gait analysis may be efficient in detecting Parkinson's disease. Engin et al, [6, 18] used several neural networks based on gait patterns analysis for PD diagnosis. Pan et al. [7] used tremor data based on intra-operative microelectrode recording of Local Field Potential (LFP) signals to build PD. Tahir and Manap [8] build a set of gait features extracted from 12 PD patients and 20 normal persons.

However, traditional PD diagnosis methods based on tremor and voice features analysis are effective only if approximately 70% of vulnerable dopaminergic neurons died due to PD [9]. Hence, another diagnosis approach with better chance to detect PD is needed.

MRI analysis is another way to build a PD diagnosis approach. Sateesh Badu et al [12] used MRI scans and build projection based learning for meta-cognitive radial basis function network coupled with recursive feature elimination approach (PBL-McRBFN-RFE) to build PD classifier and define brain areas responsible for PD.

Gene analysis can be also implemented for Parkinson's Disease diagnosis. Scherzer et al. [9] discovered a difference in between genes expression information for PD patients and normal persons. Taccioli et. al. [10] collected a gene expression information data base (ParkDB) for Parkinson's disease research. Sateesh Badu et al. [11] proposed an efficient PD classifier by using genes' expression data taken from ParkDB data base. Authors used projection based learning for meta-cognitive radial basis function network (PBL-McRBFN) to build a PD classifier. PBL-McRBFN follows general human meta-cognitive learning principals.

Gene expression data is mostly redundant. Thus, direct use of the complete set of genes' data given in ParkDB to build a PD classifier does not guarantee high classification performance. Hence, search for an optimal subset of genes with ability to build PD classifier with better generalization performance is

Accepted : Decembe 30, 2016
** School of Information, Communication and Electronics Engineering, Catholic University
email: bassvasys@hotmail.com
*** School of Computer Science and Engineering, Nanyang Technological University, Singapore
email: ssundaram@ntu.edu.sg
***** Division of Liberal Arts & Teaching, Sungkyul University
Tel: +82-31-467-8374
email: ciechoi@sungkyul.ac.kr

needed.

Search for an optimal subset of genes is widely used by researchers in many areas of science. For example, Saraswathi et. al [13] proposed an Integer Coded Genetic Algorithm and Particle Swarm Optimization coupled with Extreme Learning Machine (ICGA-PSO-ELM) to build an efficient cancer diagnosis technique. Saraswathi method searches an optimal subset of genes and uses it to build a classifier to detect 14 types of cancer. However, ICGA-PSO-ELM has a significant drawback. The number of genes in subset should be assigned manually. Later Sachnev et. al. [15] used Binary Coded Genetic Algorithm for searching an optimal set of genes from GCM data base. Authors reported about 52 discovered bio-markers from the set of 92 chosen genes, which were used to build a cancer classifier.

In this paper a novel PD diagnosis approach based on proposed SBGA-ELM is presented. Proposed PD diagnosis scheme contains two major steps: 1) "feature selection" based on proposed SamplesBased Genetic Algorithm and 2) "classification" based on Extreme Learning Machine. Proposed Samples Balanced Genetic Algorithm is a completely automatic approach for searching an optimal subset of genes from ParkDB data base. Chosen subset of genes is, then, used to build an ELM classifier. Proposed SBGA-ELM produces a large number of genes' subsets which must be evaluated in a "classification" step. Thus, machine learning technique for a "classification"step should have a fast learning procedure and acceptable classification performances. Extreme Learning Machine is a best choice for a proposed PD diagnosis approach. ELM is extremely fast and accurate. ELM[14] is a feed-forward single hidden layer neural network, where the weights of the input neurons and bias of the hidden neurons are chosen randomly, and the output weights are calculated analytically.

After extensive experiments with ParkD data base using SBGA-ELM we found a large number of genes subsets which create ELM classifiers with absolute maximum classification training and testing accuracies 100%. It causes extra difficulties for searching an optimal subset of genes. Searching is possible only if each examined subset of genes is possible to rank using corresponding ELM classifier. If ranks are same and equal 100% search is not possible anymore. Hence, new approach with ability to efficiently evaluate subset of genes which create ELM classifiers with maximum classification accuracy is needed.

The paper is organized as follows: In Section II ParkDB gene expression data base for PD research is presented. The framework and detail explanation of the proposed SBGA-ELM is presented in Section III. Section IV displays experiments in details. Section V concludes a paper.

2. ParkDB

A publicly available data base (ParkDB) presented by Taccioli et. al. [10] is the first well-organized data base with gene expression information for Parkinson's Disease research. ParkDB collects gene expression information of 22283 genes extracted from RNA of blood samples taken from 22 normal patients and 50 PD patients at early stages.

