International Journal of Future Generation Communication and Networking
Vol. 14, No. 1, (2021), pp. 714–719

Heart Disease Prediction Using Knn

K.Uma[1], Dr. K. F.Bharati[2]

[1]M.Tech Computer Science, Department of Computer Science and Engineering, JNTU College of Engineering, JNTU University, Ananthapuramu, Andhra Pradesh.
[2]Assistant Professor, Department of Computer Science and Engineering, JNTUA College of Engineering, JNTU University, Ananthapuramu, Andhra Pradesh.

Abstract

Heart disease is the primary cause of death now-a-days. Treatments of heart disease patients have been advanced, for example with Machine-To-Machine (M2M) technology to enable remote patient monitoring. M2M is used to take care of a remote heart disease patient, medical condition should be measured periodically at home. Thus, it is difficult to perform complex tests which need physicians to help. Meanwhile, heart disease can be predicted by analyzing some of the patient's health parameters. With the help of data mining techniques, heart disease prediction can be improved. In this project improved k-Nearest Neighbor (KNN) algorithm is used. The experimental results show that the proposed KNN algorithm gives better results than the existing system in terms of accuracy.

1. INTRODUCTION

Classification in medical diagnostics can aid in disease diagnosis and predicts outcomes in response to the treatment. Many efforts have been made to improve the classification performance. For instance, in the traditional methodology for classification, logistic regression, the Dichotomous Classification Tree was applied in diagnosing breast cancers. Non-parametric empirical Bayes algorithm was developed for integrative genetic risk prediction of complex diseases with binary phenotypes. A hierarchical Support Vector Machine (SVM) based algorithm was employed in the Electro Encephalon Gram (EEG) based motor imagery classification task. Bionic algorithms were also introduced in the classification of medical data. A self-adaptive niche genetic algorithm with the random forest was proposed to build a model for sepsis patient's stratification, Classification rules were extracted byant-miner algorithm and thereby applied in diagnosing heart disease, etc.

However, in the practice of medical classification, data are usually class- imbalanced, which means the distributions of classes are not uniform. In binary- classification a case, the class with larger distribution is named as the majority while the other is named as the minority. Dealing with the class-imbalanced data, conventional algorithms are prone to consider tend to minority observation as noise or outliers and ignore them in the classifying, thereby tend to classify samples into the majority class. Consequently, the predictive accuracy for the minority class will be much lower than that for the majority class. To diagnose a particular disease, a physician has to explore the patient's data and consider many factors (e.g. family history, age, body mass index, etc.).A physician's diagnosis can be subjective and is highly dependent on the experiences.

Hence, many automated classification systems that use machine learning approaches have been developed to help physicians obtain an objective second opinion for diagnosis decisions. A variety of classifiers have been utilized for diagnosis, such as artificial neural networks, support vector machines, Naïve Bayes, Decision trees, nearest- neighbor, etc. Besides, hybrid models that harness the power of different classifiers have also been proposed. However, the diagnosis decision based on the classification result of a single classifier or a hybrid model only might be weak. Different classifiers probably offer contradictory classification results while providing complementary information.

ISSN: 2233-7857IJFGCN

Therefore, it is helpful to combine the decisions of multiple classifiers. If the decision-making is based on a group of classifiers that considers the individual opinion of each classifier, the misclassified data - especially the patients who are undiagnosed by a certain classifier might be correctly diagnosed due to the correct decisions of other classifiers. There are some methods for combining classifiers, including the mixture of experts, voting, boosting, bagging, etc.
A few of them have been adapted for the diagnosis of diabetes. Some of these methods do not consider the weight of classifiers or each classifier has equal weight. But in fact, the weights of classifiers should be different and should be counted in the final decision. It makes more sense to give larger weights to classifiers that often make correct decisions and smaller weights to classifiers that usually make wrong decisions. On the other hand, some other methods adjust the weights of the classifier based on their power of prediction. In the meanwhile, they iteratively adjust the weights of instances, meaning that hard-to-classify instances get higher weights, which again influences the predictions of classifiers. The iterative interference between classifiers and instances makes the decision-making procedure complicated and time-consuming.

