IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 6, January 2017)

Diagnosis of Diabetes Mellitus Using PCA, Neural Network and Cultural Algorithm

Anjali Khandegar
khandegaranjali@gmail.com

Khushbu Pawar
khushii25@gmail.com

Abstract –**Diabetes Mellitus (DM) is a chronic metabolic disorder characterized by the hyperglycaemia with the disturbances of body metabolism resulting from the defects in insulin secretion, insulin action, or both. The chronic hyperglycaemia of diabetes is associated with the long term damage, dysfunction and failure of various vital organs like, the eyes, kidneys, nerves, heart and the blood vessels leading to micro and macro vascular complications.**

Data mining gives a diversity of methods to investigate large data keeping in mind the end goal to find hidden knowledge. This study is an effort to plan and execute a descriptive data mining approach and to devise association standards to envisage diabetes behaviour in arrangement with particular life style parameters, including physical activity and emotional states, especially in elderly diabetics. Proposed methodology is based on Principal Component Analysis (PCA), Neural Network (NN) and Cultural Algorithm (CA).

Keywords – **CA, Data mining, Diabetes mellitus, NN, PCA.**

I. INTRODUCTION

Diabetes is one of the most dangerous diseases, also known with the name “silent killer”. This disease is a major health problem in industrialized and developing countries, and its incidence is increasing with over 220 million people with diabetes worldwide. It is the fourth leading cause of death. The increase of diabetes is so fast that the World Health Organization (WHO) has identified as an epidemic.

The World Health Organization defines diabetes as a metabolic disorder of multiple etiologies characterized by chronic hyperglycemia with disturbances of carbohydrate metabolism, lipid and protein resulting in insulin secretion defects, action insulin, or both [1].

Most of what we eat is broken down into glucose to be used by our cells to produce energy. However, glucose cannot enter cells without the presence of insulin, a hormone produced not pancreas. After eating, the pancreas automatically secretes enough insulin to transport glucose from the blood to cells and decrease the sugar level in the blood. A person who has diabetes suffering from hyperglycaemia that is to say, the amount of glucose in the blood is too high. This is because the body does not produce enough insulin, produces no insulin or the cells do not respond properly to insulin produced by the pancreas.

Causes of Diabetes

The prevalence of this disease has increased fivefold in less than fifty years. This gradual increase is due to several factors [2]:

- The overall aging of the population,
- Increased life expectancy of diabetic
- Increasing the fertility of diabetic women
- The increase in obesity,
- Incrementing the consumption of refined sugars.

As well as other factors which may serve as a trigger such as:

- Sedentary lifestyle,
- Diets rich in fat and protein,
- the reduced fiber consumption,
- A diet deficient in complex carbohydrate and vitamin E.
- Chronic stress.
- Smoking can cause the onset of insulin resistance.

Classification of Diabetes

The criteria for the diagnosis and classification of diabetes mellitus were developed by an expert committee of the American Diabetes Association (ADA) and by a WHO committee.

The classification of diabetes is mainly based on its aetiology and pathophysiological feature. Diabetes is classified into four types:

- Diabetes type 1 (DM1);
- Diabetes type 2 (DM2);
- Other specific types of diabetes;
- Gestational diabetes (DMG).

Frequently people with DM2 eventually need insulin at a stage in their life, on the other hand, some DM1 patients may progress slowly or have long periods of remission without the need for insulin. It is because

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 6, January 2017)

of these cases that the insulin and non-insulin dependent terms were eliminated [3].

II. Diagnosis of Diabetes

For several years, the level of glucose in the blood to diagnose diabetes was high but in 1997 the standard level for a normal glucose level was reduced because many people suffering from the complications of diabetes even they were not diabetic according to the standards. In 2003 the standard was changed again. After much debate, the ADA published the new standard for diagnosis, based on the following criteria [4]:

Random Plasma Glucose (RPG): The glucose level when the patient has eaten normally before the analyses. It must be ≥ 200 mg / dl and accompanied by classic symptoms of diabetes.

