Mody University International Journal of Computing and Engineering Research
3 (2), 2019, 77-79
http://www.publishingindia.com/micer

Impact of Machine Learning and Data Mining in Healthcare

Somil Jain[1*] and Puneet Kumar[2]

[1]Department of Computer Science & Engineering, SET, Mody University of Science & Technology, Lakshmangarh, Sikar, Rajasthan, India. Email: somiljn90@gmail.com
[2]Department of Computer Science & Engineering, SET, Mody University of Science & Technology, Lakshmangarh, Sikar, Rajasthan, India. Email: professor.pkumar@gmail.com
*Corresponding Author

Abstract: To diagnose the chronic disease with the traditional approaches is getting obsoleted now a days and not providing the effective results. In order to get rid from the traditional approaches various machine learning and data mining approaches turning to be efficient one for diagnosing the chronic diseases like kidney failure, cancer, heart disease at a very early stage which is turning to be helpful for the decrement in the mortality rate. In this paper an attempt is made to provide the working mechanism of the various methods of machine learning and how they helpful in the healthcare sector.

Keywords: Classification, Clustering, Logistic regression, SVM.

I. Introduction

In today's era enormous amount of data is being generated from the various sources of different sectors to handle this type of data and to gain something fruitful out of this data is a very challenging task [1-2]. Here the traditional approaches are not useful as they do not possess the capability to handle such kind of things. In this regard machine learning and data mining plays a important role and act as a helping hand for the various stakeholders of different fields. If we specifically talk about the applicability of these techniques in the healthcare sector then they are attaining new heights by providing facilities like early detection of disease, providing better medical aid, best treatment methods and cost efficient treatments [3-4].

II. Machine Learning and Data Mining in Healthcare

Number of chronic disease like kidney failure, cancer, heart disease are spreading worldwide with a rapid rate due to this the mortality rate is also increasing for any developing nations. To control the mortality rate due the chronic diseases various methods of machine learning are turning helpful by providing the recommendation systems, predicting the outcome of the disease at a very early stage [5]. This will help the medical practitioners in proper diagnosis and to detect the symptoms of the disease in at initial stage [6]. Generally there are two modeling approaches associated with these techniques:

A. Descriptive Modeling

This approach is mainly focused towards identifying the different patterns of data with the help of various clustering and association rule technique.

B. Predictive Modeling

This approach is mainly used for prediction purpose by analyzing the different datasets through classification and regression.

III. Different Methods of Data Mining and Machine Learning

A. Clustering

It is a type of unsupervised learning which is generally used to club the objects of similar type in a group called "clusters". It partitions the data by grouping the similar type of elements and forms a tighter relationship between the objects in order to get the effective results [7]. K-Means is the most widely used clustering algorithm, it is a distance based clustering algorithm where distance is the measure to check the similarity between the objects means smaller will be the distance greater will be the similarity. In K-Means algorithm the number of clusters is to be pre-specified before the use of algorithm. It works mainly in two parts in the first part the random selection of centroids is done for getting a constant value of K and in the second part data point to the closest center is assigned.

B. Classification

It is a type of supervised learning approach which is mainly used to predict the outcome by classifying the data into predefined classes [8]. Various methods of classification are as follows:

- *Naïve Bayes*

  It is a classification method which provides results in probabilistic manner and it is generated from the Bayes theorem. This method provides a form of independence between the different attributes of the dataset [9]. This method is considered as the optimal method as it is helpful in reducing the generalized errors and its ability to build the classification model in lesser time and with high effectiveness and efficiency [10-11].

  $$P(A|B) = P(B|A)*P(A)/P(B) \quad (1)$$

- *J48*

  It is decision tree based classification method which builds the decision tree from the labeled training data. This classifier works on the concept that each attribute plays an important role in decision making and it is achieved by dividing them into smaller subsets. In J48 the attributes are selected on the basis of their information gain priority. J48 can handle the missing value and has the ability to manage the categorical and continuous attributes in order to generate an efficient decision tree [12-13].

- *Logistic Regression*

  This method comes under statistical regression and it is capable of forming relationship between variables which are dependent and independent as well. It has a wide applicability in various fields like health, image processing etc. There are two categories of logistic regression generally exists Binomial and Multinomial. In the prediction field where the value is either 0 or 1 binary logistic regression is helpful. Multinomial regression is used where the dependent variable has two or more categories [14]. This can be stated as:

  $$\text{Log}\,[p/1\text{-}p] = \beta_0 + \beta_1 X_1 + \ldots\ldots\ldots + \beta_n X_n^{\,2} \quad (2)$$

  Where β_0 is the constant and $\beta_1, \beta_2 \ldots\ldots \beta_n$ are the coefficient of regression $X_1, X_2 \ldots\ldots X_n$ are predictor variables.

- *SVM*

  SVM stands for support vector machine it is the supervised learning method which is widely used for the recognition of patterns, classification tasks etc. In this algorithm a hyper plane is created in the multidimensional space of two classes. This is helpful in reducing the classification error by training the tuples and margins which are generally called as support vectors [15].