2. LITERATURE SURVEY

An integrated framework K-Nearest Neighbor (KNN) with Ant Colony Optimization (ACO) technique [3]. The outcomes are compared with four dissimilar algorithms and the integrated framework shows accuracy, i.e., 70.26%. Few authors proposed an ensemble framework using hierarchical majority voting and multi-layer classification for the classification of disease and analysis using a data mining approach.

The method used in assembling is majority vote based and it is designed for every data set that belongs to the heart disease domain. The experiment prediction of data sets from different resources has two benchmarks. The accuracy of the ensemble model is90%. Experimental observation shows that the best combination is when one of its classifiers is a Naïve Bayes with an accuracy of 92%. In 2015, the researchers improved the bagging technique and integrated it with the weighted voting scheme. They presented a novel classifier ensemble for the analysis and examination of heart disease.
The approach used 5 heterogeneous classifiers named Naïve Bayes, SVM, linear regression, instance-based learner, QDA (Quadratic Discriminant Analysis), and obtained an accuracy of 84.16%. Extreme Learning Machine (ELM) is used to perfect attributes like age, sex, blood sugar, cholesterol, etc. The technique can substitute expensive medical checkups with a cautionary message for the patient which shows the probability of heart disease. This technique is applied toreal-world data where approximately 300 patients data have been collected by the Cleveland clinic foundation. The accuracy shown by this model is 80%.
The researchers developed an intelligent, disease prediction classifier for heart disease prediction and analysis. By combining five different machine learning classifiers, an ensemble model results that produce the prediction information for heart disease. Five different sets of attributes were used from five different data sets. They were assembled by a majority voting method for training and testing. The experimental result showed that MV5 predicts with high accuracy, i.e., 88.52% as compared to other techniques.

The hybrid model in which major risk factors are used for the analysis of heart disease [8,9]. The hybrid model involves two data mining tools, one is a neural network and other one is genetic algorithm. A genetic algorithm initializes the weight of a neural network. Adapting power of this model is fast and as compared to other models prediction accuracy is 89%.

ISSN: 2233-7857IJFGCN

3. PROPOSED SYSTEM

Flow chart for the proposed system is implemented as shown in figure3.1.

