International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-4 Issue-9, February 2015

Hybrid Artificial Bee Colony Algorithm and Semi Supervised Learning Prediction Model for the Risk of Cardiovascular Disease in Type-2 Diabetic Patients

P. Radha, B. Srinivasan

Abstract*— Cardiovascular disease (CVD) factor is one of the important causes of mortality among diabetes patients. Statistics shows that more than 22% of people with type 2 diabetes mellitus suffer from CVD and which in turn leads to cardiovascular disease. But still some of the works doesn't mainly focus on the semisupervised learning methods with feature selection methods to enhance the prediction accuracy of the classification methods. The aim of this research was to identify significant CVD factors influencing type 2 diabetes controls to improve prediction accuracy. In proposed methods the preprocessing and dimensionality reduction of the patients records is done by using Kullback Leiber Divergence(KLD) –Principal component analysis (PCA) ,then attribute values measurement is completed by using kernel density estimation (KDE) which measures the attributes values based on probability mass function with Gaussian kernel function, feature selection is performed by using artificial bee colony with differential evolution (ABC-DE). Hybrid prediction model Improved Fuzzy C Means (IFCM) clustering algorithm aimed at validating chosen class label of given data and subsequently applying semisupervised Modified Self-Organizing Feature Map Neural Network (MSOFMNN) classification algorithm to the result set. The proposed method examines the behavioral factors that contribute to CVD risk factors among those with type 2 diabetes (T2D) with higher prediction accuracy, less error rate.

Index Terms*— Artificial bee colony (ABC), Classification, Hybrid Prediction Model, Kernel density estimation (KDE), Modified Self-Organizing Feature Map Neural Network (MSOFMNN).

I. INTRODUCTION

People with type 2 diabetes have a twofold increased risk of CVD [1]. Guidelines for the management of type 2 diabetes advocate calculating CVD risk to guide the initiation of appropriate treatment [2-3]. Over the past decades many prediction models have been developed to predict CVD, of which only a small number have been specifically developed for people with type 2 diabetes [4]. Type-2 diabetes (T2D) is caused by relative insulin deficiency. Pancreas in Type-2 diabetes still produces insulin but it may not be effective or may not produce sufficient amount of insulin to control blood glucose.

Manuscript published on 28 February 2015.
***Correspondence Author(s)
P. Radha, Ph.d Scholar, Department of Computer Science, Karpagam University, Coimbatore, Asst. Prof., Vellalar College for Women, Erode, India.
Dr. B. Srinivasan, Department of Computer Science, Gobi Arts and Science College, Gobichettipalayam, India.

T2D is the most common type of diabetes [6]. The pathophysiology of CVD in diabetes is complex and not dependent on the effects of hyperglycaemia alone. In Type 2 Diabetes Mellitus (T2DM) a constellation of risk factors contribute to the development of early CVD, including hypertension and dyslipidaemia. Many people with T2DM are also hypertensive which contributes to the premature development of vascular disease. Diabetes is associated with a typical dyslipidaemia comprising mildly elevated levels of small dense low-density lipoprotein (LDL), reduced levels and altered composition of high-density lipoprotein (HDL) and increased triglyceride-rich lipoprotein particles. Glycated, small dense LDL is associated with increased oxidative stress within the vasculature, while reduced concentrations of altered HDL are less able to participate in atheroprotective functions such as reverse cholesterol transport. Thus, early identification of insulin resistance and impaired endothelial function may identify those at particular risk of CVD and enable targeting of aggressive risk factor control to those who will most benefit. Among these stages analysis of type 2 diabetes with CVD risk factors, important features in the dataset are not selected ,irrelevant data in the T2D patients records are also removed so it degrades the performance of the T2D patients prediction results. The aim of this study was to analyze CVD risk factors in type 2 diabetic patients. The major important steps of the proposed works as follows: Preprocessing of the data using dimensionality reduction KLD -PCA method it is also used for dimensionality reduction to reduce the complexity of the dataset. Once the dimensionality is reduced in the data then risk factors of CVD are analyzed using KDE. The purpose of this study is to propose a prediction methods based on SSL. Before that to reduce the irrelevant or unimportant features in the type 2 diabetes patient records from the KDE similarity measurements results for CVD risk factors ,then build a Hybrid Prediction Model that should perform unsupervised based on IFCM accurately and semisupervised classification methods based on MSOFMNN to classify newly diagnosed patients into type 2 diabetes or not . The proposed hybrid prediction model important features in the T2D patient's records are selected using the ABC-DE algorithm.

II. BACKGROUND STUDY

In modern medicine, large amounts of data are generated, but there is a widening gap between data collection and data comprehension.