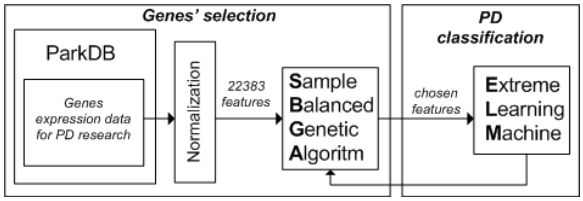
ParkDB data base is created using Robust Multi-array Analysis method proposed by Smyth [15] one of the most famous microarray technique. Extracted genes' expression information is then normalized. A set of normalized genes expression information of 22283 genes of 72 patients finally builds a publicly available ParkDB data base.

Proposed PD diagnosis approach uses nor-

malized micro-array genes expression data from ParkDB with the accession number E-GEOD-6613 to build an efficient PD classifier.

3. Proposed SBGA-ELM approach for PD diagnosis.

Presented PD diagnosis approach based on proposed Samples Based Genetic Algorithm coupled with Extreme Learning Machine. Proposed SBGA-ELM is used to build an efficient PD classifier which efficiently classifies PD patients in early stages. Proposed method contains 2 major steps: 1) "Genes' selection" and 2) "PD classification" (See Figure 1).



(Figure 1) Framework of the proposed PD diagnosis based on SBGA-ELM

In the first "Genes' selection" step a proposed PD diagnosis approach normalizes genes'expression information from ParkDB for further classifying using machine learning technique (ELM). Given genes' expression information from ParkDB needs to be normalized to a range of [0; 1] for a correct use in ELM. Normalized genes expression information is then processed by proposed Samples Balanced Genetic Algorithm. SBGA searches the best subset of genes/features among 22283 genes available in ParkDB. A set of features/genes chosen by SBGA is, then, used to build an Extreme Learning Machine (ELM) classifier for accurate PD diagnosis.

Proposed SBGA-ELM unifies searching procedure based on Samples Balanced Genetic Algorithm and PD classification based on Extreme Learning Machine (ELM) into one framework. Efficiency of the PD classification mostly depends on chosen genes/features. Thus, the set of features which creates ELM classifier with high accuracy for PD classification can be slightly modified to create other sets of genes/features with promising performances. Modified set of genes may create ELM classifier with even better classification accuracy. At the same time sets of genes/features which create ELM classifiers with low accuracies for PD classification will not be used to create other sets of genes/features. Such evaluation based approach is a key component of the famous Genetic Algorithm (GA).

The detail explanations of the proposed Samples Based Genetic Algorithm coupled with Extreme Learning Machine are presented below.

1.1 Samples Balanced Genetic Algorithm (SBGA)

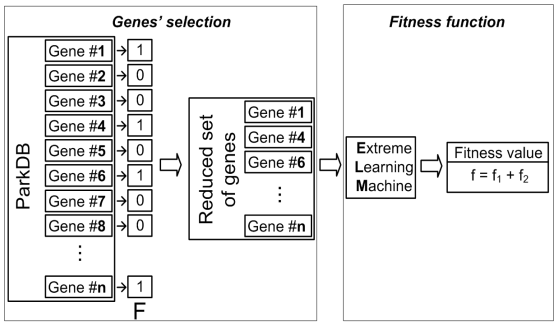
Proposed Samples Balanced Genetic Algorithm is a modified version of the Genetic Algorithm designed to search for an optimal set of genes/features for PD classification problem.

GA is a self-adaptive evaluation based approach for solving various optimization problems. GA mimics a chromosome exchange mechanism existed in nature. In GA framework each optimization problems transforms to a set of meaningful chromosomes. Manipulations with chromosomes helps GA to find solution closed to optimal solution. However, due to random manipulation with chromosomes similar to natural processes, GA mostly fails to find a global optimum. Thus, GA is a heuristic iterative optimization tool for fast searching of suboptimal solutions for complex optimization problems with large number of variables.

Proposed Samples Balanced Genetic Algorithm contains following functional units: Binary string representation, Genetic operators: Samples Balanced Crossover and Mutation, Adapted fitness function, Selection procedure, and Termination criteria.

Binary string representation: As was described before any optimization problem in GA should be first transformed to a set of relevant chromosomes. The set of chromosomes builds a single solution for a given problem. Each chromosome represents a meaningful unit, mostly number, which has an important impact to optimization problem. In GA each solution, or set of chromosomes, is a set of numbers. Each solution is then evaluated by fitness function. Fitness function processes examined solution by computing a fitness value, which numerically evaluates an examined solution.

