EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

K-Nearest Neighbor Learning based Diabetes Mellitus Prediction and Analysis for eHealth Services

Iqbal H. Sarker*[1,2], Md. Faisal Faruque[1], Hamed Alqahtani[4,5] and Asra Kalim[3]

[1]Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chittagong-4349, Bangladesh.
[2]Swinburne University of Technology, VIC-3122, Australia.
[3]Jazan University, Saudi Arabia.
[4]King Khalid University, Saudi Arabia.
[5]Macquarie University, NSW-2109, Australia.

Abstract

Nowadays, eHealth service has become a booming area, which refers to computer-based health care and information delivery to improve health service locally, regionally and worldwide. An effective *disease risk prediction* model by analyzing electronic health data benefits not only to care a patient but also to provide services through the corresponding *data-driven* eHealth systems. In this paper, we particularly focus on predicting and analysing *diabetes mellitus*, an increasingly prevalent chronic disease that refers to a group of metabolic disorders characterized by a high blood sugar level over a prolonged period of time. K-Nearest Neighbor (KNN) is one of the most popular and simplest *machine learning* techniques to build such a disease risk prediction model utilizing relevant health data. In order to achieve our goal, we present an *optimal K-Nearest Neighbor* (Opt-KNN) learning based prediction model based on patient's habitual attributes in various dimensions. This approach determines the *optimal* number of neighbors with *low error rate* for providing better prediction outcome in the resultant model. The effectiveness of this machine learning eHealth model is examined by conducting experiments on the real-world diabetes mellitus data collected from medical hospitals.

Received on 01 September 2019, accepted on 05 January 2020, published on 15 January 2020

Keywords: health data analytics, diabetes mellitus, data science, machine learning, k-nearest neighbor, predictive analytics, classification, intelligent systems, eHealth, IoT services.

doi:10.4108/eai.13-7-2018.162737

1. Introduction

Healthcare service is one of the most important application areas in the context of Internet-of-Things (IoT), which benefits patients, families, physicians, hospitals and insurance companies. eHealth can make a valuable contribution to optimising health service across the lifespan and the globe as well. eHealth literacy is defined as the ability to seek, find, understand and appraise health information from electronic sources and apply knowledge gained to addressing or solving a health problem. In a broader sense, the term characterizes not only a technical development, but also a state-of-mind, a way of thinking, an attitude, and a commitment for networked, global thinking about healthcare services. Using the internet and related technologies, eHealth seeks to improve health services locally, regionally and worldwide, which can be used within the health sector for clinical, educational, preventative, research and administrative purposes, both on-site and remotely.

*Corresponding author: Iqbal H. Sarker. Email: msarker@swin.edu.eu

In this paper, we aim to focus on building *data-driven model* for the purpose of providing eHealth services.

eHealth services are urgently needed as the incidence of chronic disease is increasing almost all over the world because of changing the living standard of us. According to the report by McKinsey [1], 50% people of USA are suffering from chronic diseases and 80% of USA medical care fee is spent on their treatment, i.e., an average of 2.7 trillion USD annually. As a result, it affects around 18% of their annual GDP. In another statistics in China [2], chronic diseases conditions such as hypertension, diabetes mellitus, stroke, cardiovascular disease, and cancer, are the leading causes of death in China, accounting for 86.6% of all deaths in 2012 [2]. In this work, we particularly focus on diabetes that might be affected to heart disease, kidney disease, nerve damage, blindness, and even death.