It is often impossible to process all of the data available and to make a rational decision on basic trends. Thus, there is a growing pressure for intelligent data analysis such as data mining to facilitate the creation of knowledge to support clinicians in making decisions. Despite this evidence of effectiveness, many countries use a rationing approach to the prescription of cardiovascular risk reduction treatment, with national guidelines suggesting that patients should have their risk of CVD calculated, to ensure therapy is targeted to patients at highest absolute risk. Multivariate risk scores have, therefore, been used to predict CVD risk in individuals with diabetes. Nissen and Wolski [11] have indicated that rosiglitazone, an oral hypoglycaemic drug, increases cardiovascular risk compared to other therapies or placebo, by providing a meta-analysis of treatment trials. Their analysis was at trial-level rather than patient level. There was no standard method for validating outcomes across trials and the total number of events was relatively small. If a classification model was able to identify such risks in a clinical database, it would clearly enhance its efficacy as a practical tool, and add to the evidence. Classification algorithms can be used to compare the effects of feature selection with no feature selection. Three classification methods were used in this research: a probabilistic learner, Naive Bayes [12], a decision tree learner, C4.5 [12]. These algorithms have proved effective in practice [13] and in particular in the clinical domain [14]. Su et al. [15] used four data mining approaches (neural network, decision tree, logistic regression and rough sets) to select the relevant features for the diabetes diagnosis, and also evaluated their performance. Feature selection involves searching through various feature subsets and evaluating each of these subsets using some criterion [16]. In the treatment of diabetes, hundreds of attributes are routinely collected but only a small number is used, i.e. the clinicians routinely perform ad-hoc feature selection. All the features in the database were not specified for blood glucose control prediction and much irrelevant information has been collected. Dalakleidi et al [17] proposed a novel combined use of a genetic algorithm (GA) and a nearest neighbour's classifier for the selection of the critical clinical features which are strongly related with the incidence of fatal and non fatal CVD in patients with T2DM. Huda et al [18] proposed a feature selection; a hybrid of filter by Maximum Relevance(MR) and wrapper by Artificial Neural Net Input Gain Measurement Approximation(ANNIGMA) , MR-ANNIGMA would be used. The combined heuristics in the hybrid MR-ANNIGMA takes the advantages of the complementary properties of the both filter and wrapper heuristics and can find significant features. The selected features set are used to generate a new set of rules for detection of CAN. But all of these feature selection methods the CVD risk factors are not evaluated and data preprocessing is not done this work thus decreases the prediction accuracy of T2D patients with CVD risk factors.

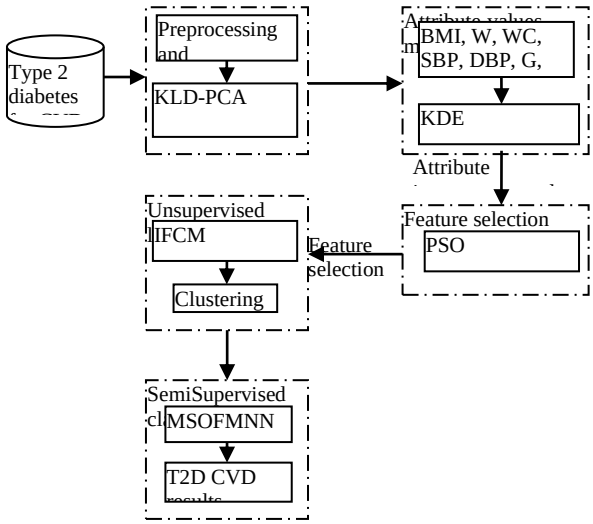

Fig. 1: Architecture of the proposed methodology

KLD-PCA Kullback Leiber Divergence- Principle Component Analysis
KDE Kernal Density Estimation
PSO Particle Swarm Optimization
IFCM Improved Fuzzy Clustering Method
MSOFMNN Modified Self Organizing Feature Map Neural Network

III. PROPOSED METHODOLOGY

Cardiovascular complications are now the leading causes of diabetes-related morbidity and mortality. The public health impact of CVD in patients with diabetes is already enormous and is increasing. In T2D a constellation of risk factors contribute to the development of early CVD, including hypertension and dyslipidaemia. Selection of the important features in the T2D also becomes a difficult task, classification of T2D patients features also becomes major important issue since most of the existing classification or learning methods use a supervised or unsupervised learning ,combination of supervised and unsupervised learning that is semi supervised learning is not supported by anyone of the existing T2D classification methods ,in order to overcome these issues in supervised and unsupervised learning and select most important features in the CVD risk factors. This work majorly focus on the classification task or detection of T2D patients records with CVD risk factors, before that important features of the T2D patient records are selected using ABC-DE. The major objective of this proposed work is to examine the common clinical and behavioral factors that contribute to CVD risk among those with type 2 diabetes and perform hybrid semisupervised learning based on MSOFMNN prediction model. The major steps involved in the proposed system are: preprocessing of the T2D data with CVD risk using KLD-PCA in which the weight values of the PCA are estimated using the Kullback Leiber divergence it is named as KLD-PCA and dimensionality reduction is also performed using KLD-PCA. Then CVD risk factors are estimated based on the KDE, to reduce unimportant features in the data feature selection is performed using hybrid to enhance the prediction accuracy results.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The selected feature with estimated CVD factors are used for unsupervised classification using IFCM clustering methods, which data is used prediction of T2D for CVD risk factors. Then perform semisupervised classification task for prediction of type 2 diabetes patients with CVD risk are predicted using MSOFMNN.

3.1 DATASET INFORMATION

The dataset collected from real patient records which includes the following attributes for diabetes patients records Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test , Diastolic blood pressure (mm Hg) ,Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml) ,Body mass index (weight in kg/(height in m)^2) ,Diabetes pedigree function ,Age (years) ,Class variable (0 or 1).These data are collected with the following CVD risk factors which includes BMI (Body Mass Index) , Weight (kg) ,Waist circumference (cm) , Systolic blood pressure (SBP) (mmHg) , Diastolic blood pressure (DBP) (mmHg) ,Glucose (mg/dl) ,Total cholesterol (mg/dl) , High-Density Lipoprotein cholesterol (HDL-c) (mg/dl) , Low-Density Lipoprotein cholesterol (LDL-c) (mg/dl) ,Triglycerides (mg/dl) ,HbA1c (glycosylated hemoglobin) (%) Fibrinogen (mg/dl), ultrasensitive C reactive protein (us-CRP) (mg/L). If the value of each and every attributes values are changed to analysis the risk factor of CVD for type 2 diabetes (T2D). Managing the numerous risk factors responsible for CVD in T2D represents an ongoing challenge for primary care clinicians, strongly influencing their decisions about treatment approaches for this complex disease.

