Role of K-nearest neighbour in detection of Diabetes Mellitus

Roshi Saxena[a], Dr. Sanjay Kumar Sharma[b] Manali Gupta[c]

[a]Research Scholar Department of computer science, SOICT,Gautam Buddha University, Greater Noida-201312, India
[b]Professor and Dean Department of computer science, SOICT, Gautam Buddha University, Greater Noida-201312, India
[c]Research Scholar Department of computer science, SOICT, Gautam Buddha University, Greater Noida-201312, India

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: Diabetes is one of the enduring and continuing illness in the world. According to world health organisation, it was approximately 104 million were suffering from diabetes in 1980 and in year 2014, the figure has risen to 422 million in the world and is expected to double by year 2030. In this paper, we have applied supervised K-nearest neighbour machine learning algorithm on PIMA Indians diabetes dataset. K- nearest neighbour algorithm works on the similarity between presented data and already stored data. We have shown that after the application of proposed algorithm, accuracy has risen from 70.1 % to 78.58% which is an increase of 8.48%.

Keywords: LGS, mathematics, teacher, difficulty Classification, k-nearest neighbour, machine learning algorithm, diabetes, accuracy

1. Introduction

Diabetes Mellitus is a disease which does not allow our body to use the energy which we get from the food properly. When the carbohydrates are being eaten by us, our body turns the energy which we get from food into glucose and passes it to bloodstream. Diabetes can occur in the following form. Either the little amount of insulin is being produced by pancreas or the adequate insulin is being produced but it is not working as it should work i.e., insulin resistance. When the human body eats food, it is broken down into glucose which gives the energy to perform our daily activities. Sugar is transported from the blood vessel to the liver cells. Insulin is being released by pancreas which helps to decompose the glucose. Without the production of insulin, body cells do not get glucose and it can't be used as energy and it gets stored in the stomach thus raising the levels of glucose in the blood. Diabetes is of four types: Type 1 diabetes when the beta cells which produces the insulin are damaged. To control this type of diabetes, insulin from outside sources is provided to the body through injections which helps the food to break down into glucose thus lowering the amount of glucose in the body. Type 2 diabetes when either the enough insulin is not produced, or it does not work properly. This type of diabetes generally occurs in the age group above than 30 and it can be controlled by proper diet, exercise and oral medications. Gestational diabetes occurs during pregnancy when the baby in womb needs more glucose. Gestational diabetes disappears after child birth but the women have higher chance of developing type 2 diabetes later in the life. Due to so many risk factors associated with the diabetes, it is essential to predict the diabetes at preliminary stage. In this article, we have used K-nearest neighbor to predict the diabetes. KNN algorithm is a supervised machine learning algorithm. When the new data is fed to the KNN machine learning algorithm, it checks the similarity between the new data and the already present data. New data is put into most likely similar category.

2. Related Work

In recent years, a good amount of research work has been done to forecast the diabetes using machine learning technique. Maniruzzaman [2] has predicted diabetes by proposing different cross- validation techniques and dimensionality reduction by selection of appropriate features. Dimensions has been reduced by implementing QDA (Quadratic discriminant analysis) [4], LDA (Linear Discriminant analysis) [3] and few machine learning methods such as support vector machine [7], gaussian process classification [6], decision tree [11], artificial neural network [8] , naïve bayes[5], adaboost [9], random forest [12], logistic regression [10] were run on the dataset. Outliers were rejected by author and missing values were replaced by median. AUC were chosen as the parameter to predict the diabetes for various machine learning models and highest AUC were 0.930 for the combination of random forest classifier and logistic regression machine learning model. Support vector machine, decision tree and naïve bayes machine learning classifiers were applied by Sisodia et.al [1] on PIMA Indians diabetes dataset and author has achieved maximum accuracy of 76.30 % through naïve bayes method. Hasan et.al [13] has proposed few feature selection methods and made use of K-NN, decision tree, random forest, Adaboost, naïve Bayes, multilayer perceptron and ensemble of few classifiers and proved that ensemble of Adaboost and gradient is best in determining AUC.

The organization of the remaining paper is as follows: "Materials and Methods section represents materials and methods, including dataset description, tool description prediction algorithms and classifiers evaluation. Results section discusses the results of all classifiers applied. Conclusion discusses the summary of current work and future work.

