Annals of R.S.C.B., ISSN: 1583-6258, Vol. 24, Issue 2, 2020, Pages. 1257 - 1261
Received 24 October 2020; Accepted 15 December 2020.
.

Medical Diagnostic System based on Data Mining Classification Techniques

A. Peter[1*], K. Manoj[2], P. Kumar[3]
[1,2]Department of Statistics, ManonmaniamSundaranar University, Abhishekapatti, Tirunelveli, India
[3]Centre for Information Technology & Engineering, ManonmaniamSundaranar University, Abhishekapatti, Tirunelveli, India
*[1]apeter1978@gmail.com, [2]manojstatms@gmail.com, [3]kumarcite@gmail.com

ABSTRACT

Data mining means a vast volume of data to be extracted and mined. We need to categorise the data after extracting this information. One of the key responsibilities in the area of healthcare is data mining categorization. Diagnosis of health problems in the realm of medical science is an essential and hard endeavour. In medical research, there are various sorts of illnesses. Diabetes illness is one of the major diseases of human health that is highly dangerous. Diabetes mellitus is a collection of metabolic diseases that are generally called diabetes when blood sugar levels are high over a lengthy period. The categorization of diabetes illness is one of the major medical difficulties because it has to do with the health conditions of the human body directly; accurate identification and careful treatment of this kind of disease may be resolved. Different writers worked on diabetes categorization and a different model for classification accuracy. In this study three classification approaches were Discriminant Analysis; for the diabetes classification, Multilayer Perceptron and KNN classifications were employed. The main objective of this research is to compare the classification technology with classification accuracy.

Keywords
Data mining,Discriminant, Healthcare, KNN, Multilayer Perceptron

Introduction

Data mining is a process in which huge data warehouses are automatically found to identify patterns and trends that go beyond simple analyses. The data collection employs advanced mathematical algorithms to separate the data and assess the likelihood of future events. Classification is a technique for data mining that applies to the categories or classes of objects in a collection. The classification is to forecast the target class properly in the data for each example. For instance, a categorization model that identifies borrowers as low, medium or high-risk credit. Diabetes is a condition that has too high amounts of blood glucose or blood sugar. The food you eat is glucose. Insulin is a hormone that helps your cells gain energy from glucose. Your body does not produce insulin with type I diabetes. Your body does not manufacture or utilise insulin well with type II diabetes, the most prevalent kind. Glucose remains in your blood without enough insulin. Pre-diabetes may also be present. This indicates your blood sugar isn't high enough to name diabetes but is higher than normal. Pre-diabetes makes you more likely to acquire diabetes of type II.

Review of Literature

The literature study reveals several diagnostic outcomes for data mining applications. Many researchers have recently employed various categorization algorithms for the diagnosis of diseases. Several researchers have concentrated on medical research utilising diabetes data sets [1],[2] for data mining. The capacity or potential of diabetes illnesses to cause harm to distinct areas of the human body is mentioned as follows from those human parts impacted by diabetes: the human heart, eye, kidney and nerves. How many chronic, hazardous diseases that shorten human lives is easily imagined, as it is indicated. Different algorithms such as Glucose, Blood pressure (BP), Skin thickness (ST), Insulin, Body max index (BMI), Diabetes pedigree function (DPF) and age have been described. It didn't contain all parameters. Only sample data utilised for small samples. Diabetes dataset was used for ANN, EM, GMM, Logistic Regression, and SVM. Better precision and performance than previous algorithms have been offered to ANN (artificial neural network).
A study of two distinct complicated illnesses, including heart disease and cancer disease, based on classification [3]. In several modes of diabetic intervention controls, data mining techniques based on classification have been proposed [4]. Data mining approaches based in the nearest neighbour have been debated and compared [5]. The most

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 24, Issue 2, 2020, Pages. 1257 - 1261
Received 24 October 2020; Accepted 15 December2020.

successful and used systems for classification, identification, and segmentation include machine learning techniques. A Perceptron model based on neural networks has been created to provide security for medical devices in real-time [2].

Methodology

UCI Repository has gathered a multi-dimensional medical services dataset. This dataset includes 100 perceptions with 7 distinct classes of diabetes factors (PG Concentration, Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques 185: Diastolic BP, Tri-Fold Thick, Serum Ins, BMI, DP capacity, age and illness).