Proposed PD diagnosis approach uses set of genes'expression information from ParkDB to build a PD classifier. Each gene represents a number in ParkDB. Proposed Samples Balanced Genetic Algorithm has to pick few genes from ParkDB to build a reduced set of genes for further analysis using ELM classifier. Each genes can be either picked or not to build a reduced set of genes. Thus, each solution for SBGA should represent a status of each gene from ParkDB: pick/NOT-pick, "true"/"false", or binary "1"/"0" (see Fig. 2). If binary coefficient is "1" then corresponding gene from ParkDB is selected to build a reduced set of genes, if "0" corresponding gene from ParkDB is skipped. Each solution for PD classification problem in the SBGA-ELM framework is a set of 22283 binary coefficients which represents a status of each gene from ParkDB in the reduced set of genes (see Fig. 2).



(Figure 2) Binary string representation for SBGA-ELM

Genetic operators: Samples Balanced Crossover and Mutation. Genetic Algorithm uses genetic operators (crossover and mutation) to update solutions in each population (see Fig. 3). Crossover and mutation modify solutions of the given optimization problem in a way similar to natural process of exchanging chromosomes. In nature crossover builds a new genome by permuting chromosomes from two sources randomly. After crossover operation new genome contains genetic materials from both sources. As a result, new genome may have properties from both sources. In nature mutation modifies chromosomes randomly, such that mutated chromosomes produce new properties, which did not exist in both sources. Mutation, in general, may not affect new genome significantly, but sometime may cause very unique properties. Combination of crossover and mutation build a basis of genetics in nature.

Genetic Algorithm mimics key processes from crossover and mutation existed in nature. GA crossover generates new solution by exchanging chromosomes from 2 randomly chosen solutions using fixed exchange strategy. Popular GA crossovers are single point crossover, 2 point crossover, uniform crossover, arithmetic crossover. Popular GA mutation is bit inversion mutation.

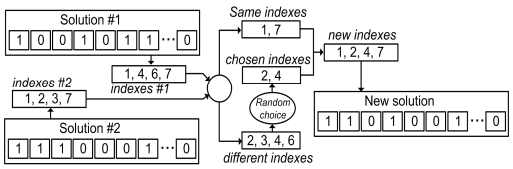
In general, proper choice of the genetic op-

erators (crossover and mutation)defines efficiency of the Genetic Algorithm. Different optimization problems may need different crossover and mutation, or combination of few crossovers and mutations for more efficient search. Proper choice of the genetic operators is problem specific and depends on given data. Optimization problem can be solved by using GA as follows: Assume that there are 4 well-known crossovers in the list - single point crossover, 2 point crossover, uniform crossover, arithmetic crossover. Every time when crossover is needed, one crossover among given 4 is picked randomly. Thus, GA always generates new solutions using all listed crossovers. Such strategy is called hybrid crossover. GA based on hybrid crossover may solve various optimization problems without choosing a proper crossover.

Some optimization problem may not be solved efficiently by using GA with hybrid crossover. It may happen because necessary crossover was not in the list. Necessary crossover may not be listed or it may not exist. Search for an optimal set of genes using GA based on hybrid crossover for PD classification problem is failed. GA failed because all examined crossovers are not suitable for a stable search using GA. Hence, new crossover designed specifically for PD classification problem is needed.

The problem with crossover issue for GA in PD classification problem has been examined in details. Each binary solution for PD classification problem contains relatively small number of binary coefficients "1". It implies that each binary solution links to 20 100 genes among 22283 genes in ParkDB data base. Note that, only coefficients "1" /"true" are important in the proposed scheme. However, all examined crossovers create new solutions with almost uncontrolled number of "1". The total number of "1"in new solutions significantly changes compared to initial solutions, which causes significant performance loss for PD classification. Sometimes crossovers generate solutions with all zeros, which is unacceptable(means there is no chosen features to build classifier). Hence, new crossover designed specifically for PD classification problem is needed. In this research a new samples balanced crossover to deal with genetic data from ParkDB is proposed.

Samples Balanced Crossover is designed to control the number of binary coefficient "1" for presented PD classification problem. Samples Balanced Crossover is displayed in Fig 3.



(Figure 3) Samples Balanced Crossover.

Samples Balanced Crossover extracts indexes of all binary coefficients "1" from input "solution #1"and "solution #2". Then indexes are divided into "same indexes" and "different indexes". "Same indexes" are then moved to new solutions solution directly. Few indexes from "different indexes" are selected randomly to build "chosen indexes". The set of "new indexes" is created by unifying "same indexes" and "chosen indexes". Finally new solution is created using indexes from "new indexes". Split parameter s in "Random choice" is limited in the range of 0.25 - 0.75.

