Diabetes Disease Prediction Using Artificial Intelligence

Muntather Ayad
Department of Computer and Communication Engineering
Islamic University of Lebanon
PO Box 30014, Werdanyeh, Lebanon
(+96171511238)
muntather7778@gmail.com

Hussein Kanaan
Department of Computer and Communication Engineering
Islamic University of Lebanon
PO Box 30014, Werdanyeh, Lebanon (+96176082259)
dr.kanaan1984@gmail.com

Mohammad Ayache
Department of Biomedical Engineering
Islamic University of Lebanon
PO Box 30014, Werdanyeh, Lebanon (+9613375703)
Mohammad.Ayache@iul.edu.lb

***Abstract*— for a long time, the major problem area for researchers is disease diagnosis and the main interest of the medicine is an accurate diagnosis. Many engineering techniques have been developed in the past to help the medical staff with a diagnosis tool. There are many traditional methods of disease diagnosis, but the application of machine learning techniques has given a new dimension to this area. In this work, two different approaches have been used for the purpose of classification between diabetic and non-diabetic, using Pima Indian Diabetes Dataset. Principal Component Analysis has been used in the purpose of feature dimension reduction before applying any proposed classifier. Support Vector Machine (SVM) and Naïve Bayes (NB) are the two classifiers used in our study. 94.14 % and 93.88% are the accuracies obtained for the SVM and NB approaches.The results obtained are very interesting and show improvement from the previous works. With this accurate learning technique, there is enough scope for improvement considerably in this field.**

Keywords—Principal Component Analysis (PCA), Naïve Bayes (NB), Support Vector Machine (SVM), Pima Indian Diabetes Dataset.

I. INTRODUCTION

Artificial intelligence (AI) greatly assists human society, weather forecasting, facial recognition, fraud detection, and genomic deciphering. AI analytics promote the predictive medicine application, particularly in the challenging setting of chronic diseases marked by multi-organ intervention, abrupt irregular incidents, and latencies with long development of the disease. Learn how to identify non-decipherable patterns utilizing biostatistics by manipulating massive data sets (big data) thru layered mathematical models (algorithms) [1].

Diabetes is a long-lasting disease and import permanent damaging to the limbs and vital organs in the body. Utilizing artificial intelligence tools could enhance the detection techniques and disease control that will be of great help to the physicians [2].

Machine learning technologies (MLT) were utilized at an earlier stage of safe human life to simulate the medical datasets. Through various data sources which utilized to be in the real-world application, massive medical databases are available [3].

The aim of this paper is to classify between diabetic and non-diabetic using Pima Indian Diabetes Dataset. 768 patients with 8 selected features have been used in the purpose of classification.

The proposed paper is organized as follows. Section 2 explains the literature review and the related works. The material and methods, and the proposed algorithm for classification are illustrated in section 3. Section 4 has been used for obtained results and discussions. We conclude in section 5.

II. LITERATURE REVIEW

Authors in [4] focus on selecting the attributes in early detection of Diabetes Miletus using predictive analysis and design a prediction algorithm using Machine learning techniques. The dataset is collected from UCI machine repository. 15 attributes have been used for the purpose of classification. Support Vector Machine, Random forest and Naïve Bayes are the classifiers used with an accuracy of 77.73 %, 75.39% and 73.48%.

Authors in [5] are being placed to design a system that aids in disease estimation including diabetes using Indians Pima Diabetes Selected-database (PIDD). In this research, three identification algorithms of machine learning, including Bayes-Naive, SVM, and Decision Tree, are utilized to diagnose diabetes at an earlier step with an accuracy of 76.3%, 65.1 % and 73.82%.

Authors in [6] proposed k – means approach in the purpose of removing the noisy data and genetic algorithms for finding the optimal set of features with Support Vector Machine (SVM) as classifier. The proposed model has attained an average accuracy of 98.79 % for reduced dataset of Pima Indians Diabetes from UCI repository.

978-1-7281-8855-3/20/$31.00 ©2020 IEEE

Authors in [7] used other identification techniques that are able to categorize diabetes, including Bayes-Naive (NB), quadratic discriminant analysis (QDA), and linear discriminant analysis (LDA) with an accuracy 81.97%.

