MACHINE LEARNING ALGORITHMS IN HEALTH CARE SYSTEMS

[1]M. Praveena Kirubabai , [2]Dr. G Arumugam

Research Scholar , Senior Professor and Head of the Department (Retd) , Department of Computer Science,
Madurai Kamaraj University, Madurai 21, TamilNadu ,India

ABSTRACT

Recent years Machine learning Techniques with their superior performance implements varieties of health care applications through Computer Aided Diagnosis with Multidimensional medical images and Databases. Machine learning is a system that learns from the data rather than algorithms. A mathematical model is generated by training the system with the data to produce more precise outputs and predictive outcomes. This paper discusses the overview of the machine learning algorithms implemented in the Health care system and presents a potential method to work with Health Care applications.

Key Words: Machine learning, Computer Aided Diagnosis, Medical images

I INTRODUCTION

Since Today we are surviving in the era of algorithms, where Machine learning and deep learning systems transform multiple industries in the field of manufacturing, Governance and Transportation .ML algorithms in Social Media and health care system are inseparable from our daily routines. Machine learning algorithm implements various techniques and tools that help the detection and diagnosis of Medical Applications. Machine learning algorithm analyses clinical parameters for predicting the diseases, extracting the medical knowledge and the detection of diseases. Machine learning algorithm assists the physicians to identify the abnormalities of the patients leading to the accurate detection of tumors and cancers. In this paper we have discussed a comprehensive literature review on various machines learning algorithms in health care systems with their advantages and also highlighting the significance of Random Forest Algorithm.

II CATEGORIES OF MACHINE LEARING ALGORITHM

In this section we discuss about the categories of Machine learning algorithm. They are supervised learning, unsupervised learning and Reinforcement learning.

Supervised learning model

It's a mathematical model contains both the input and the output. The data's are labeled that provides a special meaning to it. An optimal function allows the algorithm to correctly determine the desired output based on the inputs.

The supervised algorithm can be grouped into

Classification: Classification is implemented when the output data is of type categorical.

Example: Support Vector Machine

Regression: Regression is implemented when the output data is of type real value.

Example: Linear Regression

Example: Random Forest for both regression and classification.

Unsupervised learning model

It a model that has only inputs learn from the test data not been labeled. It's an iterative pattern works well for large dataset based on clustering and pattern recognition.

The Unsupervised algorithm can be grouped into

Clustering: It is the grouping of data based on some similarities.

Example: K-means algorithm

Association Rule: It's based on discovering rules for processing large set of data.

Example: Apriori Algorithm

Reinforcement learning

It's a model that learns from the behavior of the environment. The system performs trial and error method rather than learning from the data. It helps taking decisions sequential.

Example: Chess Game.

Support vector machine

It's a supervised machine learning algorithm for classification of data. The kernel trick is used to transform the data where it finds the optimal boundary between the possible outputs.

Linear regression

Linear regression model s the relationship between two variables namely the input and the output variables using a linear function. It was widely used in the field of statistics and now implemented in machine learning.

Random Forest

Random Forest is the way to machine learning algorithm that works through bagging approach to create a bunch of decision trees with a random subset of the data. It is considered to be one of the most effective algorithms to solve almost any prediction task. It can be used both for classification and the regression kind of problems.

Random Forest Machine Learning Algorithm maintains accuracy even when there is inconsistent data and is simple to use. It also gives estimates on what variables are important for the classification. It runs efficiently on large databases while generating an internal unbiased estimate of the generalisation error.

K-means algorithm:

K-means algorithm is an iterative algorithm partitions the dataset into predefined non overlapping subgroups. K-means popularly used in market segmentation, document clustering, image segmentation and image compression.

Apriori Algorithm:

Apriori algorithm is a classification algorithm for mining frequent item sets. It's very useful in Master Basket Analysis.