Diabetes mellitus is a chronic metabolic disorder that results in abnormal blood glucose (BG) regulations. The BG level is preferably maintained close to normality through self-management practices, which involves actively tracking BG levels and taking proper actions including adjusting diet and insulin medications. BG anomalies could be defined as any undesirable reading because of either a precisely known reason (normal cause variation) or an unknown reason (special cause variation) to the patient. It is caused because of the inappropriate working of the pancreatic beta cells. It has an impact on different parts of the body which incorporates pancreas glitch, risk of heart ailments, hypertension, kidney disappointments, pancreatic issues, nerve harm, foot issues, ketoacidosis, visual unsettling infuences, and other eye issues, waterfalls and glaucoma and so on [3]. There are different purposes behind reason like a way of life of a man, the absence of activity, sustenance propensities, heftiness, smoking, high cholesterol (Hyperlipidaemia), high blood pressure (Hyperglycaemia) etc. which fundamentally increment the risk of treating diabetes. It influences a wide range of ages, including youngsters to grown-up and matured people. Diabetes is typically characterized by a long treatment cycle, numerous complications (e.g., kidney and eye diseases), and recurrent illness. There are two main types of diabetes: type 1 and type 2. Type 1 diabetes occurs because the insulin-producing cells of the pancreas (beta cells) are damaged. In type 1 diabetes, the pancreas makes little or no insulin, so sugar cannot get into the body's cells for use as energy. People with type 1 diabetes must use insulin injections to control their blood glucose. Type 1 is the most common form of diabetes in people who are under age 30, but it can occur at any age. In type 2 diabetes (adult onset diabetes), the pancreas makes insulin, but it either doesn't produce enough, or the insulin does not work properly. Nine out of ten people with diabetes have type 2. This type occurs most often in people who are over 40 years old but can occur even in childhood if there are risk factors present. Type 2 diabetes may sometimes be controlled with a combination of diet, weight management and exercise. However, treatment also may include oral glucose-lowering medications (taken by mouth) or insulin injections (shots).

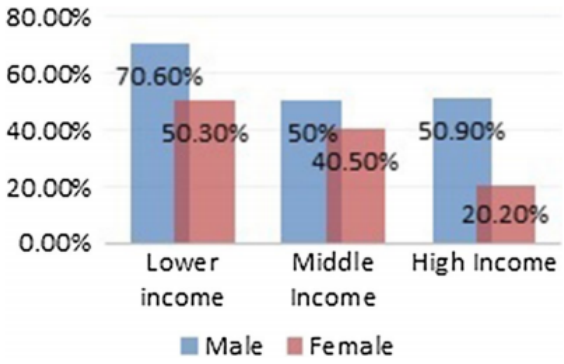

Figure 1. A survey of diabetes death rates among different category of people.

Diabetes mellitus is one of the growing extremely fatal diseases all over the world. As indicated by Canadian Diabetes Association (CDA), somewhere in the year of 2010 to 2020, the quantity of individual person figure out to have diabetic in Canada is relied upon to escalate from 2.5 million to around 3.7 million [4]. According to International Diabetes Federation, number of people having diabetes mellitus achieved 382 million out of 2013 that bring 6.6% of the world's total grown-up population [5]. Based on the world healthcare medical data it has been expected that diabetic disease will increase from 376 million to 490 million within the year 2030 [6]. According to the growing morbidity in recent years, in 2040, the world's diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes [7]. Figure 1 demonstrates the diverse individuals (gender and wage) matured between 29 and 70 years, level of passing because of hypertension [3]. Thus, there is no doubt that this alarming figure needs great attention to analysis and management.

Health management of diabetic patients is an important part of the national basic public health service. Traditionally, diagnosis of a disease are done mostly by expertise and experienced doctors, but still there are cases of wrong diagnosis and treatment. Patient have to undergo various test which are very costly and sometimes all of them are not required. So in this way it will hugely increase the bill of a patient unnecessarily. Healthcare analytics needs a technology that helps to perform a real time analysis on the massive dataset. In healthcare industry the application of

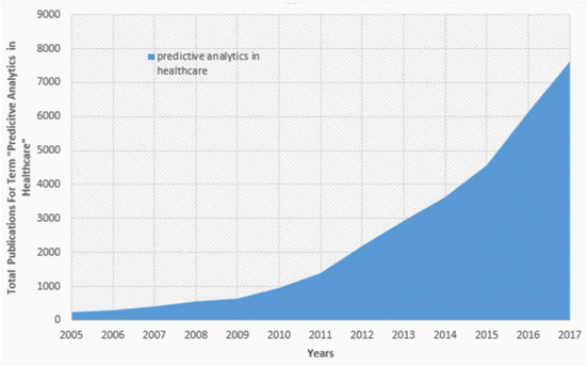

Figure 2. Research outcomes in the area of healthcare predictive analytics.

predictive analytics are significantly high. Predictions can be made about patients, which patients, areas or geographic will be affected by some disease. Due to these applications in healthcare industry predictive analytics have received a huge amount of interest from researchers in past few years. Figure 2, shows the fast increase in the number of research outcomes referring to "predictive analytics in healthcare" from year 2005 to 2017 [8].