In the example presented in Fig. 3 "solution #1" and "solution #2" have coefficients "1" in the positions {1, 4, 6, 7} and {1, 2, 3, 7} respectively. Set of "same indexes" is {1, 7}, set of "different indexes" is {2, 3, 4, 6}. The size of the set "different indexes" is 4. Split parameter s for "Random choice" is 0.5. Thus, "Random choice" chooses $4 \cdot s = 4 \cdot 0.5 = 2$ index from "Different indexes" randomly and

builds set of "chosen indexes" {2, 4}. Finally, set of "new indexes" {1, 2, 4, 7} is a combination of "same indexes" {1, 7} and "chosen indexes" {2, 4}.

Proposed Samples Balanced Crossover permutes any binary solutions and keeps a balance of the binary coefficient "1" in between initial and output solutions. In the scenario when "solution #1" and "solution #2" have just few common indexes, indexes from both solutions mostly belong to "different indexes" set. Then new solution is created mostly from "different indexes" set. In this situation new solution is significantly modified compared to initial solutions. In the scenario when most of the indexes belong to "same indexes" set, modifications in between new solution and initial solutions are minor. Thus, proposed Samples Balanced Crossover Such helps to keep convergence of Genetic Algorithm for presented optimization problems. During the first iterations, when most of the indexes belong to "different indexes" set, new generated solutions are mostly random, which is similar to chaotic search. At the end, when most of the features belong to "same indexes" set, new generated solutions are not very different compared to initial solutions, which is similar to searching an optimal solution.

Proposed Samples Balanced Mutation replaces initial solution to randomly generated new solution. The number of binary coefficients "1" in a new solution is skipped same compared to number of binary coefficients "1" in initial solution. Thus, Proposed Samples Balanced Mutation keeps the balance of the coefficients "1" in between initial and new solutions.

Fitness function evaluates each solution in GA by computing a numerical measure for further analysis. Fitness value or numerical measure defines importance of different solutions for solving given optimization problem. GA mostly searches solution with maximum or minimum fitness value. Solutions with maximum or minimum fitness value may be an optimal solution for a given optimization problem. In GA framework solutions are sorted based on fitness values. Solutions with better fitness values (high or low) are used again by crossover and mutation to build new slightly different solutions, which may have even better fitness values. Other solutions are usually ignored.

In the proposed PD diagnosis approach Extreme Learning Machine is used as a fitness function to evaluate each solution in GA. Each solution refers to the set of genes from ParkDB data base. Chosen set of genes is then used to train 20 ELM classifiers based on randomly generated weights of the input neurons and random bias of the hidden neurons (refer to subsection "Extreme Learning Machine"). Then fitness value f is calculated as follows:

$$f = \sum_{i=1}^{20} (f_1 + f_2)_i \qquad (1)$$

$$f_1 = \sum_{j=1}^{N} J_j \qquad (2)$$

where

$$J_t = \begin{cases} 1, \text{if } y^t = \hat{y}^t \\ 0, otherwise \end{cases} \quad t = 1,2,...,N$$

where f_1 is a number of correctly classified samples, $y^t$is coded class label (see Equation 5), $\hat{y}^t$ is a predicted class label, N is a number of testing samples, i is the index of the ELM classifier. If coded class label and predicted class label for t-th sample are the same, than equal 1, otherwise 0.

$$f_2 = \frac{N - \sum_{k=1}^{N} (y^k - \hat{y}^k)}{N} \qquad (3)$$

where f_2 is a penalty factor, which measures difference between predicted and coded class labels. $f_2 \in [0;1]$. Penalty factor f_2 separates cases, when the number of correctly classified samples is the same, i.e., $f_1^n = f_1^m$,

then $f^n = f_1^n + f_2^n, f^m = f_1^m + f_2^m$, finally $f^n > f^m$, if $f_2^n > f_2^m$, and vice versa.

Proposed PD classifier based on SBGA-ELM is able to find solution, which results 100 % of testing efficiency, i.e., $f_1 = N$. Note that, the proposed SBGA-ELM is also searching for a robust set of genes responsible for PD. If fitness value reaches a maximum possible value ($f_1 = N$ or 100% of overall testing accuracy), further search using GA is not possible anymore. Penalty factor $f_2$resolves this problem. Proposed SBGA-ELM with fitness value calculated using Equation 1 keeps searching for a robust set of genes, even if PD classifier produces overall testing accuracy 100%.