3.2 PREPROCESSING AND DIMENSIONALITY REDUCTION USING KULLBACK LEIBER DIVERGENCE WITH PRINCIPAL COMPONENT ANALYSIS (KLD-PCA)

The quality of the data is the most important aspect as it influences the quality of the results from the analysis. Data preprocessing in order to improve the quality of the mining result and the efficiency of the mining process [19-20]. KLD-PCA is applied for preprocessing of T2D with CVD risk factors. In PCA which data eigenvector associated with largest Eigen value is the most important vector that reflects the greatest variance for prediction process. From this point of the view the data are preprocessed and removed in the preprocessing stage. A preliminary analysis of the data indicates the usage of zero for missing data. The major problem of the PCA method is that the weight value of the PCA are randomly generated in order to overcome these problem the weight values are estimated based on the KLD . Proposed KLD-PCA for preprocessing of T2D with CVD risk factor .As mentioned above $N = (X_1, X_2, \ldots . X_n)$ is the number of type 2 diabetes patients' hospital data with the CVD risk factors and t dimension of dataset D, respectively. KLD -PCA finds a subspace of the attribute value whose basis vectors correspond to the maximum-variance direction of the original T2D data space. Let W represents the linear transformation that maps the original $t-$dimensional T2D data space into an f-dimensional reduced irrelevant and missing attribute data where normally $f \ll t$. Equation (1) shows the new reduced dimensional and reduced irrelevant data variable vectors $z_j \in R^f$

$$z_j = W^T x_j \, , j = 1, \ldots . N \quad (1)$$

$$\lambda_j e_j = Q e_j \, , j = 1, \ldots . N \, , \quad (2)$$

where $Q = XX^T \, , X = \{x_1, \ldots . x_N\}$

Here Q is the covariance matrix and λ_j the eigenvalue associated with the eigenvector e_j .The eigenvectors are sorted from high to low according to their corresponding eigen values. The eigenvector associated with largest eigen value is the most important variable and data vector that reflects the greatest variance. PCA employs the entire T2D patient hospital record variables with CVD risk factors and it acquires a set of projection attribute vectors to extract most important global variable and data vector from given training samples. The performance of PCA is reduced when there are more irrelevant data ones than the relevant T2D with CVD risk factor ones. In equation (1) the weight transformation matrix is calculated based on the KLD methods in the PCA. Therefore, the weight of attribute k, denoted as $w_{avg(k)}$, is

$$w_{avg}(k) = \sum_{l|k} \frac{\#a_{kl}}{N} KLD(C|a_{kl}) = \sum_{l|k} P(a_{kl}) \, KLD(C|a_{kl}) \quad (3)$$

where $\#a_{kl}$ represents the number of instances that have the value of a_{kl} and the N means the total number of training instances. In this formula, $P(a_{kl})$ means the probability that the attribute k has the value of a_{kl} . The final form of the weight value for attribute is denoted as,

$$w_k = \frac{\sum_{l|k} P(a_{kl}) \sum_c p(C|a_{kl}) \log\left(\frac{p(C|a_{kl})}{p(C)}\right)}{-Z \cdot \sum_{l|k} P(a_{kl}) \log\left(p(a_{kl})\right)} \quad (4)$$

The new variance of k^{th} attribute is calculated as follows:

$$\delta_{newk}(N-1) = \sum_{j=1}^{N} (n_{x_{kj}} - n\vec{x}_k)^2 \quad (5)$$

$$n = \sqrt{\frac{\delta_{newk}(N-1)}{\sum_{j=1}^{N} (x_{jk} - \vec{x}_k)^2}} \quad (6)$$

N is the number of samples and $x_{ji}, \vec{x}_i$ are i^{th} attribute of j^{th} sample and mean of k^{th} attribute respectively. After this adjustment, PCA is employed on data.

3.3 ATTRIBUTE VALUES MEASUREMENT USING KERNEL DENSITY ESTIMATION(KDE)

In this work measure the values of the attributes for prediction of the CVD risk factor in T2D patients based on Kernel Density Estimation (KDE) for prediction of T2D with CVD risks factors. For each and every attribute values select highest value which is greater than the thresholds value. BMI, Weight (kg) ,Waist circumference (cm), SBP (mmHg), DBP (mmHg), Glucose (mg/dl), Total cholesterol (mg/dl), HDL-c (mg/dl), LDL-c (mg/dl),Triglycerides (mg/dl), HbA1c (%), Fibrinogen (mg/dl) and us-CRP (mg/L). The DE algorithm uses mutation operation as initial operation to generate new feature population for T2D samples and selection operation to direct the feature selection operation toward the CVD Risk factor assessment or prediction.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The DE algorithm also uses a non-uniform crossover that can take child feature vector parameters from one feature vector T2D patient's records with CVD risk factors more often than it does from others. By using present feature of T2D patient population members to construct trial vectors, the recombination (crossover) operator efficiently shuffles T2D patient features information about successful combinations. In DE, a population for features of T2D patient's solution vectors is created randomly from ABC algorithm in employee bee phase at the start for all number of T2D patients data with CVD risk factors. Each newly created feature samples of T2D patient with CVD risk factor population feature selection solution is attained based on maximum number of cycles, $C = 1, 2, \ldots, MCN$ for bees. Mutation: Mutation operation expands the search space for searching optimal features of T2D patients with CVD risk factors mutation vector $\hat{F}_i$ obtained by (7):

