Machine Learning Classification Algorithms for Predictive Analysis in Healthcare

Ms. Manjiri Mahadev Mastoli[1], Dr. Urmila R. Pol[2], Rahul D. Patil[3]

[1]Research Scholar, Department of Computer Science, Shivaji University, Kolhapur, Maharashtra, India.
[2]Assistant Professor, Department of Computer Science, Shivaji University, Kolhapur, Maharashtra, India.
[3]Quality Assurance Engineer, Menon Bearing Ltd, Kolhapur, Maharashtra

Abstract – *Machine Learning and Artificial Intelligence has gained much attention from researchers in healthcare and medical sciences. Today volume, velocity, and variety of healthcare data rapidly increased, therefore there is a need of an efficient machine learning tool which will enhance prediction accuracy in healthcare. The main purpose of this paper is to find the best and most suitable algorithm for prediction and diagnosis of diseases and application of machine learning for heathcare systems. This paper also provides an overview of the data science concepts from data mining technique to machine learning classification algorithms.abstract summarizes, in one paragraph (usually), the major aspects of the entire paper in the following prescribed sequence.*

Key Words: Data Mining Technique, Machine Learning, Artificial Intelligence, Classification, Heathcare.

1. INTRODUCTION:

The increasingly growing number of applications of machine learning in healthcare allows us to foretaste at a future where data, analysis, and innovation work hand-in-hand to help countless patients without them ever realizing it. Soon, it will be quite common to find ML-based applications embedded with real-time patient data available from different healthcare systems in multiple countries, thereby increasing the efficacy of new treatment options which were unavailable before.

Healthcare covers detailed processes of the diagnosis, treatment and prevention of disease, [1]. The medical industry in most countries is evolving at a rapid space. The healthcare industry with rich data as they generate massive amounts of data, including electronic medical records, administrative reports and other findings [2].

Health informatics are becoming a very research-intensive field and the largest consumer of public funds. With the occurrence of computers and new algorithms, health care has seen an increase in computer tools and could no longer ignore these emerging tools. This resulted in uniting of healthcare and computing to form health informatics. This is expected to create more efficiency and effectiveness in the health care system, while at the same time, improve the quality of healthcare and lower cost. [3] Machine learning and Deep Learning algorithms will enhance prediction accuracy of any heathcare problem to a upper limit as compared to existing researches.

2. MACHINE LEARNING:

Traditionally data mining is a statistical learning approach and more effective, robust features for data analysis. From those data, it builds prediction or clustering models. There are lots of challenges on both pre-processing of complicated data and domain knowledge expertise. The latest advances in machine learning technologies provide new effective paradigms to obtain end-to-end learning machine learning models from complex data. In this research article, researchers review the literature on applying machine learning technologies to advance the health care domain.

As compared to several typical prediction algorithms, the prediction accuracy of machine learning algorithm reaches maximum and with a convergence speed which is faster than any other disease risk prediction [4]. Machine learning has resulted in important contributions to a number of disciplines in current years, with vision and natural language processing [5].

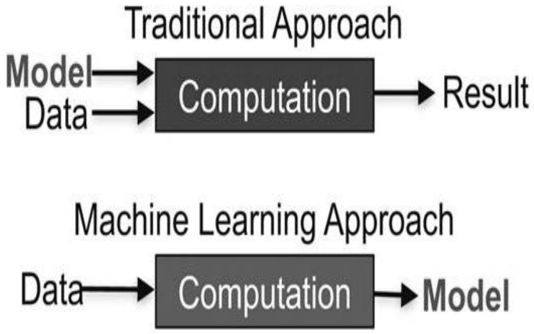

Figure 1: Traditional vs. machine learning approach. In a traditional approach to data analysis, one starts with the model as input to the machine. In an machine learning approach, one starts with the data and outputs a model that can then be applied to new data [5].