Selection procedure: Solutions generated by Genetic Algorithm produce various fitness values. Significant portion of the generated solutions are mostly useless with insignificant fitness values, some solutions produces minor (middle level) fitness values and very few solutions produces significant fitness values. In order to speed up calculation and keep convergence of the GA, selection procedure is needed. Selection procedure assigns a probability (chance) to each solution according to fitness value. Thus, solutions with significant fitness values have higher chance to be selected to generate new solution using crossover or mutation. Solutions with insignificant fitness values have negligible chance and mostly skipped.

In this paper geometric ranking method [17] is used as a selection procedure. In geometric ranking method all solutions are sorted in descending order of its fitness value. The probability (chance) of any solution j to be selected is calculated as follows:

$$P_j = q'(1-q)^{r_j - 1} \qquad (4)$$

where $q' = \frac{q}{1-(1-q)^N}$

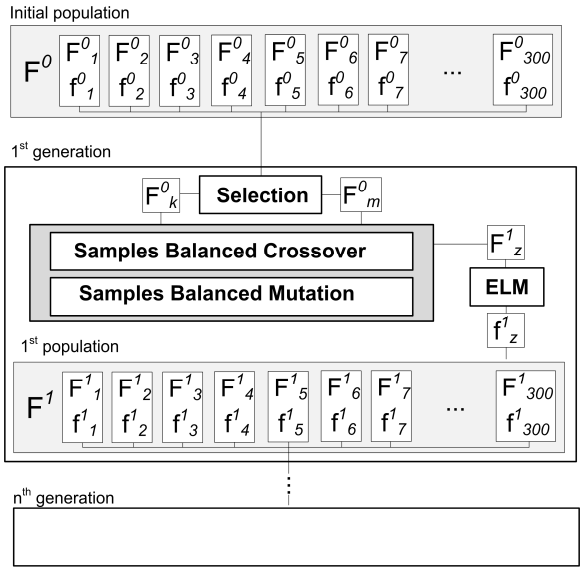
q' is selection probability, r_j is a rank of j-th solution in the partially ordered set, and N is the population size. The detail explanation of the geometric ranking method is given in [17]. Parameter q is 10^{-3}.

Termination criteria: Genetic Algorithm stops if the maximum number (100) of generation is reached.

SBGA-ELM framework contains following functional units: Binary solution - F, fitness value f, ELM classifier/fitness function block "ELM", Selection procedure block "Selection", Crossover block "Samples Balanced Crossover", mutation block "Samples Balanced Mutation" (see Fig 4.). Proposed SBGA-ELM processes 100 generations starting from initialization step. Each n-th generation uses binary solutions from population F^{n-1} to create a set of new binary solutions for a new population F^n. Each solution $F_i^n$from population $F^n$is evaluated by fitness function/ELM classifier. Binary solution links to the set of genes from ParkDB data base. The set of chosen features is used to train a set of 20 ELM classifiers based on random parameters. Fitness value $f_i^n$is then calculated using Equation 1. Finally, each binary solution F_i^n from n-th population F^ncouples with corresponding fitness value f_i^n.

SBGA-ELM starts from the initialization procedure (see Fig. 4). Initial population F^0 contains 300 randomly generated binary solutions and corresponding fitness values computed using Equation 1. The 1st generation of SBGAELM starts from selection procedure. Selection procedure manages a chance of each binary solution from initial population F^0 to be selected to generate a new binary solution for population F^1 using Samples Balanced Crossover or Mutation. Each generated solution F_i^1 from population $F^1$is evaluated using fitness function, and corresponding fitness value f_i^1 is calculated using Equation 1. In each

generation crossover creates 70% or 210 new solutions, mutation creates rest 30% or 90 new solutions.



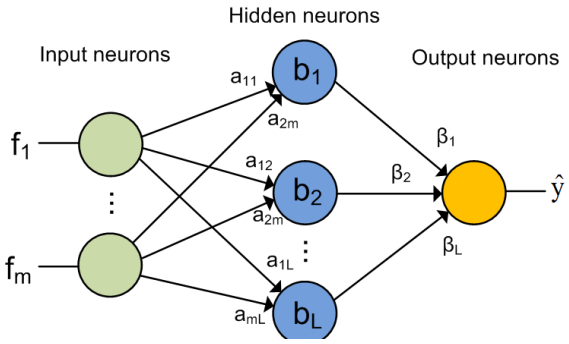
(Figure 4) Framework of the proposed SBGA-ELM

1.2 Extreme Learning Machine (ELM)

Presented PD classification problem is 2 classes classification problem (or binary classification problem) with limited number of samples (72 samples: 50 PD patients and 22 normal persons) and high dimensional features space (22283 features). Proposed SBGA-ELM efficiently solves PD classification problem by reducing a problem feature space and applying Extreme Learning Machine to build a PD classifier.