$$\hat{F}_i = F_{r_1} + F(F_{r_3} - F_{r_2}) \quad (7)$$

where F is the scaling factor having values in the range of [0,1] and solution vectors F_{r_1}, F_{r_2} and F_{r_3} are randomly chosen and must satisfy (8):

$$F_{r_1}, F_{r_2}, F_{r_3} | r_1 \neq r_2 \neq r_3 \neq i \quad (8)$$

Crossover: The parent feature vector of T2D patients with CVD risk factor vector is mixed with the mutated vector to produce a trial vector by (9):

$$y_i^j = \begin{cases} \hat{F}_i^j & R_j \leq CR \\ F_i^j & R_j > CR \end{cases} \quad (9)$$

where CR is crossover constant and R_j is a randomly selected real number between [0,1] and j denotes the j^{th} component of the corresponding array. For initialized T2D patient features from DE the bees are selected based on the highest fitness value from equation (15). Onlooker bees perform the global investigation for discovering new T2D patient feature results and updates global optimum feature selection results. A scout bee discovers the new features selection for T2D patient's features which are not focused by the employed bees. These three steps are continued until a termination criterion is satisfied. The fitness value for each one of the features for T2D patients is associated with fitness function.

$$fit_i = \frac{1}{1 + f_i} \quad (10)$$

Where f_i corresponds to risk factor value from equation (11) . An artificial onlooker bee chooses a T2D patients feature solution depending on the probability value p_i, calculated by the following expression,

$$p_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \quad (11)$$

where fit_i is the fitness value of the feature solution i and SN is the number of feature selection in onlooker bees. In order to produce a candidate food position from the old one in memory, the ABC uses the following expression,

$$v_{ij} = x_{ij} + \theta_{ij}(x_{ij} - x_{kj}) \quad (12)$$

where $k \in \{1, 2, \ldots, SN\}$ and $j \in \{1, 2, \ldots, D\}$ are randomly chosen indexes. ϕ_{ij} is a random number between $[-1, 1]$. If a parameter value produced by this operation exceeds its predetermined limit, the parameter can be set to an acceptable feature selection value. In this work, the value of the parameter exceeding its limit is set to its limit value. The feature selection of which the nectar is abandoned by the bees is replaced with a new food source by the scouts. In ABC-DE, if a position of the current T2D patient features cannot be improved further through a predetermined number of cycles, then that selected feature is assumed to be abandoned. Assume that the abandoned source is AF_i and $j \in \{1, 2, \ldots \ldots D\}$, then the scout discovers a new features source to be replaced with $\mathcal{F}_i$. This operation can be defined as in (13)

$$\mathcal{F}_i^j = \mathcal{F}_{min}^j + rand(0,1)(\mathcal{F}_{max}^j - \mathcal{F}_{min}^j) \quad (13)$$

After each candidate selected features of T2D patient position v_{ij} is produced and then evaluated by the artificial bee, its performance is compared with that of its old one. If the new feature selection results for T2D patient data are equal or better nectar than the old feature selection source, it is replaced with the old one in the memory. Otherwise, the old one is retained in the memory.

Algorithm 1: Artificial Bee Colony –Differential evolution (ABC-DE) optimization

1. Initialize the population of solutions $F_i, i = 1, \ldots \ldots SN$,each population as number of features F_1
2. Evaluate the population
3. Repeat
4. Mutation by equation
5. Recombination by equation
6. Crossover by equation
7. until requirements are met
8. Set cycle = 1
9. Repeat
10. Produce new feature selection solutions v_i for the employed bees (features) and evaluate them best feature
11. Apply the greedy selection process for the employed bees are considered as features
12. Calculate the probability values P_i for the feature solutions F_i
13. Produce the new feature solutions v_i for the onlookers from the solutions X_i selected depending on P_i and evaluate them
14. Apply the greedy selection process for the onlookers are considered as features
15. Determine the abandoned feature solution for the scout, if exists, and replace it with a new randomly produced solution $\mathcal{F}_i^j$ by
16. Memorize the best solution achieved so far
17. cycle = cycle + 1
18. until cycle = MCN

3.4 IMPROVED FUZZY C MEANS CLUSTERING (IFCM)

As the first step, before the application of the Classification algorithms, aim at validating the chosen classes using the unsupervised methods.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

This work uses an IFCM clustering to validate the preprocessed dataset, and then assign class labels to similar cluster, the clustering algorithm. In normal FCM clustering methods distance measure only evaluates the difference between two individual data points. It ignores the global view of the data distribution. However the density of data points in a cluster could be distinctly different from other clusters in a data set. A regulatory factor based on cluster density is proposed to correct the distance measure in the conventional FCM. It differs from other approaches in that the regulator uses both the shape of the data set and the middle result of iteration operation. And the distance measure function is dynamically corrected by the regulatory factor until the objective criterion is achieved. Given a CVD risk factor data for T2D with selected features $FS_r = (fs_1, \ldots . fs_n)$ for every data point fs_i, the dot density is usually defined as:

$$z_i = \sum_{j=1, j\neq i}^{n} \frac{1}{d_{ij}} \; d_{ij} \leq e, 1 \leq i \leq n \quad (14)$$

