Comparative Analysis of Classification Techniques for Diagnosis of Diabetes

Paramjot Kaur and Ramanpreet Kaur

Abstract Diabetes is a disease with which many people are affected, and diagnosing diabetes is becoming an important task. Machine learning algorithms are widely used for detection and classification process. In this work, we have used five classifiers to diagnose disease. The dataset, Pima Indian diabetes database, used to validate our work is taken from an online repository. We evaluated different machine learning algorithms for their accuracy. The classification accuracy was comparable to the state-of-the-art ranging from 70.12 to 79.22%. In this work, we suggested that the Naïve Bayes algorithm is an optimal algorithm, which is good in terms of accuracy as well as running time complexity.

Keywords Machine learning · Classifier · Confusion matrix · ANN · SVM · Diabetes

1 Introduction

Diabetes is a chronic condition when the body does not produce the accurate amount of insulin. Diabetes increases the risk of many diseases, namely heart disease, kidney malfunctioning, nerve system damaging, and so on. For diagnosing a diabetic patient, an expert relies on his past experiences. Several studies are carried out to diagnose diabetes based on various parameters. There are four kinds of diabetes: Type1, Type2, gestational diabetes, and pregestational diabetes.

Several machine learning, neural network, and deep learning techniques have been applied in various fields, including medical data. These algorithms include linear discriminant analysis, support vector machine, K-nearest neighbor, multilayer neural network, and so on. Dogantekin et al. proposed a diagnosis system based on linear discriminant analysis (LDA) and adaptive network-based fuzzy system

P. Kaur (✉)
Shaheed Darshan Singh Pheruman Memorial College for Women, Rayya, Amritsar 143112, India
e-mail: Jotchandi1@gmail.com

R. Kaur
Chandigarh University, Mohali 140413, India

L. C. Jain et al. (eds.), *Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals*, Advances in Intelligent Systems and Computing 1064,
https://doi.org/10.1007/978-981-15-0339-9_17

for diabetic data classification [1]. Another approach focused on diabetes diagnosis with the use of generalized discriminant analysis (GDA) and least square support method (LS-SVM) [2]. Sean N. Ghazavi and Thunshun W. Liao obtained good classification accuracy using fuzzy K-nearest neighbor algorithm and the adaptive network-based fuzzy inference system [3]. The GDA-LS-SVM technique had an accuracy of 82.05%. Mostafa Fathi and Mohammed Sainee Abadeh [4] made use of ant colony optimization for extraction of fuzzy rules for building a fuzzy classification system for diagnosis of diabetes disease. Kayaer and Yildirim [5] used the LM algorithm for the neural networks on a Pima Indian dataset and achieved an accuracy of 77.08%. Okan et al. investigated two different feed-forward artificial neural networks (FFANN) for diagnosing diabetes and concluded that Newman–Watts small world feed-forward artificial neural network (SW-FFANN) gives better results as compared to Watts–Strogatz SW-FFANN [6]. Izhan addressed class imbalance in various medical databases on clinical predictions using the method of bagging neural network [7]. One method used genetic algorithm as a feature selection algorithm and Naïve Bayes for classification purpose [8]. Another work explored popular machine learning algorithms for diabetes identification and validated the model against 10-fold validation [9]. One approach investigated multilayer perceptron learning on Pima Indian diabetes database and concluded that the algorithm is giving optimal results [10]. A recent work used multilayer perceptron classifier with R Studio platform for classification of diabetes data, and this work was also verified with MATLAB [11]. Particle swarm optimization for reducing number of attributes of data followed by support vector machine and fuzzy decision tree on Pima Indian diabetes dataset improved the accuracy for diabetes classification [12].

This work focuses on classifying UCI diabetes data with given attributes using five different learning algorithms and getting comparable accuracy as of other methods. We find that Naïve Bayes classifier outperforms all other classifiers with 79.22% accuracy. Further, this paper contains the proposed system, dataset description, classification method, results, and conclusion.

2 Proposed System

In this work, we trained various machine learning algorithms and validated their results on the test set. We took a benchmark dataset with eight predictors and one binary outcome variable. Then this dataset is partitioned randomly into 90% training set and 10% testing set. Then various classification models were trained and computed performance metrics in terms of accuracy.

A typical machine learning process is depicted in Fig. 1. In machine learning classification process, at first, classifier gets trained on certain training data samples and then it is validated against test data samples.