Fasting Plasma Glucose (FPG): The glucose level when the patient is not consumed during the eight hours before analysis. It must be ≥ 126 mg / dl.

Oral Glucose Tolerance Test (OGTT): The analysis is performed two hours after ingesting 75g of glucose by oral route. It must be ≥ 200 mg / dl.

III. Diabetes Mellitus

Diabetes is a standout amongst the most well-known non-transmittable diseases in the world. It is assessed to be the fourth or fifth cause for death in most of the nation [1].

Diabetes is an unending malady that happens when the body cannot sufficiently deliver insulin or cannot utilize it adequately which brings about BG (Blood Glucose) not to be ingested appropriately by the body cells and stays circulation in the blood.

Categories of diabetes:

- Type-1 diabetes
- Type-2 diabetes
- Gestational diabetes

Type 1 diabetes incorporates diabetes that is basically an aftereffect of pancreatic beta cell devastation that prompts insufficient creation of insulin. This type can influence any age yet typically happens in youngsters and youth. These diabetics can lead a typical life through mix of a day by day insulin treatment, solid eating regimen, close checking and normal physical activity [2].

Type 2 diabetes is the most well-known, one that generally happens in adult peoples yet is progressively seen in kids and young people, as well. This type is otherwise called an insulin resistance in light of the fact that, in this type the body can create insulin yet possibly it is not adequate or the body can't react to its impact leading glucose remains flowing in the blood. This type likewise incorporates LADA (Latent Autoimmune Diabetes in Adults), depicting a lessened number of individuals with diabetes type 2 who seem to have an insusceptible intervened loss of pancreatic beta cells [3]. Numerous type 2 diabetics can control their BG level through a healthy eating routine and an expanded physical movement.

Gestational diabetes mellitus alludes to a glucose intolerance with onset or first acknowledgment amid pregnancy because of ineffectively oversaw BG. This collection must be nearly checked to control their BG level and minimize the danger for the child. This could be possible by healthy eating regimen, moderate physical activity and at times insulin treatment or oral drug [2].

Proposed framework helps the patients from experiencing different tests like blood tests, checking diastolic and systolic blood pressure and so on. In the wake of having this frameworks, the patent itself is capable to analyze whether he/she is experiencing Diabetes Mellitus or not. These estimations are in view of the indications' which happens in right on time phases of Diabetes Mellitus and on some physical conditions.

IV. Database Explanation

Title: Pima Indians Diabetes Database

Sources:

Original owners: National Institute of Diabetes and Digestive and Kidney Diseases

Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu) Research Centre, RMI Group Leader, Applied Physics Laboratory, The Johns Hopkins University, Johns Hopkins Road, Laurel, MD 20707 (301) 953-6231(c)

Date received: 9 May 1990.

Past Usage: **Smith et al,** [4] investigated the diagnostic, binary-valued variable whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona, USA.

Results: Their ADAP algorithm makes a real-valued prediction between 0 and 1. This was transformed into a binary decision using a cut-off of 0.448. Using 576 training instances, the sensitivity and specificity of their algorithm was 76% on the remaining 192 instances.

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 6, January 2017)

Relevant Information: Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogy of perceptron-like devices. It is a unique algorithm; see the paper for details.

Number of Instances: 768
Number of Attributes: 8 plus class
For Each Attribute: (all numeric-valued)

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

Class Distribution: (Class value 1 is interpreted as "tested positive for diabetes")

Class Value	Number of instances
0	500
1	268

Brief Statistical Analysis:

Attribute Number	Mean	Standard Deviation
1	3.8	3.4
2	120.9	32.0
3	69.1	19.4
4	20.5	16.0
5	79.8	115.2
6	32.0	7.9
7	0.5	0.3
8	33.2	11.8

60% of data used for training
40% of data for testing.