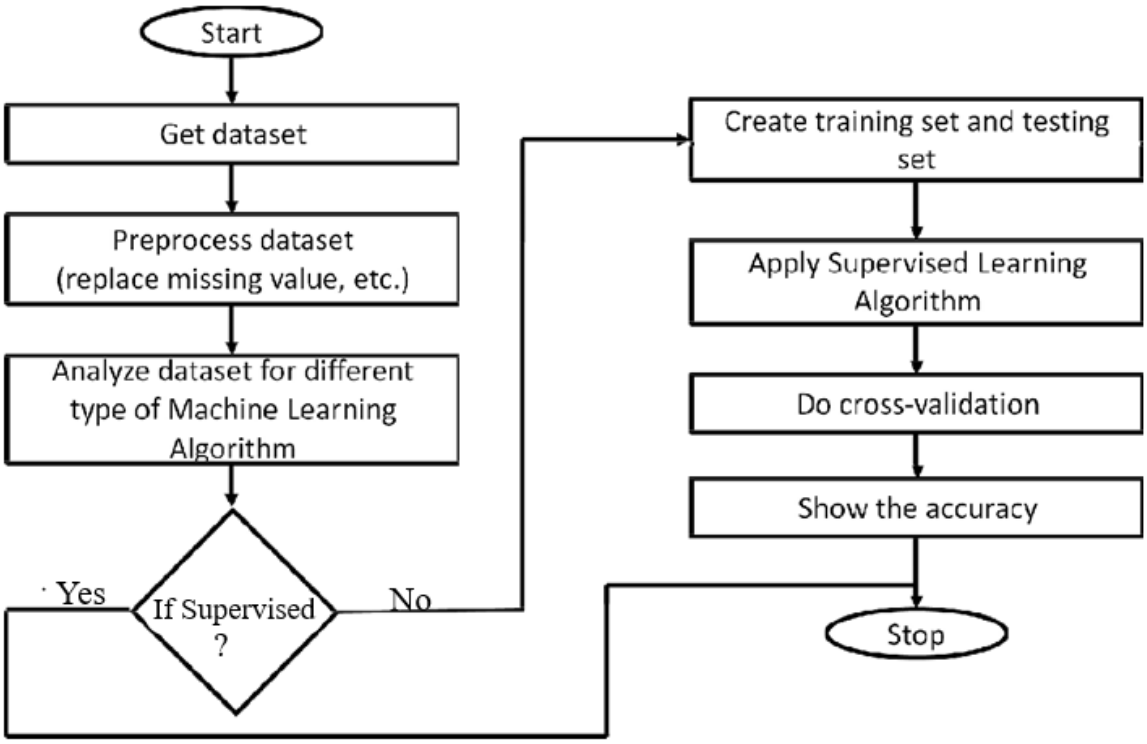

Fig 3.1: Flow Chart Implementation

3.1 DATA PREPARATION

Data Preparation is the act of manipulating (pre-processing) raw data into a form that can readily and accurately be analyzed. Some datasets have values that are missing, invalid, or otherwise difficult for an algorithm to process. If data is missing, the algorithm can't use it. If data is invalid, it causes the algorithm to produce less accurate or even misleading outcomes.

3.2 DATA MODELLING

After collecting and preparing data the next step in the process is modeling the data, which means divide the data into training and testing sets, choose the algorithm and pass the data to the algorithm then make predictions. In this project instead of using the train/test split method, here K-Fold cross-validation in this procedure a single parameter called K is of nearest neighbors. It is used to calculate the distance between the query-instance and all the training samples. Sort the distance and determine nearest neighbors based on the K-th minimum distance.

Algorithm

1. Start
2. Collect the dataset.
3. Select the dependent attribute.

Analyze dataset for different type of machine learning algorithm.

ISSN: 2233-7857IJFGCN

If supervised:
Create a training set and testing set using 10- fold cross validation.
Apply supervised learning algorithm.

4. Preprocess the data.
5. Do testing using the test set.
6. Show the accuracy.
7. Stop

4. EXPERIMENTAL RESULTS

Formulas:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d}$$

$$\text{True Positive rate (or) Recall (or) Sensitivity} = \frac{d}{c+d}$$

$$\text{Specificity, True negative rate} = \frac{a}{a+b}$$

$$\text{Precision, Predicted positive value} = \frac{d}{b+d}$$

$$\text{False positive rate} = \frac{b}{a+b}$$

$$\text{False negative rate} = \frac{c}{c+d}$$

$$\text{F1 Score} = \frac{2*(\text{Recall}*\text{Precision})}{\text{Recall}+\text{Precision}}$$

