INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB
e-ISSN : 2582 - 4376
Open Access
RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Special Issue of First International Conference on Advancements in Engineering, Management, Science & Technology (ICEMST 2021)

A Survey: Effective Machine Learning Based Classification Algorithm for Medical Dataset

G.Mahalakshmi[1], ShimaaliRiyasudeen[2], Sairam R[3], Hari Sanjeevi R[4]and B. Raghupathy[5]
Teaching Fellow[1], Department of Information Science and Technology, Anna University Chennai
[2,3,4,5] MCA, Department of Information Science and Technology, Anna University Chennai
mlakshmig27@gmail.com[1]

Abstract

Machine Learning is defined as nothing but as like humans learn from their experience likewise machine learns from experience. Here experience is nothing but the training and testing the machine with the. There are many techniques are available to train and test the system like Data Mining algorithms, Machine Learning Algorithms and Deep Learning Algorithms. It is not that all the algorithms will provide better results. Also there are many kinds of datasets is available. In this paper, the main focus is on Medical Dataset which requires more attention nowadays. And the algorithm we focus is on Machine Learning which contains different classification algorithms. Applying all the classification algorithms to the dataset and finding the best algorithm for the medical dataset with the highest accuracy. We have trained the dataset using seven classification algorithms called Naive Bayes, Random Forest, Support Vector Machine, Decision Tree, KNN, KSVM. After the implementation of each algorithm, we came up with a conclusion that Random Forest is the best algorithm for medical dataset which gives 100%accuracy among these seven Classification algorithms.

Keywords: **Naive Bayes, Random Forest, Support Vector Machine, Decision Tree, and Kernel Support Vector Machine**

1. Introduction

A medical dataset has vast amount of data that includes disease surveillance, prescription drugs, hospitalization, patient insurance etc. in the following paper we are going to look at “chronic kidney disease”. Factors determining CKD are diabetes mellitus, hypertension, age, cardio vascular disease, family history of CKD. Considering this we here use machine learning algorithms to decide whether an individual has been affected by CKD based on the complications and level of the factors.

1.1 About Machine Learning

As we humans learn from our experience, likewise same Machine learns from experience. Here Experience means training given to machine. To illustrate technically, where a machine can do things by its own without the instructions given by human, that’s what we call as machine learning. It’s done only by the training we give to the model. Higher the algorithm you use the more efficient the model would be. It has some basic steps to be followed to attain a perfect model.

1.2 Dataset

The Medical Dataset is considered for our proposed work. Here we used chronic kidney disease dataset, which gives us the details of the patient health such as blood pressure, Haematoglobulin (Haemoglobin) level, WBC counts and RBC counts and so on. We will be taking some of the main features which will be related to kidney disease, and the last column of

the dataset is class, where we will be predicting the disease is there or not.

2. Literature Survey

D.G. Huang proposed [1] performance comparison between K- nearest neighbor, random forest and decision trees for classifying unknown drugs. The classification model was evaluated in terms of accuracy, recall and precision. The proposed model achieved 77.7% accuracy while using K-nearest neighbor which is better than that of decision tree and random forest with one decision tree and worse than that of random forest with 500 decisions trees. Ifthikar Ahmad proposed [2] performance comparison between random forests, support vector machine, Kernel-SVM and Extreme Learning Machine (ELM) for intrusion detection dataset. they run the dataset as full sample, half sample and ¼ th sample into the algorithms. The model achieved 99.5% accuracy while using ELM for full sample (65,335) rather than SVM, RF, Kernel SVM. Support vector machine achieved 99.2% accuracy on ¼ th samples (18,383) then the remaining algorithms whereas Kernel SVM achieved 98.6% accuracy on half sample (32,767). Random forest has the worst performing algorithm as per their approach. Krishnaveni K.S [3] proposed a faculty rating system based on student's feedback. the author uses naive Bayes algorithm to derive class from input and it uses high level dataset and build model fast .it helps in rating process. Mrs. Jayashree [4]proposed a heart disease prediction system based on user details (patient details). It is completely android application system and it is implemented only on KNN algorithm. K - Nearest neighbors collect all those information's from the user and check with all the parameters mentioned in the page and then it provides accuracy. Based on the accuracy it indicates the user (patient), whether the disease is there or not. Masashi Sekiya [5] has proposed the system which use a Linear Logistic Regression Model and estimated muscle activity from movement data. This model has been judged based on high generalizing capability. In addition to that, this method has shown more potential for recognizing a system using less volume of data when compared to ANN (Artificial Neural Network) and has exhibited the highest performance than any other model.

