ORIGINAL ARTICLE e-ISSN: 2349-0659 p-ISSN; 2350-0964

Hybrid Type-2 Diabetes Prediction Model Using SMOTE, K-means Clustering, PCA, and Logistic Regression

Atul Kumar Ramotra, Vibhakar Mansotra

Abstract

Early prediction of diabetes is very important as diabetes can turn out to be life threatening for the patients in the later stages. In this paper, a hybrid framework for the prediction of type-2 diabetes is developed. In the first step, imbalance dataset is balanced using Synthetic Minority Over-sampling Technique. Then, clustering is applied using k-means clustering technique and all the incorrectly clustered entries and outliers are removed. Principal component analysis is then used for dimensionality reduction of the dataset. In the final step, classification is done using logistic regression (LR), naïve Bayes, support vector machine, and k-nearest neighbors classification techniques. Experimental analysis shows that 98.96% of accuracy is achieved by the proposed hybrid model using LR. The results are validated using 10-fold cross-validation.

Keywords: Classification, Clustering, Data mining, Diabetes prediction, Principal component analysis, Synthetic Minority Over-sampling Technique

Asian Pac. J. Health Sci., (2021); DOI: 10.21276/apjhs.2021.8.3.23

Introduction

Diabetes is a chronic disease which is increasing the global burden of health problems year by year. Diabetes is categorized into three types: Type-1diabetes, type-2 diabetes, and gestational diabetes. Type-1 diabetes is caused due to lower amount of insulin or no insulin production in the body. The cause of type-2 diabetes is the incapability of the body to use the insulin produced. Gestational diabetes is caused when the amount of blood glucose level is increased during pregnancy. Among all the three types of diabetes, type-2 diabetes shares the highest 90% of the cases.[1]

As per the report of International Diabetes Federation of the year 2019, 463 million people under the age of 20–79 years which amounts to 9.3% of the population globally were suffering from diabetes. By 2030, these figures are estimated to increase to 578 million people representing 10.2% of global population and to 700 million people amounting to 10.9% of global population by the year 2045.[2] India with 77 million people suffering from diabetes is at second place after China, in the list of top 10 countries with highest number of diabetic patients in 2019. These numbers are expected to rise in India to 101 million people and 134.2 million people by the year 2030 and 2045, respectively, which is a major health issue.[3] Early stage detection of diabetes is important as people who are suffering are unaware of it in most of the cases.

In medical domain, machine learning techniques are playing an important role in prediction of many chronic diseases by offering many supervised and unsupervised learning techniques.[4,5] Applications of machine learning techniques on the large health-care datasets have been beneficial by reducing the disease diagnostic costs using minimal human resources with increased diagnostic accuracy.[6] However, the accuracy of machine learning techniques decreases due to many challenging problems and one of them is the data imbalance problem.

In this paper, a hybrid framework for prediction of diabetes is developed using class balancing technique, unsupervised learning method, dimensionality reduction method, and supervised learning technique. The class distribution of the dataset is balanced using Synthetic Minority Over-sampling Technique (SMOTE). Unsupervised learning includes the clustering technique which is a process of assigning different groups to the observations according to certain similarities between them. k-means clustering technique is used for the clustering of the dataset and principal component analysis (PCA) is used to reduce the dimensionality of the dataset. Classification is a process of categorizing observations with unknown classes based on training done on subset of observations whose classes are already known. Logistic regression (LR), naïve Bayes (NB), support vector machine (SVM), and k-nearest neighbors (KNN) algorithms are used for the classification at the last step and the algorithm showing the highest accuracy is used for the development of the final framework.

Department of Computer Science and IT, University of Jammu, Jammu, Jammu and Kashmir, India

Corresponding Author: Atul Kumar Ramotra, Department of Computer Science and IT, University of Jammu, Jammu, Jammu and Kashmir, India. E-mail: ramotraatul@gmail.com

How to cite this article: Ramotra AK, Mansotra V. Hybrid Type-2 Diabetes Prediction Model Using SMOTE, K-means Clustering, PCA, and Logistic Regression. Asian Pac. J. Health Sci., 2021;8(3):137-140.

Source of support: Nil

Conflicts of interest: None.

Received: 21/04/2021 **Revised:** 22/05/2021 **Accepted:** 15/6/2021

Methods

Machine learning offers different techniques and depending on the nature of problem domain and the type of dataset, these techniques can be applied for the predictive modeling. In this section, various machine learning techniques employed for the development of the proposed hybrid framework are explained.