Health industry typically contains very large and sensitive data that needs to be handled very carefully. Such medical data makes enable to study on data-driven intelligent eHealth services and systems, in which health professionals and health consumers create and seek information. For the purpose of the study, we collect diagnostic dataset having several attributes diabetic of patients. These attributes are age, diet, hyper-tension, problem in vision, genetic etc. Effectively analyzing such data and corresponding data-driven intelligent eHealth system can be used to provide better services to health care communities. Healthcare analytics refers to the systematic use of these healthcare datasets for business insights, decision making, planning, learning, early prediction and detection of diseases by using different statistical, predictive and quantitative models and techniques. Thus, mining the diabetes data in an efficient way is a crucial concern. With advances in the informatization of medicine, medical industries with large amounts of complicated patient data are keen to extract information from this data to assist the development of these industries. The use of machine learning and other artificial intelligence methods for the analysis of medical data in order to assist diagnosis and treatment is one of the manifestations of smart medicine with the most practical significance. Thus, early prediction of diabetes can be controlled over the diseases and save human life. To achieve this goal, this research work mainly explores the early prediction of diabetes by taking into account various risk factors related to this disease. An early prediction of *disease risks* and corresponding data-driven eHealth model based on relevant disease and patient's habitual attributes can improve the system by enhancing the performance of patient management tasks, in which we are interested.

In the area of machine learning and data science, classification is one of the most important techniques for building a prediction model. Different machine learning techniques are useful for examining the data from diverse perspectives and summarize it into valuable information. K-Nearest Neighbor (KNN) [9] is one of the most popular and simplest machine learning classification technique to predict disease using health data. KNN based model consists only of storing the training data set. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set — its "nearest neighbors". However, in many cases the effectiveness of KNN depends on the K-values, such as K = 1, 2, 3, and so on. Such K-value is defined statically in the traditional KNN based prediction model. As a result, in many cases such static value of K, e.g., K = 1 may cause the *lower prediction accuracy* [5]. Thus, we present an *optimal K-Nearest Neighbor* (Opt-KNN) based disease risk prediction model based on relevant disease and patient's habitual attributes. This Opt-KNN based model dynamically determines the *optimal* number of neighbors for providing better prediction outcome. The effectiveness of this machine learning eHealth model is examined by conducting experiments on the real-world diabetes data collected from medical hospitals. For comparison purposes, experimental work has been carried out using various classification algorithms, such as Decision Tree(DT), Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) and Adaptive Boosting (AdaBoost) on our collected diabetes dataset.

The rest of the paper is organized as follows. Section 2 reviews the related work. We present our methodology in Section 3. We report our experimental results in Section 4. Finally, Section 5 concludes this paper.

2. Literature Review

Diabetes is a chronic disease or group of metabolic disease where a person suffers from an extended level of blood glucose in the body, which is either the insulin production is inadequate, or because the body's cells do not respond properly to insulin. Nowadays, this disease is leading to long-term complications and serious health problems. A report from the World Health Organisation [10] addresses diabetes and its complications that impact on individual physically, financially, economically over the families. The survey says about 1.2 million deaths due to the uncontrolled stage of health lead to death. About 2.2 million deaths occurred due to the risk factors of diabetes

like a cardiovascular and other diseases [3]. The Pima Indian diabetic database at the UCI machine learning research facility has turned into a standard for testing information mining calculations to see their expectation exactness in diabetes information arrangement [3]. The proposed KNN machine learning based model presented in this paper utilizes a real diabetic dataset collected by us directly from the medical hospitals.

In the area of data science and computing, a number of approaches could be useful to build the relevant data-driven predictive models based on machine learning techniques [11] [12] [13] [14] [15] [16] [17]. In particular, a number of researchers use various machine learning classification techniques in their study on health data. Naive Bayes (NB) [18] in one of the most popular classification algorithms, that uses class membership probabilities [19]. In order to analyze health data, Yeh et. al. [20] use naive Bayes classifier in their study. Support Vector Machines (SVM) [21], is another popular classification technique used widely for various predictive analytics. A number of researchers [22] [23] use SVM in order to build a disease prediction model. Regression is widely used in medical field for predicting the diseases or survivability of a patient, which was proposed by Le et. al. [24]. In [25], the authors use logistic regression for the estimation of relative risk for various medical conditions such as Diabetes, Angina, stroke etc.