III REVIEW OF LITERATURE

Support vector machine is one of the popular tools for machine learning tasks that involve classification and regression. The optimal classifier paper presented an automatic diagnostic system to classify the patients of breast cancer by implementing a non linear optimal classifier with SVM. They developed CAD system that improved the performance by reducing the misdiagnosis and the time required to implement a diagnosis [1].This paper proposed an approach to classify the persons with and without common diseases. The variables selected were 14 related to the common risks involved in diabetes. The variable selection was done using an automatic approach with special nonlinear function called kernels to transform the input into multidimensional space. The SVM algorithm was implemented for classification and the test data was given as the input. A 10-fold cross-validation was implemented to measure the robustness of the estimate. The SVM method was efficient without any assumptions related to the distribution and independency of the data. The results provided them an effective data analysis preprocessing and preprocessing algorithms [2].The purpose of this research is to develop a diagnostic method of diabetes based on standardized tongue image using support vector machine (SVM). The training samples were collected and the extracted features are tested and classified using SVM classifier with Optimization of Kernel Parameters by *GA* [3].The authors proposed an SVM model for the data classification of Heart Failure patients. This study aims to identify predictors of medication adherence in HF patients. Mathematical simulations were performed for the identification of variables that would best predict medication adherence. The leave-one-out cross-validation (LOOCV) was experimented on the data set to evaluate the robustness. The accuracy was around 77.6%[4].This paper was proposed to detect a correlation between the personal medical expenses using linear regression model s thereby comparing with ANOVA . The prominent predictors named smoking , higher BMI and smoking had an high impact on correlation with greater medical expenses contributing to high medical expenses.

Since there was more than one dependent variable multiple regressions was used to create different models. The prediction accuracy obtained was more than 75% accuracy charges [5].In this paper data mining techniques were used to detect the risk factors of the patients susceptible to the infections based on the characteristics, treatments and invasive device in ICU. Under sampling and oversampling are the major challenges in data analysis .This drawback was handled by Random forest algorithm [6]. This paper proposed a ranking based methodology to perform feature selection. The features were verified using feature ranking algorithm through 10-fold cross validation. The Random forest is applied to train and test the dataset. Extensive experiments were conducted with other classifiers and the Random Forest outperformed better than other classifiers [7]. The author proposed a novel approach called multiple-criteria decision analysis for extracting association rules from medical records. Natural Language Processing techniques are used along with Data Mining algorithms for extracting association rules. This discovers the association between diseases, diseases and symptoms, diseases and medicines. Earlier detection and prevention of disease might be performed. Similarly medications were given for various diseases might be retrieved using medicine → disease rules [8].This paper proposes the implementation of Apriori Algorithm on various healthcare datasets using machine learning tool Weka. Association rule mining identifies trends and patterns from large databases. The work analyses the different results by implementing the Association Apriori Algorithm in Weka and compares the performance with other algorithms. The results obtained achieve better results than Predictive Apriori Algorithm and Tertius Algorithm [9].The proposed paper works on Clustering of the Healthcare data employing assorted Kmeans Algorithm below weka toolkit. We have discovered that the Kmeans works well for clustering the healthcare data[10]**.**The study was conducted on medical record data of patient in RSUP Haji Adam Malik user of Security and Healthcare Security. The K-Means Clustering method was implemented on the data set and the algorithm is fairly easy to implement and run, relatively fast, easy to customize and widely used[11].The amount of data generated by media sensors in health monitoring systems during medical diagnosis are too complex and voluminous to be processed and analyzed by traditional methods. Data mining approaches offer transform these heterogeneous data into meaningful information for handling decision. The paper implemented the *k*-means clustering algorithms on large datasets and the proposed algorithm, called *G*-means, utilizes a greedy approach to create

the preliminary centroids and then takes *k* or lesser passes over the dataset to adjust these center points. Our experimental results show that *G*-means outperforms *k*-means in terms of entropy and *F*-scores. The experiments also yield better results for *G*-means in terms of the coefficient of variance and the execution time [12].This paper proposed an improved k-means algorithm, requires the pre-estimated number of clusters, k, which is the same to the standard k-means algorithm. For optimal solution, test the different value of k. The paper used three different data sets from the UCI repository of machine learning databases to measure the efficiency of the improved k-means algorithm and the standard k-means. In both the experiments, the time taken for execution of clusters and efficiency is computed. The Experimental results show that the improved algorithm shows improved execution time and is feasible[13].

IV ADVANTAGES AND LIMITATIONS OF MACHINE LEARNING ALGORITHM

The advantages and the limitations of the algorithm are discussed as follows

Algorithms	**Advantages**	**Limitations**
Linear Regression	The performance is good with small dataset.	Data Assumptions are needed to be compiled
Support Vector Machine	Works with non-linear solutions	For better performance the knowledge about the kernel should be known
Random Forest	It overcomes the problem of over fitting. It's extremely flexible with very high accuracy.	Not easily interpretable
K-Means	Easy to implement	Difficult to predict the number of clusters
Apriori Algorithm	Easy to implement	Very slow

V SIGNIFICANCE OF RANDOM FOREST

Impressive in Versatility

Either regression or classification task, random forest is an applicable model handles binary features, categorical features, and numerical features. There is very little pre-processing that needs to be done. The data does not need to be rescaled or transformed.