Supervised Learning

Supervised learning is the process of generating a function to get a desired output mapping certain inputs. The input to the function is set of samples whose classes are already known (training data)

so that the function can predict the target class of the new samples which are never seen before (testing data). Each sample consists of some attributes known as independent variables and belongs to a class known as dependent variable.[7] Classification algorithms are the most commonly used supervised learning techniques. The classification algorithms used in the study are as follows:

LR

LR algorithm maps the instances to a set of discrete classes. The assignment of the instances is based on the probability value calculated using the sigmoid function which transforms the predictions. The prime focus of classical LR approach is the minimization of error by solving the parameters of loss function. Gradient descent method is generally used for this type of problem.[8]

NB

NB algorithm uses statistical techniques for the classification of the data. Bayes' theorem is used to predict the membership class of the instances. NB algorithm works on the assumptions of class conditional independence which states that the values of all the features of the instances are considered to be conditionally independent of each other. Posterior probability, conditional probability, and the probability of occurrence are calculated.[9] The response variable with the highest value of probability of occurrence is selected.

SVM

SVM algorithm classifies the data on the basis of boundary determination technique. Using non-linear mapping, SVM maps lower dimensional data into higher dimensional data and searches for the decision boundary known as linear optimal hyperplane for separation of the classes. SVM algorithm can process linear as well as non-linear data. The main purpose of the SVM algorithm is to search for a best data boundary separating all the classes with maximum possible distance between them.[9,10]

KNN

KNN classifier also known as memory-based technique approximates the value of unknown data points using its nearest neighbors. The KNN points having the least distance from the unknown data points are considered to estimate the values. The distance is calculated using several techniques. Euclidian distance is the simplest and most generally used function for the distance measurement.[11]

Unsupervised Learning

Unsupervised learning is the process of discovering the hidden associations in dataset variables. Observations of training dataset do not require any labeling but the system itself tries to find the hidden structures. Various types of statistical and clustering methods follow unsupervised learning technique.

k-means clustering

Clustering is the method of division of the dataset into smaller groups known as clusters. The instances belonging to a particular cluster share greater similarities with each other than the instances belonging to other clusters. k-means uses clustering technique and works iteratively. If k is the number of clusters, then k sets of samples are selected and are considered as centroid of the clusters. The distance of each sample having k centroids is then calculated using Euclidean distance. The samples are further added to the clusters having the minimum distance from the centroid. These steps are repeated again and again as the samples keep on shifting to the new clusters on the basis of minimum distance till there is not further shifting in the clusters.[9]

PCA

PCA helps in reducing the dimensionality of the dataset which proves to be useful in decrease the complexity of a machine learning model.[12] PCA uses orthogonal transformation for converting various correlated features into mutually uncorrelated features called as principal components. A decreasing order is followed in terms of variance among the principal components, that is, the first principal component shows the highest variability than the second one and so on further in decreasing order and all principal components are orthogonal to each other also. PCA reduces dimensionality of the dataset by transforming a high dimensional dataset into a low dimensional dataset.[13]

SMOTE

Class imbalance is a common problem found in many health-care datasets where the amount samples belonging to a particular class are present in larger number as compared to the other classes. Hence, the dataset contains unequal distribution of the classes. The results of classification done on a class imbalanced dataset are more likely to be biased toward the majority class.[14] Undersampling and oversampling are two techniques used to balance the class imbalance datasets. SMOTE is an oversampling technique which creates synthetically generated samples of the minority class. Euclidean distance between a minority sample and a randomly selected k-nearest neighbor is calculated, multiplied by any number between 0 and 1 and new samples are synthetically placed across the line joining both of the minority samples.[15]

Hybrid framework

The architecture of the proposed hybrid framework is shown in Figure 1. Initially, data pre-processing is done on the original dataset.

After the data pre-processing, SMOTE is applied on the dataset to balance the imbalanced dataset. The balancing of the dataset is achieved by adding new instances of minority class and making them equal to the majority class instances. The results obtained before and after using SMOTE on the original dataset are compared in Table 1. k-means clustering is then applied on the balanced dataset by setting the value of k = 2. The value of k = 2 divided the dataset into two clusters. Cluster 0 contained the instances of non-diabetic

Table 1: Results attained before and after using SMOTE on original dataset

Type of entry	*Original dataset*	*Using smote on the original dataset*
Total number of instances	711	946
Diabetic class instances	235	470 (2*235)
Non-diabetic class instances	476	476