V. PROPOSED METHODOLOGY

A person suffering from diabetes mellitus cannot be cured i.e. diabetes mellitus can be controlled but there are no solution to permanently cure the disease. A person suffering from diabetes possesses a history of data related to his blood glucose level. Thus to control diabetes an individual must always check on its glucose level, to see whether he has controlled glucose level. The blood glucose level is the measure of the severity of diabetes in an individual. It helps him to take proper medicines, diet and exercise so that he has normal glucose level. If required he has to take insulin doses. The blood glucose value is not constant that is it changes with time and for a same person in a day we can have different BG values. Thus BG values are temporal in nature. Thus the changing value of BG level can be analyzed. This research work is based on the prediction of diabetes behaviour in arrangement with particular life style parameters. PCA and Neural Network classifiers are used for classification. Furthermore the Cultural Algorithm is used to optimize the Neural Network for performance improvement.

Figure 1 shows block diagram for proposed research work and the description of proposed methodology is as follows:

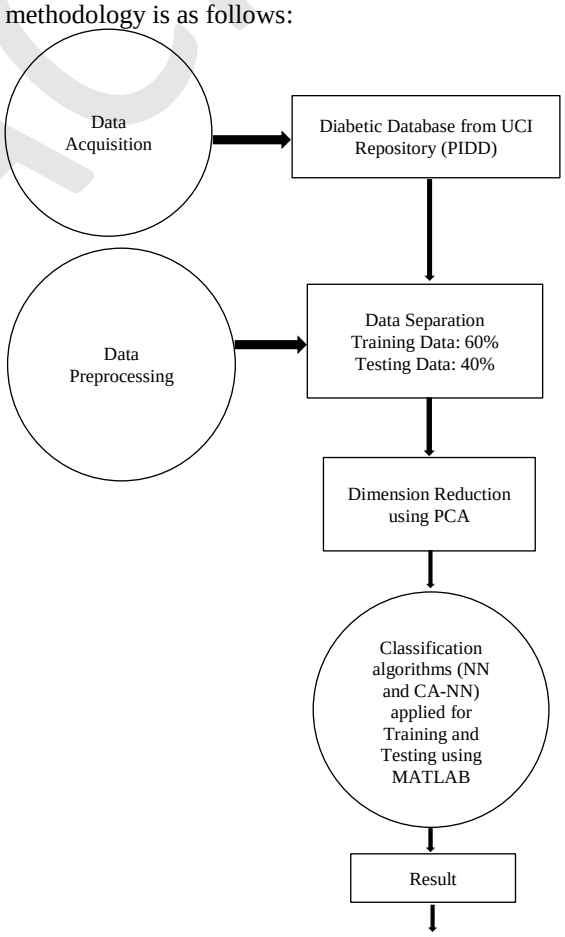

Figure 1: Block diagram for proposed research work

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 6, January 2017)

Data Acquisition Process

Data acquisition is the process of organizing and collecting information. The data is collected from Pima Indian Diabetic Database (PIDD) at the UCI (University of California Irvine) Machine learning lab. There are total 768 number of instances out of which 268 cases are of diabetic and 500 are non-diabetic.

Data Preprocessing and Data Separation

Data preprocessing is a data mining technique that transform raw data into an understandable format. There are missing outliers present in database. Data separation perform the task of dividing database into training and testing set. In proposed methodology the training data is of 60% and testing database is of 40%. So there are 460 instances which are used for training and rest 307 are used for testing of PCA-Neural Network and Cultural Algorithm Optimized Neural Network classifier algorithms.

Dimension Reduction using PCA

Principal Component Analysis is used to reduce the dimension of extracted features. As other transformation like Fourier transformation, PCA transform the data into another representation where a new set of basis vectors are used. In PCA, however, these basic vectors are not constant as in the case in many other transformation. Instead they are calculated based on the data to be transformed.

Mathematical Description of PCA

A. Statistics

The entire subject of statistics is based around the idea that we have this big set of data, and we want to analyze that set in terms of the relationships between the individual points in that data set.