A very well-known technique for prediction is decision trees [26]. The core algorithm for building decision trees called ID3 proposed by J. R. Quinlan [26]. Based on the ID3 algorithm, a modified algorithm is proposed by Quinlan, namely C4.5 algorithm [27]. A number of authors [28] [20] [29] have studied a decision tree based model for predicting disease. Another tree based classifier is Random Forest [30] that is used in [31] to prediction disease. Bahad et al. [32] study about AdaBoost and gradient boosting algorithms for predictive analytics utilizing the health data.

Among the classifiers in machine learning, K-nearest neighbors (KNN) [9] is considered as one of the simplest classification techniques. It is a type of instance-based learning that takes into account local approximation and all the computation is deferred until classification. It uses a distance function, like Euclidean distance as a similarity measure and determine the K-neighbors. Finally, the prediction result is made based on the majority voting of the neighbors. In the area of modeling disease prediction, a number of authors [33] [34] use k-nearest neighbor in diagnosing disease of the patients. However, its very difficult to assume an *optimal* K value that is needed to predefine in traditional KNN algorithm to build an effective prediction model.

In this paper, we determine the optimal number of K values with low error rate and propose an *optimal K-Nearest Neighbor (Opt-KNN)* based diabetes risk prediction model for providing better eHealth services utilizing relevant health data.

3. Dataset and Model Description

In this section, we describe the health datasets that are used in our study. Furthermore, we provide disease risk prediction model and evaluation methods.

3.1. Health Data and Attributes

For the purpose of this study, we collect real-world diabetes data of five hundred patients from the relevant medical hospitals. The dataset consists of various attributes or risk factors such as diet, hyper-tension, problem in vision, genetic etc. that cause diabetes mellitus. A sample raw dataset has been shown in Table 1. In Table 2, we have summarized the attributes and corresponding values that are used in our approach.

3.2. Data Preparing

To achieve the goal, some data pre-processing tasks have been done on the diabetes dataset. This step is one of the most important phases in the data science process. It prepares and transforms the initial raw datasets collected from the hospitals. Raw data is generally incomplete and inconsistent. Analyzing data that has such problems can produce misleading results. Thus, some data preprocessing tasks can be performed on the raw data before building the model. For instance, the exact numeric value of the attributes is not meaningful to predict diabetes. This particular dataset had both nominal and real valued attributes. As such we convert the numeric attribute values into nominal. For example, the patient's age is classified into three categories, such as Young (10-25 years), Adult (26- 50 years) and Old (above 50 years). Similarly, patient's weight is classified into three categories, such as Underweight (less than equal 40 Kgs), normal (41-60 Kgs) and Overweight (above 60 Kgs). Finally, blood pressure is classified into three categories, such as Normal (120/80 mmHg), Low (less than 80 mmHg) and High (greater than 120 mmHg). The converted categorical values are used in our approach to build the model and in the evaluation process as well, to measure the prediction accuracy of learning techniques.

3.3. Optimal KNN-based Prediction Model

K-Nearest Neighbor (KNN) [9] is a non-parametric machine learning classification technique, i.e., there is no assumption for underlying data distribution. In other words, the model structure determined from the dataset by storing all available cases and predicts new cases based on a similarity measure. It is a well-known instance-based classification technique which classifies