Discriminant Analysis

In the classification approach first established in 1936 by R.A. Fisher, Linear Discriminant Analysis (LDA) is a method. It is simple, resilient from mathematics and frequently provides models whose precise techniques are as excellent as complicated. Discriminatory analysis of functions is used to find the differences between the natural two or more categories. The discriminatory analysis can be described as.

$$\delta_k(x) = x\,\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \qquad (1)$$

Now, assuming only two classes with equal distributions, to find:

$$x = \frac{\mu_1^2 - \mu_2^2}{2\,(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2} \qquad (2)$$

Multilayer Perceptron

A Multilayer Perceptron (MLP) is a feedback mechanism for artificial neural networks, which translates input data sets to a specific output set[6]. There are two sides to the problem: understanding the network topology and the weight of the link. It has been shown that the weights given a fixed network topology are determined using a reasonably simple algorithm[7].
Perception is a linear classification, which is an algorithm that divides the input into a straight line between two categories. The input is usually a weight-multiplying function vector and a bias.
A Perceptron creates one output based on many reliable inputs by constructing a linear combination with the following input weights:

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \varphi(w^T x + b) \qquad (3)$$

The input vector is the biais and phi is the non-linear activation function. where the weight vector is mentioned. Often a concealed layer is all that is needed and by optimising anticipated accuracy a suitable number of units is established for this layer[8].

k-Nearest Neighbour

KNN is one of the easiest grading algorithms that store all known instances and classify new cases based on a similarity measurement (e.g., distance functions). As early as the 1970s, KNN was already utilised as a nonparametric approach in statistical estimation and pattern identification.
A case is categorised by a majority vote of its neighbours, which is allocated by a distance function to the most frequent class of its K closest neighbours. If so, the case will simply be allocated to the neighbour class. KNN is based on the distance from Euclidean, Manhattan, Minkowski. Furthermore, all three-distance actions are applicable only for continuous variables.

Methodology

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 24, Issue 2, 2020, Pages. 1257 - 1261
Received 24 October 2020; Accepted 15 December2020.

KNN is one of the easiest grading algorithms that store all known instances and classify new cases based on a similarity measurement (e.g., distance functions). As early as the 1970s, KNN was already utilised as a nonparametric approach in statistical estimation and pattern identification.
A case is categorised by a majority vote of its neighbours, which is allocated by a distance function to the most frequent class of its K closest neighbours. If so, the case will simply be allocated to the neighbour class. KNN is based on the distance from Euclidean, Manhattan, Minkowski. Furthermore, all three-distance actions are applicable only for continuous variables.

Result and Discussion

Table 1. Results of discriminant analysis technique

	No of Instances	Percentage (%)
Instances properly classified	86	86
Classified instances incorrectly	14	14
Comprehensive instances	100	100

Table 2. Results of Multilayer Perceptron technique

	No of Instances	Percentage (%)
Instances properly classified	90	90
Classified instances incorrectly	10	10
Comprehensive instances	100	100

Table 3. Results of K-nearest Neighbor technique

	No of Instances	Percentage (%)
Instances properly classified	83	83
Classified instances incorrectly	17	17
Comprehensive instances	100	100

Sensitivity, specificity and accuracy are assessment measures
(i) Sensitivity = TP/P
(ii) Species = TN/N
(iii) Precision = (sensitivity + spec) / 2

Table 4. Summarization of Prediction Techniques with Performance Prediction Technique

	Sensitivity	Specificity	Accuracy
Discriminatory Analysis	0.87	0.83	0.85
Perception Multilayer	0.94	0.79	0.87
K-nearest neighbor	0.71	0.48	0.60

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 24, Issue 2, 2020, Pages. 1257 - 1261
Received 24 October 2020; Accepted 15 December2020.

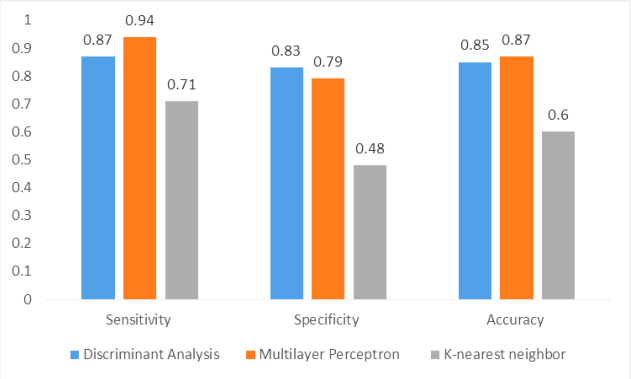

Figure 1. Comparison of Prediction Techniques

Where TP is truly positive, TN is true negative; P and T are positive and negative. A high detection, low specificities and high accuracy must be a good predictor. Table 4 summarises the comparisons of these metrics to the three prediction approaches.