B. Standard Deviation

To understand standard deviation, we need a data set. Statisticians are usually concerned with taking a sample of a population. To use election polls as an example, the population is all the people in the country, whereas a sample is a subset of the population that the statisticians measure. Here's an example set:

$$X = [1\ 2\ 4\ 6\ 12\ 15\ 25\ 45\ 68\ 67\ 65\ 98\ ]$$

X refers to this entire set of numbers. The mean of the sample is given by the formula:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad (1)$$

The Standard Deviation (SD) of a data set is a measure of how spread out the data is. SD is given by the formula

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{(n-1)}} \quad (2)$$

C. Variance

Variance is another measure of the spread of data in a data set. In fact it is almost identical to the standard deviation. The formula is simply the SD squared.

$$s^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{(n-1)} \quad (3)$$

D. Covariance

Covariance is always measured between two dimensions. If you calculate the covariance between one dimension and itself, you get the variance. So, if you had a three-dimensional data set (x, y, z), then you could measure the covariance between x and y dimensions, the x and z dimensions and the y and z dimensions. The formula for covariance is very similar to the formula for variance. The formula for variance could also be written like this,

$$var(x) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})}{(n-1)} \quad (4)$$

The formula for covariance is given by,

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{(n-1)} \quad (5)$$

The exact value of covariance is not as important as its sign (i.e. Positive or negative). If the value is positive, then it indicates that both dimensions increase together. If the value is negative, then as one dimension increases, the other decreases. If the covariance is zero, it indicates that the two dimensions are independent of each other.

E. Covariance Matrix

Covariance (cov) is always measured between two dimensions. If we have a data set with more than two dimensions, there is more than one cov measurement that can be calculated. For example, from a three dimensional data set ($dimensions\ x, y, z$) you could calculate $cov\ (x, y)$, cov (y, z), cov (y, z). In fact, for an n dimensional data set, you can calculate different covariance values.

$$\frac{n!}{(n-2)!*2} \quad (6)$$

A useful way to get all the possible cov values between all the different dimensions is to calculate them all and put them in a matrix. So, the definition for the cov matrix for a set of data with n dimensions is:

$$C^{m\times n} = \left(c_{i,j}, c_{i,j} = cov(Dim_i, Dim_i)\right) \quad (7)$$

Where, $C^{m\times n}$ is a matrix with *n* rows and *n* columns and Dim_x is the x^{th} dimension. The entry on row 2, column 3 is the cov value calculated between the 2nd dimension and the 3rd dimension.

F. Eigenvalues and Eigenvectors of Matrix

If the matrix is small, we can compute them symbolically using the characteristic polynomial. However, this is often impossible for larger

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 6, January 2017)

matrices, in which case we must use a numerical method.

An important tool for describing eigenvalues of square matrices is the characteristic polynomial: saying that λ is an eigenvalue of A is equivalent to stating that the system of linear equations $(A - \lambda I)\, v = 0$ (where I is the identity matrix) has a non-zero solution v (an eigenvector), and so it is equivalent to the determinant:

$$\det(A - \lambda I) = 0 \qquad (8)$$

The function $p(\lambda) = det(A - \lambda I)$ is a polynomial in λ since determinants are defined as sums of products. This is the characteristic polynomial of A: the eigenvalues of a matrix are the zeros of its characteristic polynomial.

All the eigenvalues of a matrix A can be computed by solving the equation $pA(\lambda) = 0$. If A is a $n \times n$ matrix, then pA have degree n and A can therefore have at most n eigenvalues. Conversely, the fundamental theorem of algebra says that this equation has exactly n roots (zeroes), counted with multiplicity. All real polynomials of odd degree have a real number as a root, so for odd n, every real matrix has at least one real eigenvalue. In the case of a real matrix, for even and odd n, the non-real eigenvalues come in conjugate pairs.

Once the eigenvalues λ are known, the eigenvectors can then be found by solving:

$$(A - \lambda I)v = 0 \qquad (9)$$

An example of a matrix with no real eigenvalues is the 90-degree clockwise rotation (equation 9) whose characteristic polynomial is $\lambda^2 + 1$ and so its eigenvalues are the pair of complex conjugates $i, -i$. The associated eigenvectors are also not real.