Data Mining Techniques:

2.1 Association:

In the *association* data mining technique pattern is discovered based on a relationship between items in the same operation. It is also known as relation technique. The association technique is used in market basket analysis. Also can be used in the association of some diseases and the symptoms of those or inter linkages between the diseases. It is also technique also used in crosses marketing, catalog design, loss-leader analysis, etc.

2.2 Classification:

It is a classic technique based on machine learning method of *classification* which is used to classify item in a set of item into one of a predefined set of clusters or group. It also uses mathematical techniques such as decision trees, linear programming, neural network, and statistics. Following are the various classification algorithms used in healthcare [5]:

K-Nearest Neighbour (K-NN)

Decision Tree (DT)

Support Vector Machine (SVM)

Neural Network (NN)

Bayesian Methods

Breast cancer is one of the dangerous diseases in women. Potter et al. has performed experiment on the breast cancer data set using Weka tool and then analyze the performance of different classifier using 10-fold cross validation method [6]. Data mining can contribute with important benefits to the blood bank sector. J48 algorithm and WEKA tool have been used for the complete research work. Classification rules performed well in the classification of blood donors, whose accuracy rate reached 89.9% [7].

2.3 Clustering:

Data mining technique, *clustering* is that makes a meaningful or useful cluster of objects which have similar properties or characteristics. The clustering technique describes the classes and objects in each class, while in the classification techniques, objects are allotted into predefined classes. It is a common descriptive task in which it seek to identify a finite set of categories or class or clusters to describe the data [8]. Rui Velosoa et. al. [9] had used the clustering approach, vector quantization method for predicting the readmissions in intensive medicine. The algorithm used in the vector quantization method is k-means.

2.4 Prediction:

Data mining technique *prediction*, as its name implied, that discovers the relationship between independent variables and relationship between dependent and indpendent variables. Adebayo Peter Idowu et al worked in predicting immunize-able diseases. The data mining model namely Mathematical Model (MM) for predicting immunize-able diseases that affect children between ages 0 - 5 years was designed by this team. The model was adapted and deployed for use in six (6) selected localized areas within an Osun State in Nigeria. [10]

2.5 Sequential Patterns:

Data mining technique sequential patterns analysis is seeks to discover or identify similar patterns, regular events or trends in transaction data over a period.

2.6 Decision Tree

The decision tree is most commonly used data mining techniques because its model is easy to understand for users. A decision tree is a structure includes a root node, branches, and leaf nodes. Data mining approaches that have been utilized for breast cancer diagnosis and prognosis Decision tree is found to be the best predictor with 93.62% Accuracy on benchmark dataset and also on SEER data set[11].

The researcher Jayanthi Ranjan presented how data mining discovers and extracts useful patterns of large data to find observable patterns. They demonstrates the ability of Data mining in improving the quality of the decision-making process in the pharma industry [12]. Jyoti Soni et al accomplished that decision trees, appears to be mainly effective for predicting patients with no heart disease compared to the other two models with 89% acuracy [14].

3. MACHINE LEARNING CLASSIFICATION ALGORITHMS

Following are the main five classifications of algorithm used for analysis in heathcare .

3.1 Classification:

Classification is technique to classify data into a desired and individual number of classes where we can assign labels to each class. Healthcare Diagnosis, Speech recognition, Handwriting recognition, Biometric identification, Document classification, etc are the applications of Machine Learning Classification. There are two categories of classifiers *Binary classifiers and Multi-Class classifiers.* The classification as two distinct classes or with two possible outcomes called as *Binary classifiers*. Classification with more than two distinct classes called as *Multi-Class classifiers.*

3.2.1 Naive Bayes or Naive Bayes Classifier:

Naive Bayes is a probabilistic classifier stimulated by the Bayes theorem. With the simple assumption that attributes are conditionally independent. The main advantage of Naive

Bayes algorithm requires a small amount of training data to estimate the necessary parameters. This classifiers are extremely fast compared to more sophisticated methods.