Conclusion

In this study, Discriminatory analysis, Multilayer Perceptron and k-Nearest Neighbour were implemented for the data about Diabetes. The research shows that classification development will be entirely different for categorization methods. The histograms show that the accuracy of the discriminating analytical data is 0.85, the accuracy of the multilayer perceptron is 0.87 and the accuracy of the KNN is 0.60. Multilayer Perceptron is greater than Discriminant and Nearest Neighbor.

References

[1] S. A. Saji and K. Balachandran, “Performance analysis of training algorithms of multilayer perceptrons in diabetes prediction,” in 2015 International Conference on Advances in Computer Engineering and Applications, 2015, pp. 201–206.

[2] H. Rathore, L. Wenzel, A. K. Al-Ali, A. Mohamed, X. Du, and M. Guizani, “Multi-Layer Perceptron Model on Chip for Secure Diabetic Treatment,” IEEE Access, vol. 6, pp. 44718–44730, 2018.

[3] R. Sharma, S. N. Singh, and S. Khatri, “Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey,” in 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT), 2016, pp. 687–691.

[4] A. A. Aljumah, M. K. Siddiqui, and M. G. Ahamad, “Application of Classification based Data Mining Technique in Diabetes Care,” J. Appl. Sci., vol. 13, no. 3, pp. 416–422, Mar. 2013.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 24, Issue 2, 2020, Pages. 1257 - 1261
Received 24 October 2020; Accepted 15 December2020.

[5] K. Manoj, K. S. Kannan, and E. Sakthivel, “A Comparative Study on Nearest-Neighbor Based Outlier Detection in Data Mining,” Int. J. Eng. Futur. Technol., vol. 8, no. 8, pp. 28–32, May 2016.

[6] L. D. X and R. T, “Analysis of multilayer perceptron machine learning approach in classifying protein secondary structures,” Biomed. Res., vol. 0, no. 0, pp. 166–173, Sep. 2016.

[7] I. H. (Ian H. Witten, E. Frank, and M. A. (Mark A. Hall, Data mining: practical machine learning tools and techniques. Morgan Kaufmann, 2011.

[8] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[9] Furstenau, L.B.; Sott, M.K.; Homrich, A.J.O.; Kipper, L.M.; Al Abri, A.A.; Cardoso, T.F.; López-Robles, J.R.; Cobo, M.J. 20 Years of Scientific Evolution of Cyber Security: A Science Mapping. In Proceedings of the International Conference on IndustrialEngineering and Operations Management, Dubai, UAE, 10–12 March 2020.

[10] Kipper, L.M.; Furstenau, L.B.; Hoppe, D.; Frozza, R.; Iepsen, S. Scopus scientific mapping production in industry 4.0 (2011–2018):A bibliometric analysis. Int. J. Prod. Res. 2019, 58, 1605–1627.

[11] Wang, X.; Guo, J.; Gu, D.; Yang, Y.; Yang, X.; Zhu, K. Tracking knowledge evolution, hotspots and future directions of emergingtechnologies in cancers research: A bibliometrics review. J. Cancer 2019, 10, 2643–2653.

[12] dos Santos, B.S.; Steiner, M.T.A.; Fenerich, A.T.; Lima, R.H.P. Data mining and machine learning techniques applied to publichealth problems: A bibliometric analysis from 2009 to 2018. Comput. Ind. Eng. 2019, 138, 106120.

[13] Sott, M.; Furstenau, L.B.; Kipper, L.M.; Giraldo, F.D.; Lopez-Robles, J.R.; Cobo, M.J.; Zahid, A.; Abbasi, Q.H.; Imran, M.A.Precision Techniques and Agriculture 4.0 Technologies to Promote Sustainability in the Coffee Sector: State of the Art, Challenges and Future Trends. IEEE Access 2020, 8, 149854–149867.

[14] López-Robles, J.R.; Otegi-Olaso, J.R.; Cobo, M.J.; Bertolin-Furstenau, L.; Kremer-Sott, M.; López-Robles, L.D.; Gamboa-Rosales, N.K. The relationship between Project Management and Industry 4.0: Bibliometric Analysis of Main Research Areas through Scopus. In Proceedings of the 3rd International Conference on Research and Education in Project Management—REPM 2020,Bilbao, Spain, 20–21 February 2020; p. 56.

[15] Kipper, L.M.; Iepsen, S.; Forno, A.J.D.; Frozza, R.; Furstenau, L.; Agnes, J.; Cossul, D. Scientific mapping to identify competenciesrequired by industry 4.0. Technol. Soc. 2021, 64, 101454.