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \qquad (10)$$

Classification

Neural Network (NN) is used as the classifier. Here, we are showing comparison of PCA-NN and PCA-CA-NN approaches.

Neural Network

Back Propagation Neural Network (BPNN) generates complex decision boundaries in feature space. BPNN in specific circumstances resembles Bayesian Posterior Probabilities at its output. These conditions are essential to achieve low error performance for given set of features along with selection of parameters such as training samples, hidden layer nodes and learning rate. In else case, the performance of BPNN could not be evaluated. For W number of weights and N number of nodes, numbers of samples (m) are depicted to correctly classify future samples in following manner:

$$m \geq O\left(\frac{W}{\epsilon}\log\frac{N}{\epsilon}\right) \qquad (11)$$

The theoretical computation of number of hidden nodes is not a specific process for hidden layers. Testing method is commonly entertained for selection of these followed in the constrained environment of performance [5].

Back-Propagation Algorithm

The back-propagation algorithm for a 3-layer network (only one hidden layer) is as follows:

initialize the weights in the network (often small random values)
do
  for *each data i in the training set of*
    database O = neural-network-output(network, i)
    T = desired output for i
    calculate error (T - O) at the output units;
    calculate δ_h for all weights from hidden layer to output layer;
    calculate δ_i for all weights from input layer to hidden layer;
    update the weights to minimize error in the network; ***until*** *some stopping criterion satisfied*
return *the network*

In this proposed NN, the value of weight and bias are random and to correct these values a fitness function is employed for Cultural Algorithm.

Fitness Function: The fitness function is a function of weight and bias with the objective of minimizing the mean square error between the predicted and target classes of the training data.

$$\min F(w, v) = \sum_{t=1}^{q}[c_t - (wx_t + v)]^2 \qquad (12)$$

Where, x_t is input and c_t is target output.

Fitness function in equation (12) is minimized using Cultural Algorithm to optimized weight w and bias v values.

Cultural Algorithm

Inspired by the process of social and cultural changes, the CA was developed to enhance evolutionary computation. Besides the population component that evolutionary computation approaches have, there is an additional peer component belief space and a supporting communication protocol between these two components, which makes CAs perform better in some special optimal cases than other evolutionary algorithms (EAs). The following figure presents the basic CA framework.

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 6, January 2017)

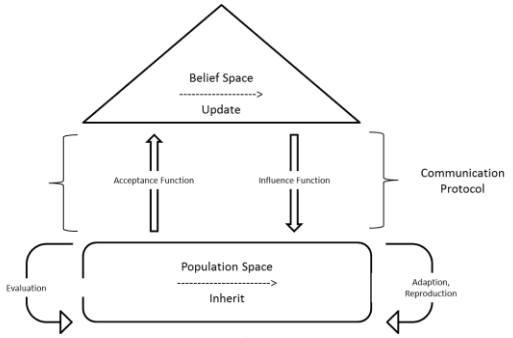

Figure 2: CA framework [6]

As Figure 2 shows, the population space and the belief space can evolve respectively. The population space consists of the autonomous solution agents and the belief space is considered as a global knowledge repository. The evolutionary knowledge that stored in belief space can affect the agents in population space through influence function and the knowledge extracted from population space can be passed to belief space by the acceptance function.

In the process of the CA evolution, the population space is initialized with candidate solution agents at random, meanwhile, the initial knowledge sources in the belief space are built. At first the two spaces evolve independently. Then the selected agents from the population space are used to update the belief space. After the knowledge sources being updated, the belief space will reversely guide the evolution of the population space. These procedures repeat till a termination condition has been reached. The CA pseudo code presented by [6] is given as follows:

t=0;
Initialize Population POP(t);
Initialize Belief Space BLF(t);
Repeat
Evaluate Population POP(t);
Adjust (BLF(t), Accept(POP(t)));
Adjust (BLF (t));
Variation(POP (t) from POP (t-1));
Until termination condition achieved

Result

Finally we get the accuracy of proposed method. Chances of patient being diabetic in form of percentage will be calculated also, so that corrective measures can be taken timely.