3. Materials and methods

3.1 Dataset Description

In this research paper, we have made use of PIMA Indians diabetes dataset which contains eight feature attributes and one class attribute. Eight attributes are number of times a woman is pregnant, diastolic blood pressure, diabetes pedigree function, plasma glucose concentration, two-hour serum insulin, age of the patient, triceps skinfold thickness and last but not the least body mass index. The description of the dataset is shown in table 1. We have made use of tool weka to categorize the pregnant women into diabetic and non-diabetic one.

Table 1: Description of PIMA Indian diabetes Dataset

S.No	Attributes	Standard Deviation	Min/Max Value Mean
1	No. of times pregnant	3.4	1/17 3.8
2	Plasma glucose concentration	32	56/197 120.9
3	Diastolic Blood Pressure	19.4	24/110 69.1
4	Triceps skin fold thickness(mm)	16	7/52 20.5
5	2-hour serum insulin	115.2	15/846 79.8
6	Body mass index(kg/m2)	7.9	18.2/57.3 32
7	Diabetes pedigree function	0.3	0.0850/2.32 0.5
8	Age	11.8	21/81 33.2
9	Class	Tested Positive:	Diabetic
		Tested Negative:	Non-Diabetic

3.2 Methodology

In the proposed methodology, we have applied our method on K-nearest neighbor on PIMA Indians diabetes dataset and a brief study of K-nearest algorithm is given below.

3.2.1 K-Nearest Neighbour Algorithm. It is a supervised machine learning algorithm and can be used for classification as well as regression problems. It is also known as lazy learner as it uses all data for training while classifying the data. It looks for the similar features between the new data presented and already present data. New data point is classified based upon similarity feature. Algorithm does not learn from trained data; direct action is performed on the dataset. Dataset is stored and when the new data is fed to the algorithm, new data is classified to that category which is the most similar one.

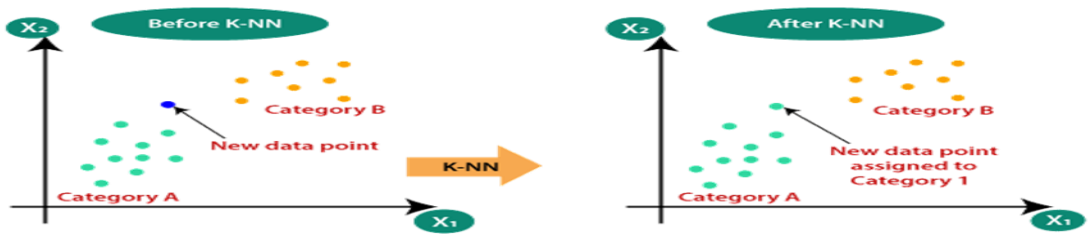

Fig 1: Pictorial Description of KNN algorithm [14]

In this article, we have tried to predict the diabetes through KNN classifier. We have selected few significant features, rejected outliers and optimized certain parameters. After applying proposed algorithm, accuracy has improved by 8%. Confusion Matrix of K-nearest neighbour algorithm is given as:

Table 2: Confusion Matrix of K-nearest neighbour algorithm

	Diabetic	Non-diabetic
Diabetic	142	126
Non-diabetic	103	397

3.3 Proposed Algorithm

1. Load the dataset in weka tool.
2. Run the K-nearest Classifier and note down precision, accuracy and recall. The accuracy is 70.1 %.
3. Select significant features by applying co-relation-based feature selection method.
4. Pre-process the dataset by rejecting outliers and replacing missing values by mean of the data.
5. Optimize the following parameters

a.) Numbers of Neighbours =45.
b.) Batch size =100.
c.) Algorithm= Linear Search
d.) Distance Function= Manhattan Distance

6. After application of steps 3-5, again run the K-nearest neighbour classifier and check the accuracy. The accuracy is now 78.58 %.

We can see that after application of proposed method, there is 8.48 % increase in accuracy. Confusion Matrix after proposed algorithm is:

Table 3: Confusion Matrix of K-nearest neighbour after proposed method

	Diabetic	Non-diabetic
Diabetic	128	114
Non-diabetic	40	437

4. Results and Discussions:

We have applied K-nearest neighbour machine learning classifier on PIMA Indians diabetes dataset in this research work. Experiments are performed using 10-folds cross validation technique. Parameters such as precision, recall, accuracy, f-measure, receiver operating curve (ROC) and area under the curve (AUC) are used for the classification. Results are discussed in table 4. Graphical representation of comparison of results is shown in figure 2 and figure 3.