Where e is the effective radius for density evaluation . Using the cluster density, the distance measure is corrected

$$\hat{d}_{ij}^2 = \frac{\left\| fs_j - v_i \right\|^2}{\hat{z}_i} \; 1 \leq i \leq c, \quad (15)$$

$$1 \leq j \leq n,$$

Thus, the optimization expression can be written as follows :

$$J_{FCM-CD}(U, V, FS) \quad (16)$$

$$= \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m \left\| fs_j - v_i \right\|^2$$

$$\frac{\sum_{k=1}^{n} \alpha_{ik} w_{ik}}{\sum_{k=1}^{n} \alpha_{ik} w_{ik} \, z_k},$$

Applying Lagrange Multiplying Method to obtain the two update equations .

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m} \; 1 \leq i \leq c \quad (17)$$

$$u_{ij} = \frac{\hat{d}_{ij}^{-2/(m-1)}}{\sum_{j=1}^{n} \hat{d}_{kj}^{-2/(m-1)}} \quad (18)$$

3.5 SEMISUPERVISED LEARNING FOR CLASSIFICATION

From this results finally the cluster are formed either class label yes or class label no for classification of type 2 diabetes patients reduced dimensionality data in the KLD-PCA. Finally perform classification task for unsupervised class labels results from IFCM clustering. The clustered results are taken as input to semi supervised learning of MSOFMNN requires a small amount of labeled patterns for selected features results from ABC-DE. The labeled T2D patients with CVD risk factor results have been collected in several ways. In the present technique, for experimental purpose, labeled T2D patient feature samples results from the ground truth for both the classes with equal percentage. After collection of labeled feature selection samples, weight initializations for the labeled and the unlabeled neurons are done differently. If the class label of the selected T2D patients features is known , then the weight vector for the j[th] output neuron, denoted as W_{FSC_i}, is initialized with the normalized feature values of the corresponding labeled pattern; whereas weight vector for others is initialized randomly between[0, 1]. During training, the input feature selected cluster samples FSC_i are passed to the **MSOFMNN** consecutively. Each time, the dot product, $d(FSC_i)$ between FSC_i and W_{FSC_i} is calculated as,

$$d(FSC_i) = FSC_i . W_{FSC_i}, 0 < i < k \quad (19)$$

At the beginning of the training phase, the connection weights of the network are updated in the following manner using the labeled T2D patients with CVD risk factors only. If the class label of clustered features samples of the T2D patient with CVD risk factors is known, then the weight vector of the output neuron is updated using Eq. (20)

$$w_{kl}(itr + 1) = w_{kl}(itr) + h_{kl,i}(itr)\eta(itr)\left(FSC_i - w_{kl}(itr)\right) \quad (20)$$

where $\eta(itr)$ denotes the learning rate in the itr[th] iteration and it decreases with the increase of itr. The weight vectors of the wining neuron and its neighborhood neurons gradually move towards the input clustered T2D features samples under consideration. Learning labeled clustered T2D features selected samples is continued iteratively until convergence O at each iteration 'itr' is calculated as,

$$O = \sum_{d(FSC_i) \geq \theta} d(FSC_i) \quad (21)$$

where θ is a pre-defined threshold value for CVD risk factors of T2D patients. Weight updating is preformed until the difference between output O in two consecutive iterations is less than δ, where $\delta \in [0,1]$ is a small positive quantity. The components of the weight vector$\vec{w}_{kl,k}$ are normalized in the following way so that the dot product $d(FSC_i)$ lies in [0, 1]:

$$W_{FSC_i,k} = \frac{W_{FSC_i,k}}{\sum_{k=1}^{y} W_{FSC_i,k}} \quad (22)$$

After each training step, the unlabeled clustered T2D features patients data are presented to the network and their soft class labels are calculated using the concept of fuzzy set theory. Let us consider that there exist two fuzzy sets: one for the changed CVD risk factor class and the other for the unchanged CVD risk factor class . The membership values of each unlabeled clustered T2D features patients data for CVD risk factor class and the other for the unchanged CVD risk factor class can be determined. For each unlabeled samples $d(FSC_i)$ is computed and $d(FSC_i) \geq \theta$ is more likely to belong to the changed CVD risk factor class than the unchanged CVD risk factor class , otherwise it is from the not under category of CVD risk factor class. Let, $\mu(FCS_i) = [\mu_1(FCS_i)\, \mu_2(FCS_i)]$be the membership value of the FCS_i unlabeled cluster sample , where$\mu_1(FCS_i)$ and $\mu_2(FCS_i)]$are the membership values of the FCS_i unlabeled cluster sample in the CVD risk factor class and the other for the unchanged CVD risk factor class, respectively. These values can be calculated as,

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$$[\mu_1(FCS_i)\ \mu_2(FCS_i)] = \begin{cases} [\min(d(FCS_i), 1 - d(FCS_i)) \\ , \max(d(FCS_i), 1 - d(FCS_i))] \\ [\max(d(FCS_i), 1 - d(FCS_i)) \\ \min(d(FCS_i), 1 - d(FCS_i))] \end{cases} \quad (23)$$