Authors in [8] used genetic algorithms to predict diabetes disease applied on the Pima Indian Dataset with an accuracy of 78.26%.

Authors in [9] suggested a system where will be categorized into a multi-Diabetes disease categorize model with a three-hierarchy layer. A combination of SVM and BPNN (back propagation neural network) has been used with an accuracy of 88.04%.

Authors in [10] give a descriptive diagnosis of Pima diabetes disorder. To this end, a multiplayer neural network structure trained by the algorithm Levenberg – Marquardt (LM) and a deterministic neural network structure has been utilized with an accuracy of 82.37%.

Authors in [11] developed relatively good performance models to identify persons with diabetes across three aging ranging in the Canadian inhabitance utilizing bagging adaBoost and J48 decision tree. The selected-data group used in this analysis is derived from the Sentinel Surveillance Network of Canadian Primary Care (CPCSSN) selected-database. An accuracy of 89.07% has been obtained.

Authors in [12], cluster analysis technique is utilized to group the objects depending on their similarity. Studies were placed to compare the various techniques of categorization that were developed so far like Categorizers Naïve Bayes, Decision tree (c4.5), k-Means, SVM and k Nearest Neighbor classifier with 77.8646%, 78.2552%, 77.474% and 77.7344% of accuracy.

Authors in [13] used Learning Vector Quantization (LVQ) to categorize the diabetes selected-data group with Chi-Square for character chosen. Chi-square is employed to recognize the character's essential for the prediction of diabetes. It could also help to sort the characteristics from the most prominent to the least. Using chi-square feature selection and LVQ can be obtained at 80% and 90% data training.

III. MATERIALS AND METHODS

Identification algorithms are commonly utilized in different medical uses. Data classified is a two-phase method where the first stage is the training step in which the identification algorithm constructs identification with the training set of tuples and the second stage is the identification period in which the model is used for identification. Its output is evaluated with the test sample of tuples.

The using of classifier technologies in medical diagnosis is incremental. There is no reason to think that even the most significant factors of treatment are an interpretation of patient evidence and expert decisions. Nevertheless, expert systems and various detection methods for artificial intelligence do support experts a considerable amount. Several of the machines learning research in the diabetes diagnosis area is based on the research of the Indian Pima Diabetes database in the UCI repository. For Diabetes forecasting, two classification methods of Primary Component Evaluation are applied and concluded with better predictive methodology with optimum precision. The strategies employed are described below.

Data Definition

The applied raw dataset contains 768 samples corresponding to 768 female patients. The data set is distributed as follows: 500 female non-diabetic patients and 268 female diabetic patients. The dataset has eight attributes, namely; several times pregnant, plasma glucose concentration a two-hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skinfold thickness (mm), two-Hour serum insulin (mu U/ml), body mass index (weight in kg/(height in m)), diabetes pedigree function and age. The dataset has two classes and the classes are coded as 1 and 0 for diabetes and healthy, respectively [14].

Data Pre-Processing

Data pre-processing is a strategy that removes abnormal values which creates unfeasible patterns of data. Feature extraction and selection creates a subset of features from the raw dataset that helps to analyses data for screening. Important features are assigned more weight than others. The most prominent attribute for the female can be the number of times pregnant and the age in years. The concentration of glucose, Blood Pressure, Skinfold thickness, insulin, BMI and Pedigree function is the other eminent attributes. Missing values are substituted with the median values of all patients in the dataset [15].

Normalization Method

The data normalization is considered as the most important preprocessing step using neural networks. To improve the performance of multilayer neural networks, it is better to normalize the data entry such that will be found in the interval of [0 1]. Each feature and parameter in the original data presented in our study has a different scale. To normalize the data, and use it as inputs of the neural network, scaling or normalization should be realized for each attribute. The eight attributes, are scaled with a range of 0 and 1. There are many types of normalization that are found in the literature. The new values obtained after normalization, follow this equation:

$$New\ value = \frac{current\ value - Min}{Max - Min} \quad (1)$$

Principal Component Analysis (PCA):