Table 1. Sample dataset of diabetes patients

Sl	Age	Sex	Weight	Diet	Polyuria	Water_Cons	Excessive_Thirst	BP	Hyp_Ten	Tiredness	Problem_in_Vision	Kidney_Problem	Hearing_Loss	Itchy_Skin	Genetic	Glucose_Level_PGP	Diabetic
1	62	Male	67	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	142	Yes
2	53	Female	60	Yes	Yes	No	Yes	Normal	Yes	Yes	No	Yes	Yes	Yes	No	97	No
3	45	Female	55	Yes	Yes	Yes	Yes	Normal	Yes	Yes	No	Yes	Yes	Yes	No	80	No
4	67	Male	65	Yes	Yes	Yes	Yes	High	Yes	Yes	Yes	No	No	Yes	Yes	167	Yes
5	42	Female	52	No	No	No	No	Normal	No	No	No	Yes	No	No	No	172	Yes
6	48	Male	66	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	145	Yes
7	54	Female	65	Yes	Yes	Yes	Yes	High	Yes	Yes	Yes	Yes	No	Yes	Yes	148	Yes
8	60	Male	66	Yes	Yes	Yes	Yes	Low	No	Yes	Yes	Yes	Yes	Yes	No	78	No
9	50	Male	68	No	No	No	No	High	Yes	No	No	Yes	No	No	No	95	No
10	66	Male	62	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	No	Yes	Yes	Yes	156	Yes
11	61	Male	72	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	141	Yes
12	46	Female	54	No	No	No	No	High	Yes	Yes	No	Yes	No	No	Yes	185	Yes
13	71	Male	67	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	No	95	No
14	69	Male	72	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	No	88	No
15	43	Female	64	No	No	No	No	Normal	No	No	No	No	No	No	No	158	Yes

Table 2. Various attributes of the dataset

Attributes	Data Type	Values
Age (Years)	Numeric	{1 to 100}
Sex	Categorical	{Male, Female}
Weight (Kg's)	Numeric	{5 to 120}
Diet	Categorical	Vegetarian, Non-Vegetarian
Polyuria	Categorical	{Yes, No}
Water Consumption	Categorical	{Yes, No}
Excessive Thirst	Categorical	{Yes, No}
Blood Pressure (mmHg)	Numeric	{50 to 200}
Hyper Tension	Categorical	{Yes, No}
Tiredness	Categorical	{Yes, No}
Problem in Vision	Categorical	{Yes, No}
Kidney Problem	Categorical	{Yes, No}
Hearing Loss	Categorical	{Yes, No}
Itchy Skin	Categorical	{Yes, No}
Genetic	Categorical	{Yes, No}
Diabetic	Categorical	{Yes, No}

the new sample based on similarity measure or distance measure. An overview of KNN considering both the training and classification phase, is given below.

- Training phase: the algorithm stores the attribute or feature and corresponding class label of the training samples.
- Classification phase: based on the value of "k" this algorithm performs the classification for the unlabelled test sample. The test sample can be classified into the defined class by calculating the feature similarity. Finally, majority of voting occurs to finalize the classification process.

As mentioned above, a case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor. The performance of KNN for building health prediction

model may depend on K values. According to our goal for predicting binary class in our prediction model, we take into account odd values for K such as 1, 3, 5, 7 etc. For instance, Figure 3 shows an example of K-Nearest Neighbor indicating K = 3 and K = 7.

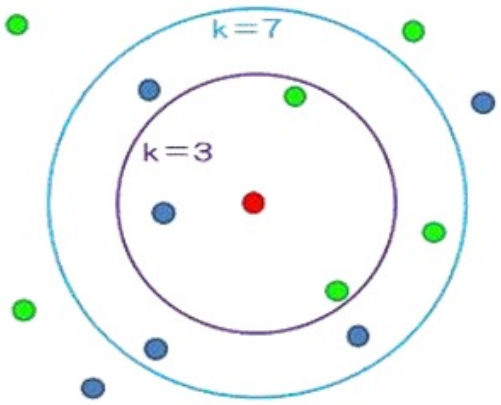

Figure 3. An example of K-Nearest Neighbor indicating K = 3 and K = 7.

In order to achieve better prediction outcome, we determine the optimal number of K value by considering the low error rate according to the patterns of risk factors (attributes) discussed above, and called an Opt-KNN based model. As we have no prior knowledge about the hidden patterns in the dataset, we iteratively use different K values (K++) and calculate the *mean absolute error* rate for each K. If F_i is the prediction, and Y_i is the true value, then the mean absolute error (MAE) is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|F_i - Y_i| \quad (1)$$

In our Opt-KNN based model, mean absolute error rate, presented in Equation 1, measures how close the predictions with the actual outcome over each iteration. The value that yields the lowest error establishes the *optimal K-value* in our Opt-KNN leaning based model. The overall process is set out in Algorithm 1.