Extreme Learning Machine with Gaussian neurons is used to approximate functional relationship between reduced set of features and coded class labels. ELM is a single hidden layer feed-forward neural network (see Fig. 5) where input weights $\{a_{11},a_{2m},a_{12},a_{2m},...,a_{mL}\}$ of input neurons and bias $\{b_1,b_2,...,b_L\}$ of the hidden neurons are randomly assigned and output weights $\{\beta_1,\beta_2,...,\beta_L\}$are estimated analytically [14].

ELM classifier is developed as follows: Training data contains N samples taken from ParkDB data base,



(Figure 5) Framework of the Extreme Learning Machine.

$\{(X^1,c^1),...,(X^t,c^t),...,(X^N,c^N)\}$. Where $\mathbf{X}$ is m-dimensional feature vector and $c^t \in 1,2$ is class label for t-th sample. The coded class label y^t is calculated as follows:

$$y^t = \begin{cases} 1, & \text{if } c^t = k \\ -1, & otherwise \end{cases} \quad k = 1,2 \qquad (5)$$

$y = \{y^1,y^2,...,y^N\}$ is a vector with all coded class labels. Then, training data is a combination of feature vectors X and coded class labels y: $\{(X^1,y^1),...,(X^t,y^t),...,(X^N,y^N)\}$.

Any classifier L is a function, which maps m-dimensional features from training set to corresponding coded class label. Discovered function predicts class labels for testing samples with certain accuracy.

Assume that is the number of hidden neurons and ELM employs Gaussian activation function. Then the response of j-th hidden neurons for the t-th sample is calculated as:

$$g_j^t = \exp\left(-\left(\frac{(X^t - A_j)^T \cdot (X^t - A_j)^T}{2 - b_j^2}\right)\right) \qquad (6)$$

where A_j and b_j are the center and width of the j-th hidden neuron, $X^t$is the set of input features for t-th sample. $A_j = \{a_{1j},a_{2j},a_{3j},...,a_{mj}\}$ is a vector with input

weights of the j-th hidden neuron. $A=\{A_1,A_2,...,A_L\}$ is the set of all input weights. $B=\{b_1,b_2,...,b_L\}$ is the set of Gaussian widths of hidden neurons.

Predicted coded class label for t-th sample is calculated as:

$$y^t=\sum_{j=1}^{L}\beta_j\cdot\ g^{t_j}(7)$$

where β_j is an output weight of j-th hidden neuron.

Matrix form of Equation 7 is:

$$\hat{y}=\beta\cdot\ G \quad (8)$$

where $\beta=\{\beta_1,\beta_2,...,\beta_L\}$ is a set of output weights, $\hat{y}$ is a set of predicted coded class labels, G is a hidden layer output matrix:

$$G=\begin{pmatrix} g_1^1 \dots g_1^N \\ \vdots \quad \vdots \\ g_L^1 \dots g_L^N \end{pmatrix} \quad (9)$$

The output weight β is computed using Moore-Penrose generalization inverse as follows:

$$\beta=y\cdot\ G^+ \quad (10)$$

Framework of the Extreme Learning Machine is summarized as follows:

1) Generate sets of input weights A and width B of hidden Gaussian neurons randomly.

2) Compute the hidden layer output matrix G (9)

3) Compute output weights β using Equation 10.

4) Compute predicted coded class labels $\hat{y}$ using Equation 8.

5) Define predicted class label as follows:

$$\hat{c_k}=\begin{cases}1, \text{if } \hat{y_k}\geq\ 0\\ 2, \text{if } \hat{y_k}<0\end{cases}$$

Accuracy of the ELM classifier in the proposed SBGA-ELM depends on the randomly chosen centers and hidden neuron bias. Each binary solution is used to create 20 ELM classifiers for more accurate analysis. Fitness value computed using Equation 1 is a combination of means of classification accuracies of 20 ELM classifiers and penalty factors. Such strategy neglects effect of randomness and keeps searching for a robust solution, which may identify PD bio-markers.

4. Experimental results

In this section experiments with proposed SBGA-ELM are presented. Experiments conduct 1) searching for a set of robust features (genes) from ParkDB data base, 2) training an efficient PD classifier and 3) searching bio-markers responsible for PD. Proposed PD classifier examines 72 samples from ParkDB data base.

Proposed SBGA-ELM discovers many genes subsets, which creates PD classifier with absolute classification accuracy (i.e. 100%).Presented SBGA-ELM design continues searching for a set of robust features even if absolute classification accuracy is reached. Due to penalty factor f_2 (see Equations 1 and 3), the proposed SBGA-ELM efficiently differentiates two solutions with absolute classification accuracies. Solution with higher penalty factor has higher fitness value and vice versa. Thus, such solution has higher chance to pass selection procedure in GA.