Principal Components Analysis (PCA) is the predominant linear dimensionality reduction technique, and it has been widely applied on datasets in all scientific domains. In words, PCA seeks to map or embed data points from a high dimensional space to a low dimensional space while keeping all the relevant linear structure intact. To improve the efficiency and accuracy of data mining task on high dimensional data, the data must be

preprocessed by an efficient dimensionality reduction method. PCA steps used in our proposed method describe as bellows:

Standardization:

The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis. Once the standardization is done, all the variables will be transformed to the same scale. This is a very critical step, since the PCA method is very sensitive to the variance of the initial variables. That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges. Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{value - mean}{standard\ deviation} \quad (1)$$

Covariance Matrix Computation:

The aim of this step is to understand how the variables of the input data (features selected from datasets) set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them. Because sometimes, variables are highly correlated in such a way that they contain redundant information. So, in order to identify these correlations, we compute the covariance matrix. The covariance matrix is a p × p symmetric matrix (where p is the number of dimensions) that has as entries the covariance associated with all possible pairs of the initial variables. The covariance matrix is a 3×3 matrix of this form:

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Compute the Eigenvectors and Eigenvalues of the Covariance Matrix to identify the Principal Components:

Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the principal components of the data. The eigenvectors of the Covariance matrix are actually the directions of the axes where there is the most variance (most information) and that we call Principal Components. And eigenvalues are simply the coefficients attached to eigenvectors, which give the amount of variance carried in each Principal Component.

Feature Vector:

In this step, we choose whether to keep all these components or discard those of lesser significance (of low eigenvalues), and form with the remaining ones a matrix of vectors that we call Feature vector. As a result, the feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep. This makes it the first step towards dimensionality reduction, because if we choose to keep only p eigenvectors (components) out of n, the final data set will have only p dimensions.

In this paper we applied PCA method to extract the best discriminate feature among several features selected form the available datasets before we apply the proposed classification method. Our proposed method was tested with and without applying PCA method. We conclude that applying PCA method improves the accuracy result.

Classifiers

Classification can be thought of as two separate problems - binary classification and multiclass classification. In binary classification, a better-understood task, only two classes are involved, whereas in multiclass classification three or more classes should be distinguished. Presented in the following are some popular classifiers.

Support Vector Machine (SVM)

SVM is one of the standard set of supervised machine learning model employed in classification. Given a two-class training sample, a support vector machine aims to find the best highest-margin separating hyper-plane between the two classes. For better generalization, hyper-plane should not lies closer to the data points belong to the other class. Hyper-plane should be selected, which is far from the data points from each category. The points that lie nearest to the margin of the classifier are the support vectors. The goal of SVM training is to find the weight vector w that maximizes the margin. Employed SVM to process the inputs and extracted the rules utilizing an eclectic approach. This approach was then utilized to predict the diagnosis of diabetes utilizing a questionnaire depending on demographic, historical, and anthropometric measures. .The SVM finds the optimal separating hyper-plane by maximizing the distance between the two decision boundaries. Mathematically, we will maximize the distance between the hyper-plane which is defined by

$$wT\,x + b = -1 \quad (2)$$

And the hyper-plane defined by

$$wT\,x + b = 1 \quad (3)$$

This distance is equal to:

$\frac{2}{\|w\|}$ (4)

This means we want to solve max.

$\frac{2}{\|w\|}$ (5)

Equivalently we want min.

$\frac{\|w\|}{2}$ (6)

The SVM should also correctly classify all x (i), which means

$yi(wt\ xi + b) >= 1$ (7)

The illustration is showing in Figure 1.