After that, the target CVD risk factors classification results are found then is updated in the same way using K-nearest neighbor technique [23]. For each unlabeled clustered feature samples of T2D patients, its K nearest neighbors is determined. To search for the K number of nearest neighbors, instead of using all clustered feature selected samples, consider only small number of cluster samples within a window around that unlabeled classification samples for CVD risk factors in T2D patients. The target CVD risk factors classification results are estimated as,

$$t(FCS_i) = \left[\frac{\sum_{FCS_{sl} \in M} \mu_1(s,l)}{K}, \frac{\sum_{FCS_{sl} \in M} \mu_2(s,l)}{K}\right] \quad (24)$$

Training of the network and re-estimation of soft class labels of the unlabeled clustered T2D features patient's data using Eqs. (28) and (29) are continued iteratively until the network is stabilized. The sum of square error, ε, after each training step as:

$$\varepsilon = \sum_{i=1}^{p}\sum_{k=1}^{2}(\mu_k(Fcs_i) - t_k(FCS_i))^2 \quad (25)$$

Learning is continued until the difference of error ε between two consecutive training steps is less than ξ (where ε is a small pos-itive quantity). The algorithmic representation of the proposed semi supervised learning methodology is given below:

Algorithm 2: Semi supervised MSOFMNN learning for classification of T2D patient with CVD risk factors

Step 1: Pick up a few labeled samples from the reference map in the cluster

Step 2: Initialize connection weights of the MSOFMNN network.

For the output neuron corresponding to each of the labeled cluster feature samples, initialize weights using the cluster feature samples of the corresponding feature samples.

For the output neuron corresponding to each of the unlabeled cluster feature samples initializes weights randomly in [0, 1].

Step 3: Update the network weight vector for the output neuron corresponding to each of the unlabeled cluster feature samples using labeled cluster feature samples only.

Step 4: Calculate the membership value (μ) of the unlabeled cluster feature samples using similarity measure (d) and the pre-fixed threshold value (θ) by passing through the network.

if $d \geq \theta$,

μ in the changed CVD risk factor class = max[d, (1 − d)].

μ in the unchanged CVD risk factor class = min[d, (1 − d)].

Else

μ in the changed CVD risk factor class = min[d, (1 − d)].

μ in the unchanged CVD risk factor class = max[d, (1 − d)].

Step 5: Assign the target value of each unlabeled cluster feature samples using the membership values of its K nearest neighbors.

Step 6: For the next training step select those unlabeled cluster feature samples for which the estimated target value in changed CVD risk factor class is greater than the unchanged CVD risk factor class one.

Step 7: Update the network weight vector for the output neuron corresponding to each of the unlabeled cluster feature samples using the labeled cluster feature samples as well as the selected unlabeled cluster feature samples.

Step 8: Repeat Steps 4, 5, 6 and 7 until convergence. At convergence, go to Step 9.

Step 9: Assign a hard class label to each of the unlabeled patterns.

IV. EXPERIMENTATION RESULTS

The data were not specifically collected for a research study. As part of routine patient management, UCHT collected diabetic patients' information from 2000 to 2004 in a clinical information system. The data contained physiological and laboratory information for 3857 patients, described by 410 features. The patients included not only type 2 diabetic patients, but also type 1 and other types of diabetes such as gestational diabetes .Some measure of evaluating performance have to be introduced. One common measure in the literature [24] is accuracy defined as correct classified instances divided by the total number of instances. The true positives (TP) and true negatives (TN) are correct classifications. A false positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A false negative (FN) occurs when the outcome is incorrectly predicted as no when it is actually yes. In this study the following equation are used to measure the accuracy Eq. (26), specificity Eq. (27), sensitivity Eq. (28)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (26)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (27)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (28)$$

These parameters can be used to measure accuracy, sensitivity and specificity, respectively. Sensitivity is also referred to as the true positive rate that is, the proportion of positive tuples that are correctly identified, while specificity is the true negative rate that is, the proportion of negative tuples that are correctly identified. The results are shown in Table 1 and are found to be better than the accuracies of other classifiers in the related studies for Pima Indian diabetes dataset.

Parameters	K-C4.5	IFCM-SVM	PSO-IFCM-ELM	ABC-DE-IFCM-MSOFMNN
Accuracy	92.3	93.8	94.5	95.68
Sensitivity	89.4	90.49	92.5	93.59
Specificity	60.8	54.7	52.1	45.89

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

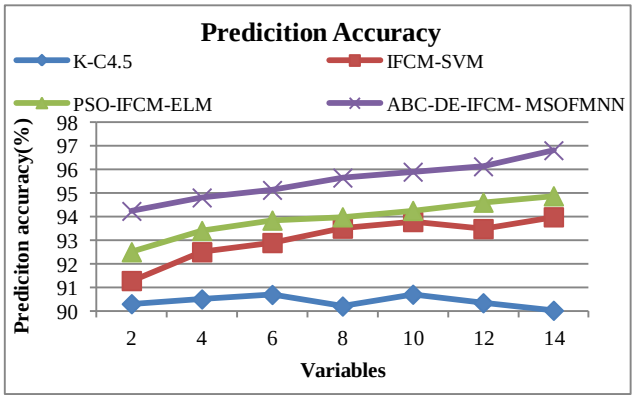

Figure 2: Prediction accuracy of the prediction methods

Prediction accuracy of the proposed ABC-DE-IFCM-MSOFMNN based prediction methods achieves higher classification accuracy than the existing classification methods PSO-IFCM-ELM ,IFCM-SVM ,K-C4.5 prediction accuracy is illustrated in Figure 2 , since the proposed methods ABC-DE is used to select the important features in the T2D for CVD risk factors after the completion of the preprocessing and dimensionality reduction KLD-PCA .In the proposed work KDE based similarity measurement is used to measure the attributes importance , this work presents a novel unsupervised learning and semi supervised when compare to existing prediction methods.