VI. Simulation And Results

The performance of proposed technique has been studied by means of MATLAB simulation.

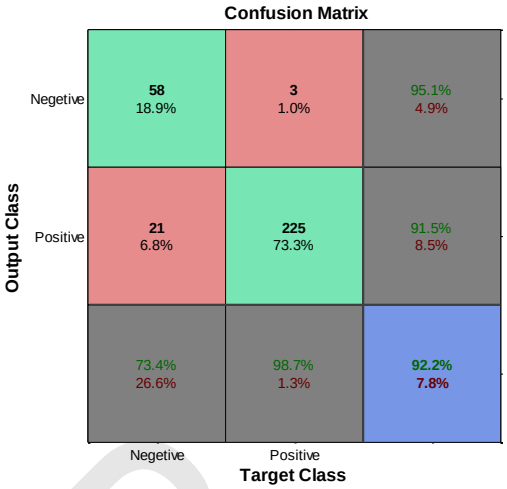

Figure 3: Confusion matrix plot for PCA-NN approach

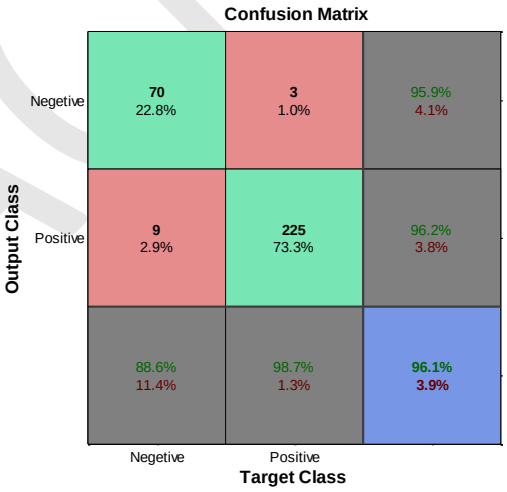

Figure 4: Confusion matrix plot for PCA-CA-NN approach

Table 1: Comparative analysis with previous work

Methods	Accuracy
NN [7]	72.9%
ANFIS [7]	70.56%
PCA-ANFIS [7]	89.2%
PCA+NN [7]	90.49%
PCA+NN (Proposed)	**92.2%**
PCA+CA+NN (Proposed)	**96.1%**

VII. Conclusion

In this paper, Cultural Algorithm Optimized Neural Network classifier has been used in diagnosis of

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 6, January 2017)

diabetes mellitus. PCA is used for dimensionality reduction of extracted features. Training overhead is reduced around 30% with the help of PCA and thus the computational complexity is minimized. On observing results it can be said that the enhancement in accuracy is achieved using Cultural Algorithm. Maximum accuracy achieved by proposed approach is **96.1%**.

REFERENCE

[1] IDF Diabetes Atlas, 6th edition, "International Diabetes Federation", 2013, Online available at: http://www.idf.org/diabetesatlas
[2] "Definition, classification and diagnosis of diabetes, Prediabetes and metabolic syndrome", Canadian Journal of Diabetes, Vol. 37, Canadian Diabetes Association. 2013. Online available at: www.canadianjournalofdiabetes.com.
[3] "Diagnosis and classification of diabetes mellitus", American Diabetes Association, Diabetes Care, vol. 35, Supplement 1, 2012.
[4] Jack W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, and R.S. Johannes, "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus", IEEE Symposium on Computer Applications and Medical Care, pp. 261-265, 1988.
[5] Christos Stergiou and Dimitrios Siganos, "Neural Networks", Report available at: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html
[6] Reynolds, R.G.: An introduction to cultural algorithms. In: Proceedings of the third annual conference on evolutionary programming, Singapore, pp. 131-139, 1994.
[7] Motka, Rakesh, Viral Parmarl, Balbindra Kumar, and A. R. Verma, "Diabetes mellitus forecast using different data mining techniques", IEEE 4th International Conference on Computer and Communication Technology (ICCCT), pp. 99-103. IEEE, 2013.