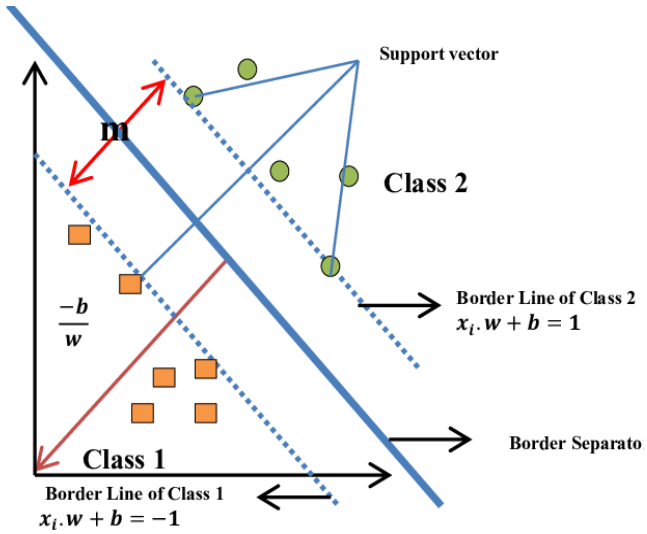

Figure 1: Illustration of SVM Methods.

$xi.w + b \geq 1$ (8)

$xi.w + b \geq -1$ (9)

W is a perpendicular vector to a separate hyper-plane and b is the C Parameter corresponding to the shortest distance from the origin of the coordinates to the hyper-plane. 1 and -1 represent two different output classes. Thus, if the calculation yields 1, then the data will be classified as class A, and if it produces -1, then it will be classified as class B. To find the boundary field or (hyper-plane) to know where the support vector can be utilized the following formula [16].

$\min 1/2 \left| W^2 \right|$ (10)

In this study, SVM classifiers are used to predict Diabetes of patients. SVM is a binary classifier that categorize the input feature vector by evaluating the classifier function f (x) as follows:

$f(x) = sgn(w^T \varphi(x) + b)$ (11)

Where x is the feature vector, b and w, are bias and the vector of SVM coefficients respectively, defines kernel function and sgn denotes the sign function. Since SVM is inherently a binary classifier, one against one approach is used in this study to predict Diabetes.

Naïve Bayes Classifier

Naive Bayes is a classification technique with a notion which defines all features are independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. Since it depends on conditional probability, it is considered as a powerful algorithm employed for classification purpose. It works well for the data with im-balancing problems and missing values. Naive Bayes is a machine learning classifier which employs the Bayes Theorem [17]. Theorem formula calculates the posterior probability for each class utilizing below formula

$p(cj \mid d) = \frac{p(d \mid cj)\, p(cj)}{p(d)}$ (12)

Where,

p (cj | d) = probability of instance d being in class cj.

P (d | cj) = probability of generating instance d given class cj.

p (cj) = probability of occurrence of class cj.

p (d) = probability of instanced occurring.

$p(d|cj) = p(d1|cj) \times p(d2|cj) \times \ldots\ldots \times p(dn|cj)$ (13)

Where

p (d | cj) = the probability of class cj generating instanced.

p (d1 | cj) = the probability of class cj generating the observed value for feature 1.

p (d2 | cj) = the probability of class cj generating the observed value for feature 2.

p (dn | cj) = the probability of class cj generating the observed value for feature n.

The Flowchart for Naïve Bayesian is shown in Figure 2.

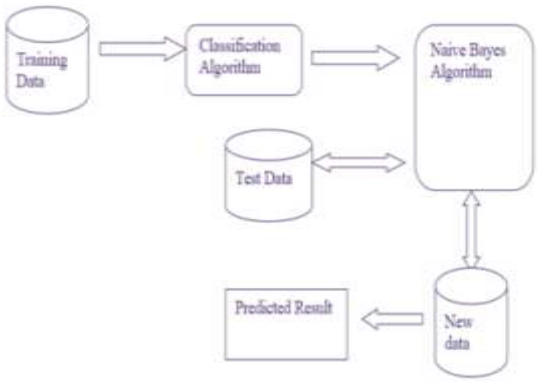

Figure 2: Flow chart for the Naïve Bayesian classification.

IV. RESULTS AND DISCUSSIONS

By normalization process for the raw inputs has a great effect on preparing the data to be suitable for the training. Next attribute reduction, identify and extract the important patterns in the dataset by the principal component analysis (PCA). Finally, step two classifier techniques, support vector machine (SVM) and naïve Bayes with Principal component analysis are implemented for the forecasting of diabetes and concluded with best forecasting techniques which have maximum accuracy.