After 100 iterations proposed SBGA-ELM collects PD classifiers with different fitness values. Hence, a set of 10 PD classifiers with fitness value closest to maximum has been collected. Each PD classifier has been trained based on unique set of genes chosen by SBGA. Thus, 10 PD classifiers merge 10 corresponding sets of genes into one appearance table (see Table 1). Appearance table also collects a number of times each gene appears in a table (see Table 1 "counts"). According to experiment logic genes, which are appeared more frequently in an appearance table and have high "count" are more important for PD diagnosis and can be considered as

bio-markers. Such genes have been chosen by SBGA-ELM more frequently to build PD classifiers with fitness values close to maximum.

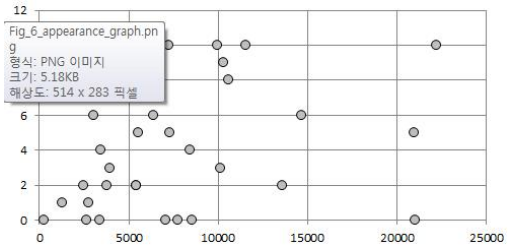
Appearance graph (see Fig. 6) displays "count" from appearance table in respect of genes ID. 4 genes have maximum possible appearance 10, which means those 4 genes appear in all Top-10 PD classifiers. 4 discovered genes are definitely bio-markers. 8 genes have appearance from 8 to 10. 15 genes have appearance from 5 to 10. Finally, 24 genes appeared more than once in appearance table are considered as discovered bio-markers. Importance of discovered genes is proven by proposed analysis.

A set of discovered bio-markers has been tested by SVM and PBL-McRBFN [11] adapted to PD classification problem. SVM is a popular machine learning technique for PD classification. 24 discovered genes have been used to build both SVM and PBL-McRBFN classifiers. Both methods show maximum possible classification accuracy 100%.

	slot #1	*slot #2*	*slot #3*	*slot #4*	*slot #5*	*slot #6*		*slot #10*		Counts
1	1	1	0	0	0	0		1		3
2	0	1	0	0	0	1		0		2
3	1	0	0	1	0	0		0		2
4	0	0	0	0	0	0		0		0
5	0	0	1	1	0	1		0		3
n	1	0	0	0	0	1		0		2

n=22,283

<Table 1> Appearance table



(Figure 6) Appearance graph

5. Conclusion

In this paper, we have presented a PD diagnosis approach and bio-markers detector based on Samples Balanced Genetic Algorithm coupled with Extreme Learning Machine. Proposed method uses gene expression data taken from ParkDB data base for experiments. Proposed Samples Balanced Genetic Algorithm designed specifically for PD classification problem searches an optimal set of robust features (genes) for further analysis. Chosen features are used to build a PD classifier based on Extreme Learning Machine. Finally proposed SBGA-ELM creates an efficient PD classifier with maximum possible classification accuracy 100% and selects a set of 24 bio-markers probably responsible for Parkinson's Disease. Discovered bio-markers have been verified by SVM and PBL-McRBFN.

Acknowledgement

This work was supported by Catholic University of Korea, Research Funds 2016

References

[1] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig. "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease." IEEE Transactions on Biomedical Engineering, vol. 56, pp. 1015- 1022, 2009

[2] M. F. Caglar, B. Cetisli, and I. B. Toprak. "Automatic recognition of Parkinson's disease from sustained phonation tests using ANN and adaptive neuro-fuzzy classifier". Journal of Engineering Science and Design, vol. 1, pp. 5964, 2010

[3] C. Sakar and O. Kursun. "Telediagnosis of Parkinson's disease using measurements of dysphonia". Journal of Medical Systems, vol. 34, pp. 591 - 599, 2010.

[4] R. Das. "A comparison of multiple classification methods for diagnosis of Parkinson disease". Expert Systems with Applications, vol. 37, pp 1568 - 1572, 2010.

[5] G. Sateesh Babu, S. Suresh, Uma Sangumathi and H.J. Kim. "A projection based learning meta-cognitive RBF network classifier for effective diagnosis of Parkinson's disease". Advances in Neural Networks ISNN 2012. Lecture Notes in Computer Science, vol. 7368, pp. 611 - 620, 2012.

[6] M. Engin, S. Demirag, E.Z. Engin, G. Celebi, F. Ersan, E. Asena, Z. Colakoglu. "The classification of human tremor signals using artificial neural network." Expert Systems with Applications, vol. 33, pp 754761, 2007.