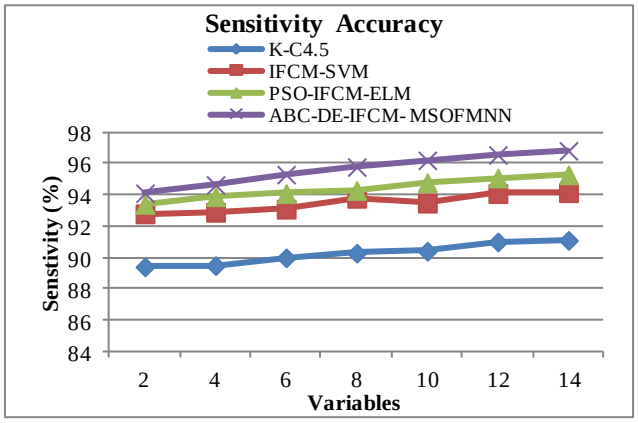

Figure 3: Sensitivity accuracy of the prediction methods

Sensitivity accuracy of the proposed ABC-DE-IFCM-MSOFMNN based prediction methods achieves higher Sensitivity than the existing classification methods such as PSO-IFCM-ELM ,IFCM-SVM and K-C4.5 .Sensitivity is illustrated in Figure 3, Sensitivity result of the proposed ABC-DE-IFCM- MSOFMNN system are high because of the feature selection (ABC-DE) is performed after the completion of the preprocessing and dimensionality reduction methods and KDE based similarity measurement is performed to improve prediction accuracy. Additionally proposed work semisupervised learning is performed using MSOFMNN, thus increases the prediction accuracy.

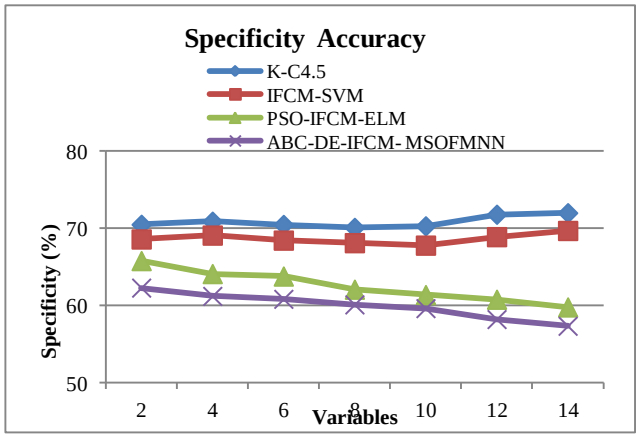

Figure 4: Specificity accuracy of the prediction methods

Specificity accuracy of the proposed ABC-DE-IFCM-MSOFMNN based prediction methods achieves lesser specificity than the existing classification methods such as PSO-IFCM-ELM ,IFCM-SVM and K-C4.5 prediction specificity is illustrated in Figure 4 , the specificity of the proposed ABC-DE-IFCM- MSOFMNN system are less ,since proposed work feature selection (ABC-DE) is performed after the completion of the preprocessing and dimensionality reduction methods and KDE based attribute similarity measurement is performed to improve prediction accuracy, additionally semisupervised MSOFMNN learning based prediction methods is performed in the proposed work .

V. CONCLUSION AND FUTURE WORK

Cardiovascular disease (CVD) is a serious preventable complication of diabetes that leads to substantial disease burden, increased health services use, and premature mortality in Type 2 diabetes. High-quality diabetes care requires first identifying patients at high risk of cardiovascular complications, and then targeting modifiable factors substantially associated with CVD risk. In type diabetes with CVD risk factors finding the most important features becomes major difficult task and thus reduces the prediction accuracy .In order to overcome these problem this work propose a novel semisupervised MSOFMNN classifiers for prediction of CVD risk factors in T2D patients . Before that preprocessing and dimensionality reduction is done using KLD-PCA, assessment of the impact of CVD risk factors based on KDE methods ,features were selected using ABC-DE ,then IFCM clustering algorithm is proposed for unsupervised learning ,finally MSOFMNN is discussed to perform prediction of T2D with CVD risk factors. The network is initially performed using only a few labeled T2D patients. Thereafter, the membership values, in both the classes, for each unlabeled T2D patients data samples are determined using the concept of fuzzy set theory. The soft class label for each of the unlabeled T2D feature selected patients data is then estimated using the membership values of its K nearest neighbors. The experimental results are also statistically validated using classification parameters. The proposed method produced promising results when compare to existing methods.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The complex relationships in diabetes among medical care costs, patient characteristics, and comorbidity are not fully understood. The present study begins to disentangle these relationships, and it provides new information about the large and unique impact of CVD on medical care costs in diabetes. Further research should explore how CVD differs between subjects diagnosed with diabetes at relatively young versus relatively old ages.