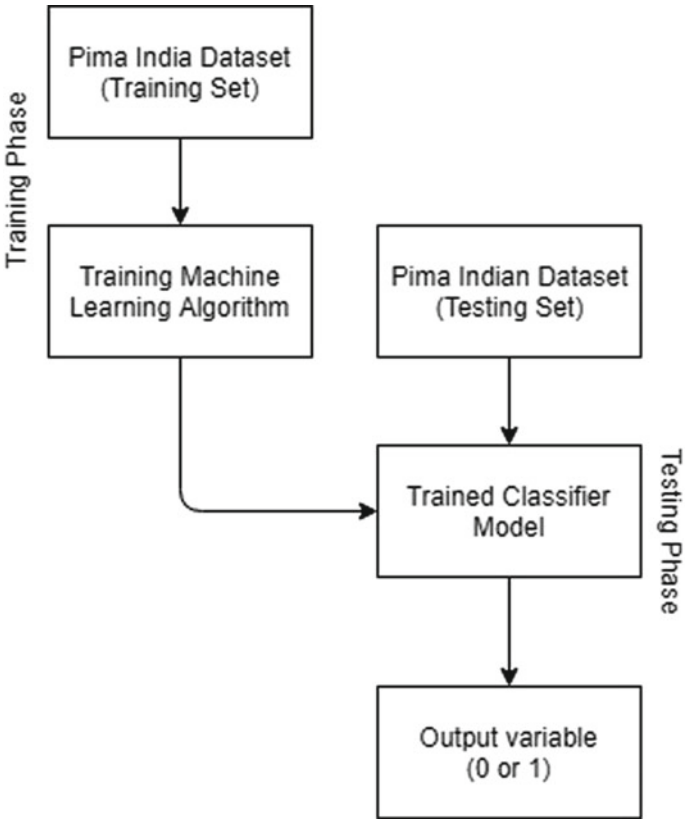

Fig. 1 Machine learning training and testing phase for Pima Indian dataset

2.1 Dataset Description

We chose a benchmark dataset, Pima Indian diabetes database, for diagnosing diabetes from Kaggle [13], which is an online machine learning repository. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset consists of 769 entries of females at least 21 years old.

There are eight predictors in the database and one binary outcome class to be predicted. These are described as follows:

- Number of time pregnant: Number of times pregnant
- Glucose concentration: Plasma glucose concentration 2 h in an oral glucose tolerance test
- Diastolic blood pressure measured in mm Hg
- Skin thickness: Triceps skin fold thickness measured in mm
- Insulin: 2 h serum insulin measured in mu U/ml

- Body mass index: Body mass index (weight in kg/(height in m)^2) measured in kg m^{-2}
- Diabetes pedigree function: Diabetes pedigree function
- Age: Age in years
- Outcome: Binary class variable (0 or 1).

2.2 Classification Techniques

A. *Artificial Neural Network* (*ANN*): Artificial neural network consists of multiple processing layers to represent data with multiple layers of abstraction [14].
B. *Decision Tree*: Decision tree algorithms are commonly based on a divide and conquer technique. The main motive of these algorithms is to divide the training set T into multiple subsets so that each subset belongs to the single target set. Decision trees are constructed by partitioning the numerical attributes. Researchers from various fields create a decision tree based on observational data [15].
C. *Random Forest*: Random forest algorithm is used for classification and regression. Random forests are a combination of decision trees and values of a random vector of each tree. The random forest is sampled independently [16]. It produces two types of information predictor variable and internal structure of data.
D. *Support Vector Machine* (*SVM*): Support vector machines are based on maximum margin principle. They construct a hyperplane to separate different categories to be classified [17].
E. *Naïve Bayes Classifier*: Naïve Bayes is a simple probabilistic model which can distinguish different objects or classes based on Bayes' theorem [18]. Naïve Bayes is a practical classifier with advanced applications like a cancer diagnosis, face recognition, and diabetes diagnosis.

3 Classifier Results

In this work, we used classifiers explained in Sect. 2.2 to conclude the results. These classifiers allow the user to validate various models for data consisting of feature space. The benchmark dataset taken from online repository was fed to respective machine learning algorithm. Then classifiers were trained on training set in various iterations until convergence is reached and give validated results in terms of accuracy on test set. The minimum accuracy achieved is 70.12% using artificial neural network and maximum accuracy achieved is 79.22% using Naïve Bayes classifier. From these classifier results, we can observe that the Naïve Bayes classifier gives the most satisfactory results in terms of accuracy.

In Table 1, average recall for each model is given. In medical diagnosis, recall is an important criterion which tells a fraction of relevant samples that have been

Table 1 Average recall corresponding to each model used for classification

Models	Average recall (%)
Artificial neural network (ANN)	70.00
Decision tree	77.00
Random forest	75.00
Support vector machine (SVM)	78.00
Naïve Bayes	79.00

correctly identified from total samples. Deep learning provides us least, while Naïve Bayes achieves most recall value.

In Table 2, f1-score is provided against each model used. f1-score is another parameter which measures classifiers accuracy. Again, the Naïve Bayes classifier is leading among other classifiers used in this work.

Figure 2 comprises an overall comparison corresponding to each classifier used in this work, which is showing that the Naïve Bayes algorithm is the best algorithm for future predictions.

The most important metric for classifier's performance is the confusion matrix which tells us the number of samples correctly classified and misclassified. The diagonal elements of the confusion matrix are correctly classified samples, while elements other than diagonal elements are misclassified samples. Tables 3, 4, 5, 6 and 7 consist of a confusion matrix corresponding to each classifier used in this work.