SMOTE: Synthetic Minority Over-sampling Technique

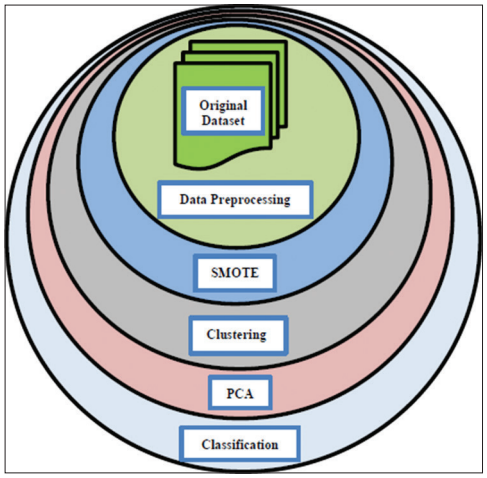

Figure 1: Architecture of the proposed hybrid framework

class, cluster 1 contained the instances of diabetic class, and both the clusters also contained some miss clustered instances. k-means clustering is used for the identification of outliers present the new balanced dataset. The next step involved the identification and removal of all incorrectly clustered instances or outliers from both these clusters. After the removal of incorrectly clustered instances, the entries of both the clusters are joined and shuffled to make a new dataset. PCA is then applied on the new dataset to decrease the dimensionality of the dataset and transforming it into various uncorrelated principal components. In the final step, classification using LR, NB, SVM, and KNN algorithms is done. After the results analysis, the classifier showing the best results has been used for the development of the proposed hybrid framework.

Dataset

The dataset used in the study is Pima Indian Diabetes Dataset collected from UCI Machine Learning Repository contributed by National Institutes of Diabetes and Digestive and Kidney Diseases in the US.[16] The original dataset contains 768 instances with eight input clinical features as: Number of times pregnant, plasma glucose concentration after 2 h in an OGTT, diastolic blood pressure (mmHg), triceps skinfold thickness (mm), 2 h serum insulin (μU/mL), BMI, diabetes pedigree function, and age (years). The output class contains two values 0 or 1.

Results

The original dataset contains 768 instances, 8 input attributes, and 2 output class values. After data pre-processing, 711 instances are considered for the study. To balance the highly imbalanced original dataset, SMOTE algorithm is applied. An increase of 50% to the minority class (diabetic class) instances is done to balance the dataset, as shown in Table 1. After applying k-means clustering and PCA on the balanced dataset, LR, NB, SVM, and KNN classifiers are applied for the prediction of diabetes. Comparison of the results is done on the basis of precision values, recall values, f1 scores, and accuracy scores obtained and the results are validated using 10-fold cross-validation. Python programming language has been used to design the proposed framework. The results attained using LR, NB, SVM, and KNN directly on the original dataset are shown in Table 2 and the results attained using these classifiers in the proposed framework are shown in Table 3. Experimental analysis shows that the highest precision, recall, f1 score, and accuracy values of 98.92%, 98.40%, 98.60%, and 98.84%, respectively, are achieved by the proposed hybrid model using LR classifier. The performance of using LR classifier in the proposed hybrid framework is shown in Figure 2 and the comparison of results attained using all the classifiers on the original dataset and the proposed hybrid framework is shown in Figure 3.

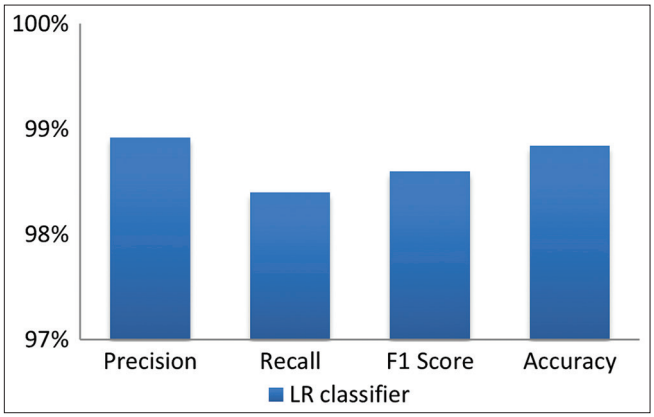

Figure 2: Performance of using logistic regression classifier in the proposed framework