- *Random Forest*

  This method serves two ways classification and regression and it is a part of ensemble learning. Decision tree acts as the base classifier for this type of method where an identical distribution of random vector is performed. This method is followed by a voting process to get most popular class of input X. Generalized error is attained on the basis of below listed parameters for which upper bound is derived [16].

  1. Individual accuracy of classifier.
  2. Form of dependency between classifiers.

- *K-NN*

  This type classifier is used to identify the characteristics of individual category in advance. It examines the total distance between the features to be classified and the features which already classified. It uses Euclidean distance to measure the distance between points [17-18].

- *Bayes Net*

  It is also called DAG i.e. Direct Acyclic Graph which contains the nodes for the variables which are random in nature and the vertex to show the connection between the variables. In this method each node has local probability distribution function which is decided from the parent state [19].

IV. Conclusion

The role of data mining and machine learning in the health care sector is very crucial especially in the prediction for the various chronic diseases. This will help the various stakeholders of this field to take decision on real time basis. Also the models developed from the various methods of machine learning can be useful for the doctor's to diagnose the disease with more precision and in a timely manner in order to save the life of patients.

References

[1] N. Jothi, N. A. A. Rashid, and W. Husain, "Data mining in healthcare – A review," *3rd Information System International Conference Procedia Computer Science*, vol. 72, pp. 306-313, 2015.

[2] H. Wu, S. Yang , Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics in Medicine Unlocked*, vol. 10, pp. 100-107, 2018.

[3] G. Ogbuabor, and F. N. Ugwoke, "Clustering algorithm for a healthcare dataset using silhouette score value," *International Journal of Computer Science & Information Technology*, vol. 10, no. 2, pp. 27-37, 2018.

[4] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *6th International Symposium on Frontiers in Ambient and Mobile Systems Procedia Computer Science*, vol. 83, pp. 1064-1069, 2016.

[5] H. K. K. Zand, "A comparative survey on data mining techniques for breast cancer diagnosis and prediction," *Indian Journal of Fundamental and Applied Life Sciences*, vol. 5, pp. 4330-4339, 2015.

[6] J-J. Yang, J. Li, J. Mulder, Y. Wang, S. Chen, H. Wu, Q. Wang, and H. Pan, "Emerging information technologies for enhanced healthcare," *Computers in Industry*, vol. 69, pp. 3-11, 2015.

[7] G. Gan, C. Ma, and J. Wu, "Data clustering theory algorithm and application," *First ed. ASA-SIAM,* 2007.

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, no. 3, pp. 37-54, 1996.

[9] M. Jan, A. A. Awan, M. S. Khalid, and S. Nisar, "Ensemble approach for developing a smart heart disease prediction system using classification algorithms," *Research Reports in Clinical Cardiology*, vol. 9, pp. 33-45, 2018.

[10] X. Wu, V. Kumar,...and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1-37, 2007.

[11] S. K. Yadav, B. Bharadwaj, and S. Pal, "Data mining applications: A comparative study for predicting student's performance," *International Journal of Innovative Technology & Creative Engineering*, vol. 1, no. 12, pp. 13-19, 2012.

[12] Y. Chauhan, and J. Vania, "J48 classifier approach to detect characteristic of Bt cotton base on soil micro nutrient," *International Journal of Computer Trends and Technology*, vol. 5, no. 6, pp. 305-309, 2013.

[13] V. Chaurasia, S. Pal, and B. B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *Journal of Algorithms & Computational Technology*, vol. 12, no. 2, pp. 119-126, 2018.

[14] A. Wang, N. An, Y. Xia, L. Li, and G. Chen, "A logistic regression and artificial neural network-based approach for chronic disease prediction: A case study of hypertension," *IEEE International Conference on Internet of Things (iThings 2014), Green Computing and Communications (GreenCom2014), and Cyber-Physical-Social Computing (CPSCom 2014)*, pp. 45-52, 2014.

[15] G. R. Kumar, G. A. Ramachandra, and K. Nagamani. "An efficient prediction of breast cancer data using data mining techniques," *International Journal of Innovations in Engineering and Technology*, vol. 2, no. 4, pp. 139-144, 2013.

[16] E. Goel, and Er. Abhilasha, "Random forest: A review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 1, pp. 251-257, 2017.

[17] K. Sabancı, and M. Koklu, "The classification of eye state by using KNN and MLP classification models according to the EEG signals," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 3, no. 4, pp. 127-130, 2015.

[18] Z. Yong, L. Youwen, and X. Shixiong, "An improved KNN text classification algorithm based on clustering," *Journal of Computers*, vol. 4, no. 3, pp. 230-237, 2009.

[19] N. Cruz-Ramírez, H. G. Acosta Mesa, H. Carrillo Carvet, L. A. Nava Fernandez, and R. E. Berientos Martinez, "Diagnosis of breast cancer using Bayesian networks: A case study," *Computers in Biology and Medicine*, vol. 37, no. 11, pp. 1553-1564, 2007.