Naive Bayes Classifier appear to be most efficient as it has the highest percentage of correct predictions rate with 86.53% accuracy for patients with heart disease, followed by Neural Network and Decision Trees[14].

3.2.2 Support Vector Machine:

Support vector machine commonly called as SVM is a representation of the training data as points in space separated into categories. A clear gap that is as wide as possible. New scenario are then mapped into that same space and predicted to belong to a category based on which side of the gap. There are three main parameters Type of kernel, Gamma value, C value. The main advantages of svm is, it is efficient in high dimensional spaces and uses a subset of training points in the decision function. It is also memory efficient.

3.2.3 K-NEAREST NEIGHBOUR (KNN):

K-Nearest Neighbour classify an item into majority vote of the item's neighbors, in the space of input parameter. The object is assigned to the class which is most common among its "k" nearest neighbor which is an integer part. This classification is computed from a simple majority vote of the k nearest neighbors of each point.

Advantages: K-Nearest Neighbour algorithm is simple to implement, robust to noisy training data and effective if training data is large.

3.2.4 DECISION TREE

Decision Tree makes decisions with tree-like model. It splits the sample into two or more homogeneous sets called leaves which based on the most significant differentiators in input variables. To choose a differentiator or predictor, the decision tree algorithm considers all features and does a binary split on them for categorical data. It will then choose the one with the least cost i.e. highest accuracy, and repeats recursively, until it successfully splits the data in all leaves or reaches the maximum depth. The main advantage of the decision tree is,it is simple to understand and visualize, requires little data preparation, and can handle both numerical and categorical data.

3.2.5 RANDOM FOREST

Random forest (RF) is a collective model that grows multiple tree and classify objects based on the votes of all the trees. RF classifier is a meta_estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size, but the samples are drawn with replacement. It could handle large data set with high dimensionality, output Importance of Variable, useful to explore the data and it could handle missing data while maintaining accuracy. Advantage of random forest is reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases.

Stephan Dreiseitl et al evaluated the discriminatory power of *K*-nearest neighbors, artificial neural networks (ANNs), logistic regression, support vector machines (SVMs) and decision tress on the task of classifying pigmented skin lesions (nevi), dysplastic nevi OR melanoma[15]. N. G. Maity et al made case study analysis on machine learning they demonstrates the use of Bayesian Inference diagnosing Alzheimer's disease based on cognitive test results and demographic data and focused on programmed classification of cell images to find out the advancement and severity of breast cancer using ANN[16].

Many machine learning applications are found in the medical related areas such as Identifying Diseases and Diagnosis, Drug Discovery and Manufacturing, Medical Imaging Diagnosis, Personalized Medicine and Outbreak Prediction. Machine learning applied in healthcare industry play an important role in prediction and diagnosis of the diseases. The respective study will find the useful and hidden knowledge from the dataset.

3. CONCLUSION

In this paper, researcher reviewed research papers on machine learning methods applied to healthcare applications. Decision Tree and Support Vector Machine are the machine learning classification algorithm used by the majority of researchers in their heathcare predictive research and are the best algorithm in case of accuracy. Machine Learning and Artificial Intelligence have virtually endless applications in the healthcare and medical domain. Machine learning is helping to streamline administrative processes, diagnosis diseases, prognosis diseases, treatment schedule and personalize medical treatments in healthcare to map and treat diseases. In future, there may be more advanced Machine Learning technique focused on the early detection, diagnosis, prognosis of diseases. Based on the analyzed work, researchers suggest that AI and Machine Learning approaches could be the vehicle for translating big heathcare data into improved human health.

REFERENCES

[1] J.-J. Yang, J. Li, J. Mulder, Y. Wang, S. Chen, H. Wu, Q. Wang, and H. Pan, "Emerging information technologies for enhanced healthcare," Comput. Ind., vol. 69, pp. 3–11, 2015.