Parallelizable

They are parallelizable, meaning that we can split the process to multiple machines to run. This results in faster computation time. Boosted models are sequential in contrast, and would take longer to compute.

Great with High dimensionality

Random forests are great with high dimensional data since we are working with subsets of data.

Quick Prediction/Training Speed

It is faster to train than decision trees because we are working only on a subset of features in this model, so we can easily work with hundreds of features. Prediction speed is significantly faster than training speed because we can save generated forests for future uses.

Robust to Outliers and Non-linear Data

Random forest handles outliers by essentially binning them. It is also indifferent to non-linear features.

Handles Unbalanced Data

It has methods for balancing error in class population unbalanced data sets. Random forest tries to minimize the overall error rate, so when we have an unbalance data set, the larger class will get a low error rate while the smaller class will have a larger error rate.

Drawbacks

1. Model interpretability: Random forest models are not all that interpretable; they are like black boxes.
2. For very large data sets, the size of the trees can take up a lot of memory.

VI CONCLUSION

The use of Machine learning algorithm has a great potential to transform health care systems. In this paper we have discussed about the various Machine learning algorithms with the overview of pros and cons. We also mentioned about the significance of Random Forest. In future Random forest can be implemented in image classification for various medical Applications.

BIBLIOGRAPHY:

1. Wei Yu*, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, Muin J Khoury, Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes,BMC Medical Informatics and Decision Making,16,2010.

2. Jianfeng Zhang,1 Jiatuo Xu,1 Xiaojuan Hu,2 Qingguang Chen,3 Liping Tu,2 Jingbin Huang,1 and Ji Cui1, Diagnostic Method of Diabetes Based on Support Vector Machine and

Tongue Images,Machine learning in multi modal medical imaging,Jan 2017

3. Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients, Youn-Jung Son, RN, Hong-Gee Kim, Eung-Hee Kim, ME2, Sangsup Choi, Soo-Kyoung Lee, Health Care Informatics Research, September 2010

4. Linear regression model for predicting medical expenses based on insurance data Akhil Alfons Kodiyan, Kirthy Francis, Engineering and Computing, Dec 2019

5. María N. Moreno García, Juan Carlos Ballesteros Herráez, Mercedes Sánchez Barba and Fernando Sánchez Hernández, Random Forest Based Ensemble Classifiers for Predicting Healthcare-Associated Infections in Intensive Care Units, Distributed Computing and Artificial Intelligence,June 2016

6. Md. Zahangir Alam1, M. Saifur Rahman, M. Sohel Rahman, A Random Forest based predictor for medical data classification using feature ranking, Informatics in Medicine Unlocked, 2019

7. Extracting Association Rules from Medical Health Records Using Multi-Criteria Decision Analysis, Science Direct**,** Procedia Computer Science 115 (2017) 290–295

8. Divya Jain, Sumanlata Gautam, Implementation of Apriori Algorithm in Health Care Sector: A Survey, Semantic Scholar, 2013

9. Jitendra Kumar Samriya, Sachin Kumar, Sunil Singh, EFFICIENT K-MEANS CLUSTERING FOR HEALTHCARE DATA, Advanced Journal of Computer Science and Engineering (AJCST), 2393-8390 (O) 4, 2016

10. Clustering of Patient Disease Data by Using K-Means Clustering Parasian D.P Silitonga

Teknik Informatika, Universitas Katolik Santo Thomas Sumatera Utara Jln. Setia Budi No. 479-F Medan, Sumatera Utara, Indonesia, International Journal of Computer Science and Information Security (IJCSIS),Vol. 15, No. 7, July 2017

11. Ramzi A. Haraty, Mohamad Dimishkieh, and Mehedi Masud, An Enhanced *k*-Means Clustering Algorithm for Pattern Discovery in Healthcare Data, International Journal of Distributed Sensor Networks, Hindawi Publishing Corporation, International Journal of Distributed Sensor Networks, 2015