Rather than arbitrarily assume the K-value in advance, this Opt-KNN determines the optimal value of K utilizing a given dataset. Thus, this optimal K-value is not static for all datasets, may differ from one to another according to the data patterns.

4. Evaluation and Experimental Results

In this section, we discuss the performance of our Opt-KNN based disease risk prediction model in various aspects using 10-fold cross validation technique [19]. In a 10-fold cross validation, we randomly divide the dataset into 10 parts and calculate the prediction results in 10 iterations. In each iteration, we train the model using 9 parts and test the resultant model using the remaining dataset.

Algorithm 1: Optimal K Nearest Neighbor

Data: Train data: DB_{train}, Test data: DB_{test}
Result: Prediction: $Class_{outcome}$

```
//initialise the value of K and MAE
K ← 1; MAE_init ← 1
foreach test point t in DB_test do
    //find the distance to all training data points
    dist ← calculateDist(t, DB_train)
    store the distances in a list dist_list and sort it
    //Get top K data from the sorted list
    neighbors ← getNeighbors(dist_list, K)
    //Get the most frequent class from neighbors
    Class_pred ← getFrequent(neighbors, class)
    //calculate mean absolute error MAE
    MAE ← calculateMAE(Class_pred, t)
    //compare MAE
    if MAE < MAE_init then
        K_optimal ← K // store as optimal
        MAE_init ← MAE // update initial MAE
         Class_outcome ← Class_pred //result
    end
    increase K // next K value
end
return Class_outcome
```

4.1. Effect of K-value in Prediction Model

In this experiment, we show the effect of K-value on the overall prediction accuracy of the model. For this, we illustrate the results for different K-values shown in Figure 4. As we have two class values for prediction, we take into account odd values of K such as K = 1, 3, 5,.., so on. We did this because of avoiding the conflict while finalizing the final outcome by determining the majority voting. The x-axis of the figure represents various K-values starting from 1 up to 15 and y-axis represents the corresponding mean absolute error (MAE) for these values.

If we observe Figure 4, we see that different K-value gives different results in terms of mean absolute error rate. Thus it is evident that a predefined assumption based K-value, can not ensure the effectiveness of the model. The setting of this K-value should be optimal according to the patterns in the dataset. As can be seen in Figure 4, we get lowest error rate when K=3. Thus, our Opt-KNN based prediction model dynamically selects K=3 as an optimal value to build an effective disease risk prediction model. As the data patterns are not identical in the real word, this optimal K-value may differ in another case.

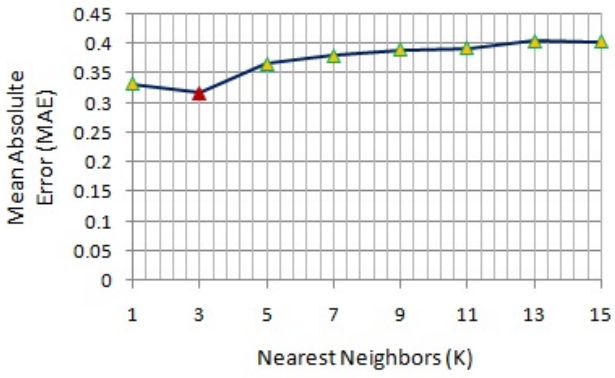

Figure 4. Optimal K-value selection with the lowest MAE in Opt-KNN model.

4.2. Traditional KNN vs Opt-KNN

In this experiment, we show the prediction results for both the traditional KNN and Opt-KNN based model utilizing the same dataset. In a traditional KNN based model a static value of K, e.g., (K=1 or K=5) is chosen as default. On the other hand, in our Opt-KNN based disease prediction model, we dynamically determine the optimal value of K based on the error rate. In Figure 5, we show the prediction results using both the traditional KNN (K=1) and our Opt-KNN based diabetes risk prediction model by considering the relevant attributes discussed above.