[2] N. Wickramasinghe, S. K. Sharma, and J. N. D. Gupta, "Knowledge Management in Healthcare," vol. 63, pp. 5–18, 2005

[3] Shortliffe, EH., Perrault, LE., (Eds.). Medical informatics: Computer applications in health care and biomedicine (2nd Edition). New York: Springer, 2000.

[4] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017. doi: 10.1109/ACCESS.2017.2694446

[5] Jenna Wiens, Erica S Shenoy, Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology, Clinical Infectious Diseases, Volume 66, Issue 1, 1 January 2018, Pages 149–153, https://doi.org/10.1093/cid/cix731

[6] S.Yamini , Dr.V.Khanaa , Dr.Krishna Mohantha - A State of the Art Review on Various Data Mining Techniques, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, Issue 3, March 2016

[7] R. Potter, “Comparison of classification algorithms applied to breast cancer diagnosis and prognosis”, advances in data mining, 7th Industrial Conference, ICDM 2007, Leipzig, Germany, pp. 40-49, (2007) July.

[8] Arvind Sharma and P.C. Gupta—Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool International Journal of Communication and Computer Technologies Volume 01 – No.6, Issue: 02 September 2012.

[9] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” AI Mag., pp. 37–54, 1996.

[10] R. Veloso, F. Portela, M. F. Santos, Á. Silva, F. Rua, A. Abelha, and J. Machado, “A Clustering Approach for Predicting Readmissions in Intensive Medicine,” Procedia Technol., vol. 16, pp. 1307–1316, 2014.

[11] International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 5– No.7,May 2013.

[12] ShwetaKharya—Using Data Mining Techniques ForDiagnosis And Prognosis Of Cancer Diseasel, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.

[13] Jayanthi Ranjan—Applications of data mining techniques in pharmaceutical industry||, Journal of Theoretical and Applied Technology, (2007).

[14] Jyoti Soni, Ujma Ansari, Dipesh Sharma, ,Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011

[15] Stephan Dreiseitl, LucilaOhno-Machado,Harald Kittler, Staal Vinterbo,Holger Billhardt, Michael Binder,A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions ,Journal of Biomedical Informatics,Volume 34,Issue 1,February 2001,Pages 28-36.

[16] N. G. Maity and S. Das, "Machine learning for improved diagnosis and prognosis in healthcare," *2017 IEEE Aerospace Conference*, Big Sky, MT, 2017, pp. 1-9.

BIOGRAPHIES

Ms. Manjiri Mahadev Mastoli pursed Master of Compter Application from CSIBER, Shivaji University, Kolhapur, Maharashtra. She is currently pursuing Ph.D. and currently working as research Scholar in Department of Computer Sciences, Shivaji University, Kolhapur, Maharashtra. Her main research work focuses on Artificial Intelligence and Machine Learning, Data Science, She has 5 years of teaching experience and 5 years of Research Experience.

Dr.Urmila Pol pursed Bachelor of Science from Shivaji University and Master of Compter Application from CSIBER, Shivaji University, Kolhapur, Maharashtra. She is currently working as Assitant Professor in Department of Computer Sciences, Shivaji University, Kolhapur, Maharashtra. She is Member of International Association of Engineer. Currently, she is guiding four Ph.D and one M.Phil student. Her main research work focuses on Artificial Intelligence, Data Science, Big Data Analytics, LMS and Open Source Technologies, She has 20 years of teaching experience and 15 years of Research Experience.

Mr. R.D. Patil pursed Bachelor of Enginerring Shivaji University, Kolhapur and Master of Buniness Administration from IGNOU. He is currently working as Quality Assurance Engineer at Menon Bearing Ltd, Kolhapur, Maharashtra. His main research work focuses on Oprational Research, Artificial Intelligence and Machine Learning. He has 8 years of Quality Assurance experience and 4 years of Research Experience.