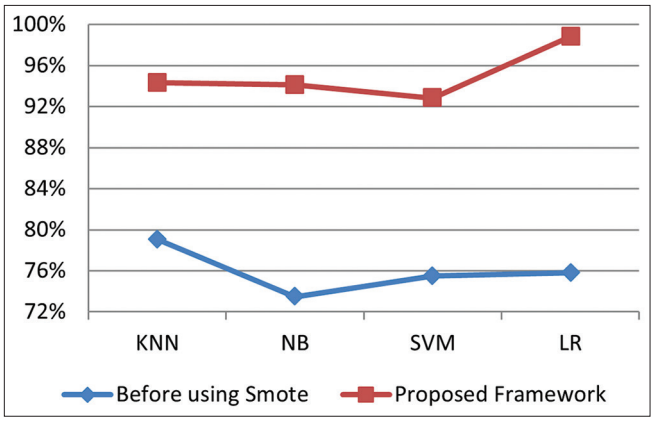

Figure 3: Comparison of accuracy attained using original dataset and the proposed framework

Table 2: Results attained using original dataset

Classifier	*Precision*	*Recall*	*F1 score*	*Accuracy*
KNN	80.72	76.38	78.30	79.07
NB	74.46	71.48	72.82	73.46
SVM	76.33	73.82	74.94	75.48
LR	76.49	74.25	75.26	75.80

KNN: k-nearest neighbors, NB: Naïve Bayes, SVM: Support vector machine, LR: Logistic regression

Table 3: Results attained by the proposed framework

Classifier	*Precision*	*Recall*	*F1 score*	*Accuracy*
KNN	92.93	94.15	93.26	94.32
NB	93.47	92.63	92.97	94.15
SVM	92.71	96.75	94.59	92.80
LR	98.92	98.40	98.60	98.84

KNN: k-nearest neighbors, NB: Naïve Bayes, SVM: Support vector machine, LR: Logistic regression

Conclusion

Diabetes is a chronic disease which is increasing the global burden of health problems year by year. Early prediction of diabetes is very important as diabetes can turn out to be life threatening for the patients in the later stages. In this paper, a hybrid framework for the prediction of type-2 diabetes is developed. Performance of LR, NB, SVM, and KNN using the proposed framework has been compared. LR classifier attained the highest accuracy of 98.84% using the proposed framework and performed better as compared to the application of the classification techniques directly on the original dataset.

References

1. Kavakiotis L, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104-16.
2. International Diabetes Federation. IDF Diabetes Atlas. 9th ed. Brussels, Belgium: International Diabetes Federation; 2019.
3. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, *et al*. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes Atlas, 9th edition. Diabetes Res Clin Pract 2019;157:107843.
4. Uddin S, Khan A, Hossain E, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 2019;19:281.
5. Maleki F, Ovens K, Najafian K, Forghani B, Reinhold C, Forghani R. Overview of machine learning Part 1: Fundamentals and classic approaches. Neuroimaging Clin N Am 2020;30:e17-32.
6. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25:30-6.
7. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlocked 2018;10:100-7.
8. McLaren CE, Chen WP, Nie K, Su MY. Prediction of malignant breast lesions from MRI features: A comparison of artificial neural network and logistic regression techniques. Acad Radiol 2009;16:842-51.
9. Han J, Kamber M, Pei J. Data mining: Concepts and techniques. In: The Morgan Kaufmann Series in Data Management Systems. Amsterdam: Elsevier; 2011.
10. Chen H, Tan C, Lin Z, Wu T. The diagnostics of diabetes mellitus based on ensemble modeling and hair/urine element level analysis. Comput Biol Med 2014;50:70-5.
11. Karegowda AG, Jayaram MA, Manjunath AS. Cascading K-means clustering and K-nearest neighbor classifier for categorization of diabetic patients. Int J Eng Adv Technol 2012;1:147-51.
12. Nilashi M, Ibrahim O, Dalvi M, Ahmadi H, Shahmoradi L. Accuracy improvement for diabetes disease classification: A case on a public medical dataset. Fuzzy Inform Eng 2017;9:354-7.
13. Quan Z, Kaiyang Q, Yamei L, Dehui Y, Ying J, Hua T. Predicting diabetes mellitus with machine learning techniques. Front Genet 2018;9:515.
14. Fernandez A, Garcia S, Herrera F. Addressing the classification with imbalanced data: Open problems and new challenges on class distribution. In: Hybrid Artificial Intelligent Systems. Berlin: Springer; 2011. p. 1-10.
15. Chawla NV, Bowyer KW, Hall LO, Kegelmeye WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-57.
16. UCI Machine Learning Repository. Diabetes Data Set; 2020. Available from: https://www.archive.ics.uci.edu/ml/datasets/diabetes. [Last accessed on 2020 Apr 10].