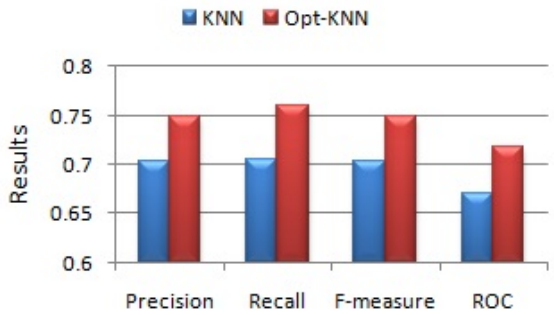

Figure 5. Prediction results of K-Nearest Neighbor (KNN) vs Optimal K Nearest Neighbor (Opt-KNN).

If we observe Figure 5, we see that our Opt-KNN based disease prediction model outperforms the traditional KNN based model. Opt-KNN based model gives better prediction accuracy in terms of precision, recall, f-measure, ROC area [19]. This results show that Opt-KNN is more effective that traditional KNN in terms of prediction accuracy and minimize the additional effort for assuming the K-value.

4.3. Accuracy Comparison with Other Classifiers

In this experiment, we show the effectiveness of our Opt-KNN based model in terms of precision, recall, f-measure, and ROC area, comparing it most popular machine learning classifiers. To do this, first we select 5 baseline classification methods, such as Adaptive Boosting (AdaBoost), Logistic Regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), and Decision Tree (DT) that are frequently used to analyze health data, discussed in Section 2. For these baseline techniques, we utilize the same training and testing datasets in order to compare the techniques fairly.

Figure 6. Prediction accuracy comparison

From Figure 6, we find that our Opt-KNN based disease risk prediction model outperforms other popular classification techniques. The reason is that our Opt-KNN based model gives the better results by taking into account the optimal value of K while building the prediction model (shown in Figure 5) and makes the disease risk prediction model more effective than traditional KNN classifier. The comparison results are also better than other classifiers as well, shown in Figure 6, when using diabetes disease risk data with patients habitual attributes.

5. Conclusion

In this paper, we have presented a k-nearest neighbor learning based diabetes mellitus prediction and analysis for eHealth services. In our model, we have taken into account an optimal number of nearest neighbors for the purpose of building the disease risk prediction model utilizing diabetes data. We have determined the optimal value of K by considering the low error rate to build an effective prediction model. The effectiveness of this model is examined by conducting experiments on the real-world diabetes data collected from medical hospitals. We have analyzed the prediction results by taking into account various risk factors related to this disease using several machine learning techniques. The experimental results could assist health care to take

early prevention and make better clinical decisions to control disease and thus save human life.

Acknowledgement

The authors would like to thank the doctors and hospital staff, and the patients for supporting to collect the real diagnostic data for the purpose of this study.

References

[1] P. Groves, B. Kayyali, D. Knott, S. Van Kuiken, The 'big data'revolution in healthcare, McKinsey Quarterly 2 (3) (2013).

[2] National health and family planning commission of the people's republic of china, National status report on nutrition and chronic disease of residents in China, 2015.

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

[4] M. Mashayekhi, F. Prescod, B. Shah, L. Dong, K. Keshavjee, A. Guergachi, Evaluating the performance of the framingham diabetes risk scoring model in canadian electronic medical records, Canadian journal of diabetes 39 (2) (2015) 152–156.

[5] M. F. Faruque, I. H. Sarker, et al., Performance analysis of machine learning techniques to predict diabetes mellitus, in: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, 2019, pp. 1–4.

[6] C. B. Giorda, V. Manicardi, J. Diago Cabezudo, The impact of diabetes mellitus on healthcare costs in italy, Expert review of pharmacoeconomics & outcomes research 11 (6) (2011) 709–719.

[7] Q. Zou, K. Qu, Y. Ju, H. Tang, Y. Luo, D. Yin, Predicting diabetes mellitus with machine learning techniques, Frontiers in genetics 9 (2018) 515.

[8] M. A. Sarwar, N. Kamal, W. Hamid, M. A. Shah, Prediction of diabetes using machine learning algorithms in healthcare, in: 2018 24th International Conference on Automation and Computing (ICAC), IEEE, 2018, pp. 1–6.

[9] D. W. Aha, D. Kibler, M. K. Albert, Instance-based learning algorithms, Machine learning 6 (1) (1991) 37–66.