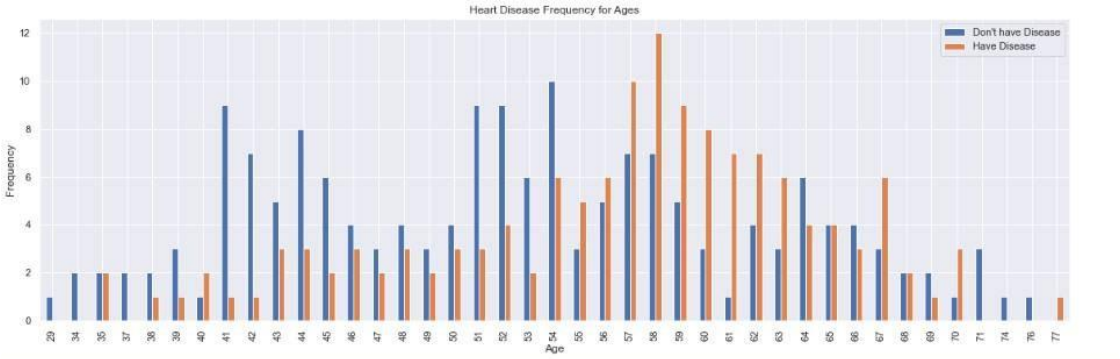

Figure 4.1: Disease vs Age

ISSN: 2233-7857IJFGCN

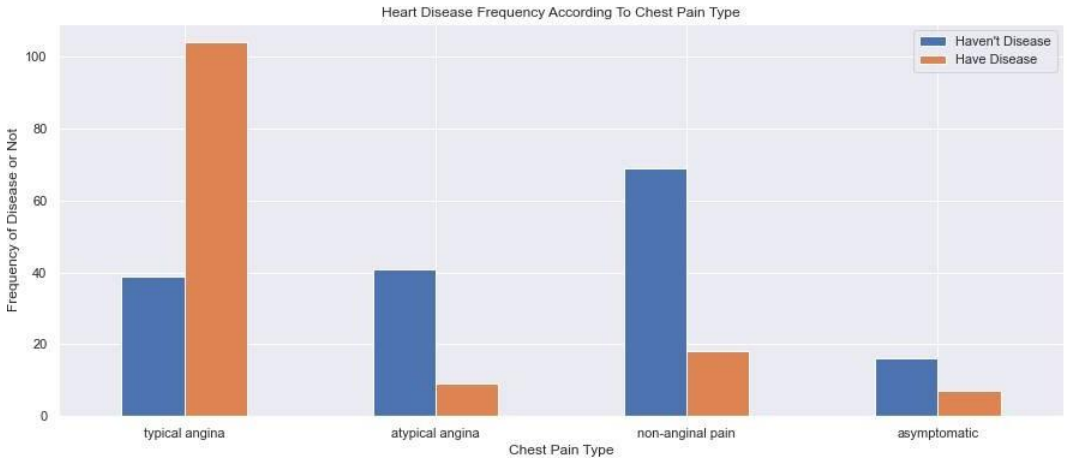

Fig4.2: Disease vs Pain type

5. CONCLUSION

Data mining techniques have been used in many fields, one of them is healthcare. This objective is to check whether heart attack prediction can be based on fewer parameters than what is recommended in previous studies. Experiments using 8 parameters with KNN shows good accuracy if compared with 13 parameters, even with other data mining algorithms like Naive Bayes and Decision Tree (in this research we use Simple CART). The benefit of the result from this study is: That 8 simple parameters are good enough to be used in heart attack prediction. Future research can be used as parameters in remote patient monitoring using Machine-To-Machine (M2M) technology, especially for patients treated at home or remote clinics. The end-to-end M2M will be built and a prediction system will be embedded as the novel feature.

6. References

[1] Zhong Xiao, Ma Shaoping, et al. “Survey of Data Mining [J]”, fuzzy recognition and artificial intelligence,2001,01.

[2]T.Imielinski and H. Mannila, “A database perspective on knowledge discovery”. Communications of ACM, 39:58-64,1996.

[3] M. S. Chen, J. Han, and P. S. Yu, “Data mining: An overview from a database perspective”. IEEE Trans. Knowledge and Data Engineering, 8:866-883,1996.

[4] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, “Advances in Knowledge Discovery and Data Mining”. AAAI/MIT Press, 1996.

[5] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”. Morgan Kaufmann,2000.

[6] John Durkin, CaiJingfeng, CAI Zixing, “Decision tree technology and its current research direction [J]”, Control Engineering, 12 (1)2005.

[7]Jiawei Han, MichelineKamber, “Data Mining Concepts and Techniques”, [M]SecondEdition,2007,3.page3-4.

[8] BoshraBrahmi, MirsaeidHosseiniShirvani, “Prediction and Diagnosis of Heart Disease by Data Mining Techniques”, Journals of Multidisciplinary Engineering Science and Technology, vol.2, 2 February 2015, pp.164-168.

[9] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, “Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration,” Journal of biomedical informatics, vol. 53, pp. 220–228, 2015.

ISSN: 2233-7857IJFGCN

[10] Marcos D. Assunção et.al. “Big Data computing and clouds: Trends and future directions”, Journal of Parallel and Distributed Computing Volumes 79–80, May2015.
[11] S. Patel and H. Patel, “Survey of data mining techniques used in the healthcare domain,” Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60,March 2016.
[12] Mr. ChalaBeyene, Prof. PoojaKamat, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques”, International Journal of Pure and Applied Mathematics, 2018.

ISSN: 2233-7857IJFGCN