[7] S. Pan, S. Iplikci, K. Warwick, and T. Z. Aziz. "Parkinson's Disease tremor classification: A comparison between support vector machines and neural networks". Expert Systems with Applications, vol. 39, pp. 10764 - 10771, 2012

[8] M.N. Tahir and H.H Manap. "Parkinson disease gait classification based on machine learning approach". Journal of Applied Sciences, vol. 12, pp. 180 - 185, 2012.

[9] C. R. Scherzer, A.C. Eklund, L.J. Morse, Z. Liao,J. J. Locascio,D. Fefer, M. A. Schwarzschild, M. G. Schlossmacher , M. A. Hauser, J. M. Vance, L. R. Sudarsky, D. G. Standaert, J. H. Growdon, R. V. Jensen, and S. R. Gullans. "Molecular markers of early Parkinsons disease based on gene expression in blood". Proceedings of the National Academy of Sciences, vol. 104, pp. 955 - 960, 2007

[10] C. Taccioli, V. Maselli, J. Tegner, D. Gomez-Cabrero, G. Altobelli, W, Emmett, F. Lescai, S. Gustincich, and E. Stupka. "ParkDB: A Parkinsons disease gene expression database". http://database.oxfordjournals.org/content/2011/bar007, 2011.

[11] G. Sateesh Babu, S. Suresh, B. S. Mahanand, "A novel PBL-McRBFN-RFE approach for identification of critical brain regions responsible for Parkinson s disease", Expert System with Applications, vol. 41 no. 2, pp. 478-488, 2014.

[12] G. Sateesh Babu, S. Suresh, B. S. Mahanand, " Parkinsons disease prediction using gene expression A projection based learning meta-cognitive neural classifier approach", Expert System with Applications, vol. 40, no. 5, pp. 1519-1529, 2013.

[13] S. Saraswathi, S. Suresh, N. Sundararajan, M. Zimmermann and M. Nilsen-Hamilton, "ICGA-PSO-ELM Approach for Accurate Multiclass Cancer Classification Resulting in Reduced Gene Sets in Which Genes Encoding Secreted Proteins Are Highly Represented", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, pp. 452 - 463, 2011.

[14] G.-B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: theory and applications", Neurocomputing, vol. 70, no. 1-3, pp. 985990, 2006.

[15] G. K. Smyth. "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments". Statistical Applications in Genetics and Molecular Biology, Article 3, 2004.

[16] S. Suresh, S. N. Omkar, V. Mani, T. N. G. Prakash, "Lift coefficient prediction at high angle of attack using recurrent neural network", Aerospace Science and Technology, vol. 7, pp. 595 602, 2003

[17] S. Suresh, S. N. Omkar, V. Mani, T. N. G. Prakash, "Lift coefficient prediction at high angle of attack using recurrent neural network", Aerospace Science and Technology, vol. 7, pp. 595 602, 2003

[18] L. V. Ma, S. H. Park, J. H. Jang and J. H. Park, "Fuzzy Decision Making-based Recommendation Channel System using the Social Network Database, " J. of Digital Contents Society, Vol.17, No.5, 2016

Vasily Sacnev

2002년 : Komsomolsk-na-Amure State Tech. University, (B.S)
2004년 : Komsomolsk-na-Amure State Tech. University, (M.S)
2009년 : Korea University (PhD)

2010년～현재: Catholic University, Assistant Professor
관심분야 : Multimedia Security, Steganography, Steganalysis, Machine learning and Bio-informatics

Sundaram Suresh

1999년 : Bharathiyar University, INDIA(B.E)
2001년 : Indian Inst. of Science Bangalore, INDIA(M.E)
2005년 : Indian Inst. of Science Bangalore, INDIA (Ph.D)

2005년～2007: Nanyang Tech. University, post-doctoral
2007년～2008: Indian Inst. of Tech. -Delhi, Assistant Professor
2010년～현재: Nanyang Tech. University, Assistant Professor
관심분야 : Computational cognitive system, Neural networks, Intelligent control, Medical image processing, Mathematical optimization and Game theory

최용수

1998년 강원대학교 제어계측공학과 공학사
2000년 강원대학교 제어계측공학과 공학석사
2006년 강원대학교 제어계측공학과 공학박사
2006년～2007년 연세대학교 첨단융합건설연구단 연구교수
2007년～2013년 고려대학교 정보보호대학원 연구교수
2013년～현재 성결대학교 교양교직부 (멀티미디어) 조교수
관심분야 : Multimedia Hashing, Information Hiding, Watermarking, Steganography, Image Forensics, Forgery Detection 등