Table 4: Results before and after the proposed algorithm

	Before Proposed Algorithm	After Proposed Algorithm
Precision	0.696	0.783
Recall	0.702	0.786
Accuracy	70.18	**78.58**
F-Measure	0.698	0.774
AUC	0.650	0.838

Figure 2: Comparison of result before and after proposed algorithm

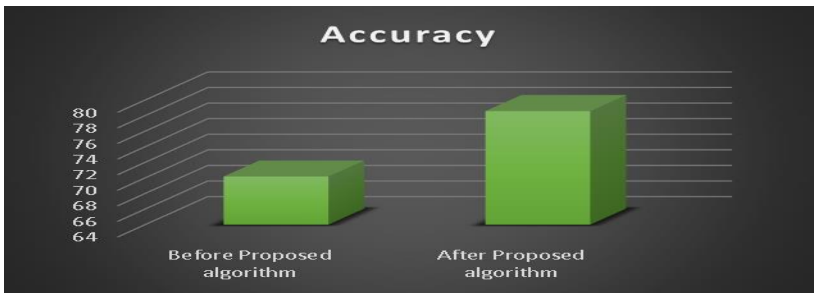

Figure 3: Comparison of accuracy before and after proposed method

5. Conclusion:

Diabetes Mellitus is a disease which can occur to any person having excessive body weight, unhealthy lifestyle, too much workload and stress. Prediction of diabetes in this article is done through k-nearest neighbor which is a lazy algorithm and we have devised a method using by applying pre-processing techniques and optimizing certain parameters of the lazy classifier. After running the proposed methodology on Weka, we found that the accuracy has been raised by 8.48% which is a remarkable increase in the accuracy.

References

1. Sisodia D and Sisodia S, Jan. 2018, “Prediction of diabetes using classification algorithms," Procedia Comput. Sci., vol. 132, pp. 1578-1585
2. Maniruzzaman M, Rahman M J, Hasan M, Suri H S, Abedin M M, El-Baz A, and Suri J S, Jan 2020 “Classification and prediction of diabetes disease using machine learning paradigm," J. health information science and system., vol. 42, no. 5, p. 92 -103.
3. Kamadi V S, Varma P, Rao A A, Mahalakshmi T S and Rao P V N , July 2014 “A Computational Intelligence approach for a better diagnosis of diabetic patients” , J. Computers and Electrical Engineering , Vol. 40 Issue 5, , p 1758-65.
4. Maniruzzaman M, Rahman M J, Hasan M, Suri H S, Abedin M M, El-Baz A, and Suri J S, May 2018. “Accurate diabetes risk stratification using machine learning: Role of missing value and outliers," J. Medical System, vol. 42, no. 5, p. 92
5. Cover T M, Jun. 1965, “Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition”, IEEE Trans. Electronic Computers, vol. EC_14, no. 3, pp. 326-34.
6. McLachlan G, Jun. 2005, “Discriminant analysis and statistical pattern recognition” J. Roy. Stat. Soc., Ser. A, Statist. Soc., vol. 168, no. 3, pp. 635-36.
7. Webb G I, Boughton J R, and Wang Z, Jan. 2005. “Not so naive bayes: Aggregating one-dependence estimators," J. Machine Learning, vol. 58, no. 1, pp. 5-24.
8. Tabaei B P and Herman W H, Nov. 2002, “A multivariate logistic regression equation to screen for diabetes: Development and validation," Diabetes Care, vol. 25, no. 11, pp. 1999_2003.
9. Reinhardt A and Hubbard T, May 1998 “Using neural networks for prediction of the subcellular location of proteins," Nucleic Acids Res., vol. 26, no. 9, pp. 2230-36.
10. Cortes C and V. Vapnik, Sep. 1995 “Support-vector networks," Mach. Learn., vol. 20, pp. 37-297.
11. Breiman L, Oct. 2001 “Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32.
12. Belhouari S B and Bermak A, Nov. 2004, “Gaussian process for nonstationary time series prediction," Comput. Statist. Data Anal., vol. 47, no. 4, pp. 705-12.
13. Hasan K, Alam A, Das D, Hussain E and Hasan M, April 2020 “Diabetes prediction using ensembling of different machine learning classifiers”, IEEE Acess, vol. 8, p 76516-31
14. Javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.