Table 2 Average f1-score corresponding to each model used for classification

Models	Average f1-score (%)
Artificial neural network (ANN)	67.00
Decision tree	77.00
Random forest	76.00
Support vector machine (SVM)	78.00
Naïve Bayes	79.00

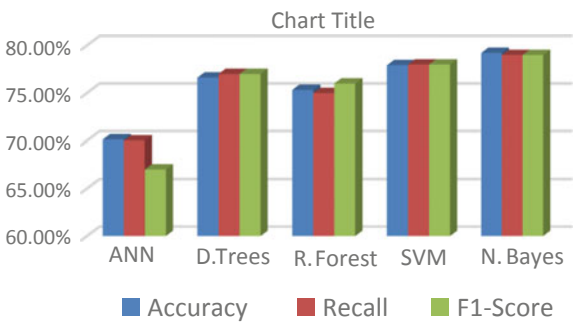

Fig. 2 Bar chart comparison of accuracy, recall, and f1-score of each classifier

Table 3 Confusion matrix for ANN classifier with accuracy 70.12%

	True Class 0	True Class 1
Pred. Class 0	44	3
Pred. Class 1	20	10

Table 4 Confusion matrix for decision tree classifier with accuracy 76.62%

	True Class 0	True Class 1
Pred. Class 0	44	5
Pred. Class 1	18	15

Table 5 Confusion matrix for random forest classifier with accuracy 75.32%

	True Class 0	True Class 1
Pred. Class 0	48	11
Pred. Class 1	8	10

Table 6 Confusion matrix for SVM classifier with accuracy 77.92%

	True Class 0	True Class 1
Pred. Class 0	40	8
Pred. Class 1	9	20

Table 7 Confusion matrix for Naïve Bayes classifier with accuracy 79.22%

	True Class 0	True Class 1
Pred. Class 0	41	5
Pred. Class 1	11	20

4 Conclusions

This work comprises various machine learning algorithm trained and tested on Pima Indian dataset taken from the online repository and observing various performance metrics. In classification accuracy, Naïve Bayes classifier performs best for classifying diabetes data. Naïve Bayes achieves 79.22% accuracy whereas artificial neural network achieves least accuracy of 70.12%. If all parameters are considered, then the Naïve Bayes algorithm is performing the best. So, based on these parameters we can easily explore the best training algorithm required for diagnosing diabetes.

References

1. E. Dogantekin, A. Dogantekin, D. Avci, L. Avci, An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS. Digit. Signal Process. (2010)

2. K. Polat, S. Güneş, A. Arslan, A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine. Expert Syst. Appl. (2008)
3. S.N. Ghazavi, T.W. Liao, Medical data mining by fuzzy modeling with selected features. Artif. Intell. Med. (2008)
4. M.F. Ganji, M.S. Abadeh, A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis. Expert Syst. Appl. (2011)
5. K. Kayaer, T. Yildirim, Medical diagnosis on pima indian diabetes using general regression neural networks, in *International Conference on Artificial Neural Networks and Neural Information Processing* (2003)
6. O. Erkaymaz, M. Ozer, M. Perc, Performance of small-world feedforward neural networks for the diagnosis of diabetes Appl. Math. Comput. (2017)
7. I. Fakhruzi, An artificial neural network with bagging to address imbalance datasets on clinical prediction, in *2018 International Conference on Information and Communications Technology, ICOIACT 2018* (2018)
8. D. Choubey, S. Paul, S. Kumar, S. Kumar, Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection. Commun. Comput. Syst. (2016)
9. S. Wei, X. Zhao, C. Miao, A comprehensive exploration to the machine learning techniques for diabetes identification, in *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, (2018), pp. 291–295
10. S.A. Saji, K. Balachandran, Performance analysis of training algorithms of multilayer perceptrons in diabetes prediction, in *Conference Proceeding—2015 International Conference on Advances in Computer Engineering and Applications, ICACEA 2015* (2015), pp. 201–206
11. S.K. Mohapatra, J.K. Swain, M.N. Mohanty, Detection of diabetes using multilayer perceptron, in *International Conference on Intelligent Computing and Applications* (2019), pp. 109–116
12. D.K. Choubey, S. Paul, K. Bala, M. Kumar, U.P. Singh, Implementation of a hybrid classification method for diabetes, in *Intelligent Innovations in Multimedia Data Engineering and Management*. IGI Global (2019), pp. 201–240
13. https://www.kaggle.com/uciml/pima-indians-diabetes.
14. J.S. Tiruan, Artificial neural network. Neuron (2011)
15. L. Rokach, O. Maimon, Decision tree, in *Data Mining Knowledge Discovery Handbook* (2005), pp. 165–192
16. University of California, L. Breiman, Random forest. Mach. Learn. **45**(5), 1–35 (1999)
17. C. Cortes, V. Vapnik, Support vector machine. Mach. Learn. (1995)
18. Naive Bayes classifier The naive Bayes probabilistic model