[10] W. H. Organization, et al., Global report on diabetes. 2016 (2017).

[11] I. H. Sarker, A. Kayes, P. Watters, Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage, Journal of Big Data 6 (1) (2019) 57.

[12] I. H. Sarker, A. Colman, J. Han, Recencyminer: mining recency-based personalized behavior from contextual smartphone data, Journal of Big Data 6 (1) (2019) 49.

[13] I. H. Sarker, Context-aware rule learning from smartphone data: survey, challenges and future directions, Journal of Big Data 6 (1) (2019) 95.

[14] I. H. Sarker, A machine learning based robust prediction model for real-life mobile phone data, Internet of Things 5 (2019) 180–193.

[15] I. H. Sarker, A. Colman, J. Han, A. I. Khan, Y. B. Abushark, K. Salah, Behavdt: A behavioral decision tree learning to build user-centric context-aware predictive model, Mobile Networks and Applications (2019) 1–11.

[16] I. H. Sarker, A. Colman, M. A. Kabir, J. Han, Individualized time-series segmentation for mining mobile phone user behavior, The Computer Journal 61 (3) (2017) 349–368.

[17] I. Sarker, Mobile data science: Towards understanding data-driven intelligent mobile applications, EAI Endorsed Transactions on Scalable Information Systems 5 (19) (2018) e4.

[18] G. H. John, P. Langley, Estimating continuous distributions in bayesian classifiers, in: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.

[19] J. Han, J. Pei, M. Kamber, Data mining: concepts and techniques, Elsevier, Amsterdam, Netherlands, 2011.

[20] D.-Y. Yeh, C.-H. Cheng, Y.-W. Chen, A predictive model for cerebrovascular disease using data mining, Expert Systems with Applications 38 (7) (2011) 8970–8977.

[21] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, K. R. K. Murthy, Improvements to platt's smo algorithm for svm classifier design, Neural computation 13 (3) (2001) 637–649.

[22] T. H. A. Soliman, A. A. Sewissy, H. AbdelLatif, A gene selection approach for classifying diseases based on microarray datasets, in: Computer Technology and Development (ICCTD), 2010 2nd International Conference on, IEEE, 2010, pp. 626–631.

[23] C.-L. Huang, H.-C. Liao, M.-C. Chen, Prediction model building and feature selection with support vector machines in breast cancer diagnosis, Expert Systems with Applications 34 (1) (2008) 578–587.

[24] S. Le Cessie, J. C. Van Houwelingen, Ridge estimators in logistic regression, Applied statistics (1992) 191–201.

[25] C. Gennings, R. Ellis, J. K. Ritter, Linking empirical estimates of body burden of environmental chemicals and wellness using nhanes data, Environment international 39 (1) (2012) 56–65.

[26] J. R. Quinlan, Induction of decision trees, Machine learning 1 (1) (1986) 81–106.

[27] J. R. Quinlan, C4.5: Programs for machine learning, Machine Learning (1993).

[28] C. Ordonez, Comparing association rules and decision trees for disease prediction, in: Proceedings of the international workshop on Healthcare information and knowledge management, ACM, 2006, pp. 17–24.

[29] J.-Y. Yeh, T.-H. Wu, C.-W. Tsao, Using data mining techniques to predict hospitalization of hemodialysis patients, Decision Support Systems 50 (2) (2011) 439–448.

[30] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[31] M. Khalilia, S. Chakraborty, M. Popescu, Predicting disease risks from highly imbalanced data using random forest, BMC medical informatics and decision making 11 (1) (2011) 51.

[32] P. Bahad, P. Saxena, Study of adaboost and gradient boosting algorithms for predictive analytics, in: International Conference on Intelligent Computing and Smart Communication 2019, Springer, 2020, pp. 235–244.
[33] M. Shouman, T. Turner, R. Stocker, Applying k-nearest neighbour in diagnosing heart disease patients, International Journal of Information and Education Technology 2 (3) (2012) 220–223.
[34] C.-H. Jen, C.-C. Wang, B. C. Jiang, Y.-H. Chu, M.-S. Chen, Application of classification techniques on development an early-warning system for chronic illnesses, Expert Systems with Applications 39 (10) (2012) 8852–8858.