REFERENCES

1. Sarwar N, Gao P, Seshasai SR, et al, "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies", Emerging Risk Factors Collaboration, vol.375,no.9733, pp.2215-22,2010.
2. Rydén, L, Standl, E, Bartnik, M, Van den Berghe, G, Betteridge, J, De Boer, MJ & Wood, D. (2007). Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology (ESC); European Association for the Study of Diabetes (EASD). Guidelines on diabetes, pre-diabetes, and cardiovascular diseases: executive summary: The Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology (ESC) and of the European Association for the Study of Diabetes (EASD). European Heart Journal, vol.28, no.1, pp. 88-136.
3. National Collaborating Centre for Chronic Conditions. Type 2 Diabetes National Clinical Guideline for Management in Primary and Secondary Care (update), 2011.
4. Chamnan P, Simmons RK, Sharp SJ, et al, "Cardiovascular risk assessment scores for people with diabetes: a systematic review", Diabetologia vol. 52, no.10, pp.2001-14, 2009.
5. Pellegrini E, Maurantonio M, Giannico IM, et al. "Risk for cardiovascular events in an Italian population of patients with type 2 diabetes", Nutrition, Metabolism and Cardiovascular Diseases,vol.21,no.11, pp.885-92,2011.
6. Acharya, UR, Tan, PH, Subramaniam, T, Tamura, T, Chua, KC, Goh, SC, et al, "Automated identification of diabetic type 2 subjects with and without neuropathy using wavelet transform on pedobarograph", Journal of Medical Systems, vol.32,no.1,pp. 21–29,2008.
7. Chapelle ,O, Scholkopf, B & Zien. A, "Semi-Supervised Learning", vol. 2. Cambridge, MA, USA: MIT press, 2006.
8. X. Zhu, "Semi-supervised learning literature survey," Department of Computer Science University of Wisconsin, Madison, WI, USA, Technical Report 1530, 2007.
9. Gaede P, Vedel P, Larsen N, Jensen GV, Parving HH, Pedersen O , "Multifactorial intervention and cardiovascular disease in patients with type 2 diabetes", New England Journal of Medicine, ,vol.348, no.5, pp.383–393,2003.
10. Gaede P, Lund-Andersen H, Parving HH, Pedersen O , " Effect of a multifactorial intervention on mortality in type 2 diabetes", New England Journal of Medicine, Med ,vol.358,no.6 ,pp.580–591,2008 .
11. Nissen, SE, & Wolski, K , " Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes", New England Journal of Medicine, vol.356,no.24, pp.2457-2471,2007.
12. George Dimitoglou, Adams, JA, Carol MJ, "Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability", Journal of Computing, vol. 4, no.8, 2012.
13. Rish I, Hellerstein, J, Thathachar, J, " An analysis of data characteristics that affect Naive Bayes performance", New York. IBM Technical Report; 2002. http://www.research.ibm.com/ PM/icml01.pdf (2007).
14. Hall M, "Correlation-based feature selection for machine learning", PhD thesis. Hamilton, New Zealand: Department of Computer Science, University of Waikato; 1999. http:// www.cs.waikato.ac.nz/_mhall/thesis.pdf (2007).
15. Su CT, Yang CH, Hsu KH, Chiu WK, "Data mining for the diagnosis for type II diabetes from three-dimensional body surface anthropometrical scanning data", Computers & Mathematics with Applications ,vol.51,no.1, pp.1075—92, 2006.
16. Yu, L, & Liu, H (2003), " Feature selection for high-dimensional data: A fast correlation-based filter solution", In Proceedings of the 20th international conference on machine learning, pp. 856–863.
17. Dalakleidi, KV, Zarkogianni, K, Karamanos, VG , Thanopoulou, AC & Nikita, KS, "A hybrid genetic algorithm for the selection of the critical features for risk prediction of cardiovascular complications in Type 2 Diabetes patients", IEEE 13th International Conference on Bioinformatics and Bioengineering (BIBE), pp.1 – 4,2013.
18. Huda, Shamsul, et al, "Exploring novel features and decision rules to identify cardiovascular autonomic neuropathy using a hybrid of wrapper-filter based feature selection" , Sixth IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), pp. 297-302, 2010.
19. Myatt, GJ, "Making sense of data a practical guide to exploratory data analysis and data mining", New Jersey: John Wiley & Sons, 2007.
20. Han, J & Kamber, M, "Data mining: Concepts and techniques", 2nd edition, Morgan Kaufmann Publishers, 2006.
21. Suguna, N & Thanushkodi, KG, "An Independent Rough Set Approach Hybrid with Artificial Bee Colony Algorithm for Dimensionality Reduction", American Journal of Applied Sciences, vol.8, no.3, pp. 261 – 266, 2011.
22. Bao, L & Zeng, Jc, "Comparison and Analysis of the Selection Mechanism in the Artificial Bee Colony Algorithm", Proceedings of IEEE 9th International Conference on Hybrid Intelligent Systems, pp.411-416, 2009.
23. Patra, S, Ghosh, S , Ghosh,A , " Change detection of remote sensing images with semi-supervised multilayer perceptron" , Fundamenta Informaticae ,vol.84 ,pp.429–442,2008.
24. Chawla, NV, Bowyer, KW, Hall, LO & Kegelmeyer, WP , "SMOTE: Synthetic minority over-sampling technique", Journal of Artificial Intelligence Research (JAIR), vol.16, pp.321–357,2002.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)