Table 4-1: all the outcomes of error rate gained depending on naïve Bayes with PCA

Classifier method	Training data %	Testing data %	Error rate % Without PCA	Error rate % With PCA	Accuracy without PCA	Accuracy With PCA
Naïve Bayes (NB)	50%	50%	15.76	15.36	84.24	84.64
	75%	25%	8.59	7.68	91.41	92.32
	80%	20%	6.64	5.86	93.36	94.14
Support Vector Machine (SVM)	50%	50%	16.02	15.23	83.98	84.77
	75%	25%	9.11	7.42	90.89	92.58
	80%	20%	7.16	6.12	92.84	93.88

The results mentioned in the table above shows that the applying of PCA method before the classifier improves the accuracy rate. This improvement is due to the extraction and selection of best feature using PCA. According to SVM approach used in our proposal , we conclude that the two classes diabetes and non-diabetes are linearly discriminated

I. CONCLUSION

The main spotlight of this thesis is to combination Pre-processing technique undertaken is Principal Component Analysis to remove the abnormal raw data and also the reduction of an attribute. The classifiers utilized for prediction of diabetes are Support Vector Machine (SVM), and Naive Bayes Classifier (NB). For future work, the same method could be considered, and many other machine learning classifiers algorithms could be considered to compare the most accurate one. This method can also be implemented on various other disease and medical datasets.

REFERENCES

[1] Miller, D.D. and Brown, E.W., 2018. Artificial intelligence in medical practice: the question to the answer. The American journal of medicine, 131(2), pp.129-133.

[2] El_Jerjawi, N.S. and Abu-Naser, S.S., 2018. Diabetes prediction using artificial neural network.

[3] Contreras, I. and Vehi, J., 2018. Artificial intelligence for diabetes management and decision support: a literature review. Journal of medical Internet research, 20(5), p.e10775.

[4] Sneha, N. and Gangil, T., 2019. Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data, 6(1), p.13.

[5] Sisodia, D. and Sisodia, DS, 2018. Prediction of diabetes using classification algorithms. Procedia computer science, 132, pp.1578-1585.

[6] Santhanam, T. and Padmavathi, M.S., 2015. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for a diabetes diagnosis. Procedia Computer Science, 47, pp.76-83.

[7] Maniruzzaman, M., Kumar, N., Abedin, M.M., Islam, M.S., Suri, H.S., El-Baz, A.S. and Suri, J.S., 2017. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Computer methods and programs in biomedicine, 152, pp.23-34.

[8] Fatima, M. and Pasha, M., 2017. Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 9(01), p.1.

[9] Zolfaghari, R., 2012. Diagnosis of diabetes in the female population of Pima Indian heritage with an ensemble of bp neural network and SVM. Int. J. Comput. Eng. Manag, 15, pp.2230-7893.

[10] Temurtas, H., Yumusak, N. and Temurtas, F., 2009. A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with Applications, 36(4), pp.8610-8615.

[11] Perveen, S., Shahbaz, M., Guergachi, A. and Keshavjee, K., 2016. Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, 82, pp.115-121.

[12] Thirumal, P.C. and Nagarajan, N., 2015. Utilization of data mining techniques for the diagnosis of diabetes mellitus-a case study. ARPN Journal of Engineering and Applied Science, 10(1), pp.8-13.

[13] Putri, NK, Rustam, Z. and Sarwinda, D., 2019, June. Learning Vector Quantization for Diabetes Data Classification with Chi-Square Feature Selection. In IOP Conference Series: Materials Science and Engineering (Vol. 546, No. 5, p. 052059). IOP Publishing.

[14] Putri, NK, Rustam, Z. and Sarwinda, D., 2019, June. Learning Vector Quantization for Diabetes Data Classification with Chi-Square Feature Selection. In IOP Conference Series: Materials Science and Engineering (Vol. 546, No. 5, p. 052059). IOP Publishing.

[15] Vijayan, VV and Anjali, C., 2015, April. Decision support systems for predicting diabetes mellitus—A review. In 2015 Global Conference on Communication Technologies (GCCT) (pp. 98-103). IEEE.

[16] Kumari, VA and Chitra, R., 2013. Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), pp.1797-1801.

[17] Iyer, A., Jeyalatha, S. and Sumbaly, R., 2015. Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv: 1502.03774.