3. Existing Approach

In order to gain insights about the algorithms performance we focus on existing papers. In paper [2] author used random forest, support vector machine, extreme learning machine for intrusion detection. Based on the investigation they come to a conclusion that ELM outperforms other algorithms in terms of accuracy and precision on full sample data and SVM produce better results on half and ¼ samples. In paper [1] the author used random forest, decision tree and K-nearest neighbor for drug classification and has come to a termination that KNN achieved higher accuracy than other algorithms

4. Proposed Approach

The proposed work uses the classification algorithms for medical dataset to find the performance of the classification algorithms. The existing approaches are done with different machine learning algorithms and different field of dataset. We implemented classification algorithms for medical dataset and our medical dataset is labelled dataset of 2 classes. The dataset is trained with seven classification algorithms such as logistic regression, naive Bayes, random forest, decision tree; support vector machine, kernel support vector machine and K-nearest neighbor. we use this dataset to conclude the performance of each algorithm in medical dataset and which algorithm has maximum efficiency to work in medical dataset.

5. Detailed Approach

In the proposed work, a medical dataset is considered on chronic chronic kidney disease which having certain attributes regarding the disease such as age, blood pressure, white blood cells count, red blood cells count, packed cell volume etc. These attributes have only numerical values. This dataset is fed into the machine learning algorithms.

5.1 Assumption

First, assume that every feature in the dataset has a standard normal range of value. For example, if blood pressure of a person increases or decreases then the standard normal range then the person is suffering from high or low bp. Similarly, each feature has a standard normal range of value in the dataset where a person is not sick. Second, it is assumed that if the dataset has null value, then it is not possiblepredict the output.

5.2 Types of Classification Algorithms Used

In Machine Learning there are different classifications algorithms are available. Some of the most widely using classification algorithm is considered for proposed work.

5.2.1 Support Vector Machine

Support Vector Machine (SVM)is a kind of supervised learning algorithms, mostly used for classification problems.it can also use for regression cases. The working of this algorithm is to separate the data items by a line called hyper plane. here each data item is called as vector. Now the hyper plane is going to separate the vectors and find the approximate class of each vector (data items).

It is important that the hyper plane must be optimal.

For that, the data items or vectors help us to find the optimal one. How to find an optimal one? The margin with greatest possible is the optimal hyper plane. So, that the data items or vectors can be classified correctly. The data point/vectors are known as support vector. These support vectors help the hyper plane to classify the vectors approximately so that this algorithm is called as support vector machine learning algorithm.

5.2.2 Kernel SVM

Kernels are the methods that are widely using in support vector machine learning algorithm. This method used to for pattern analysis which is useful to find the general relation between the variables/features in the dataset. In short form, kernels are used to transforming one dimension to another for making a clear hyper plane to separate classes of the dataset.

5.2.3 Naïve Bayes Classifier

It is a classification algorithm which based on a Bayesian theorem which is best suits for a high dimensional dataset.

Bayesian theorem:

- It is used to calculate the probability of an given event, by considering the probability of a previous events which are happened earlier.
- Bayesian formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{Eq. (1)}$$

- The fundamental assumption of an algorithms is that every attributes will make an equal and independent contribution forall the outcomes.

The algorithm works on combination prior probability and the likelihood of data item to find the class of the data item which it belongs to. Let's consider there are two groups of data which are marked by the color red and yellow in the graph. Now a new data arrives, we have to find the class of the new data. For that we have find the prior probability of the first group that is red color as well as for the second group yellow

To calculate prior probability Prior probability as total no. of data items in group/total no. of data items in the dataset. Now for calculating the likelihood, first we have marked the new data in the graph then draw a circle around the new data. The formula is likelihood of group = no. of data items belongs to group/total no. of data items in a group. After finding this value now the posterior probability is calculated as prior probability * likelihood of group . Now the new data items belong to the group which has the greatest possible value of posterior probability

5.2.4 KNN Algorithm

KNN stands for k-nearest neighbor is a prediction algorithm that is used for classification and regression problems. This algorithm works the value of k to predict the class of the data item.

5.2.5 Logistic Regression

It is a supervised classification algorithm.As it is a classification algorithm it uses discrete Values for a given feature. It is the most famous Machine learning algorithms which stands next to linear regression. Both are similar to each other but the major difference is Linear Regression-predictLogistic Regression-To classify.

5.2.6 Random Forest

Random forest consists of group of decision trees based on random selection of data (ensemble technique). this algorithm increases the accuracy of classification.by using random forest algorithm we can avoid bagging. We will be creating lot of subsets using random values.it has a split at end of each tree helps to decorrelate. predictions from multiple trees are combined and used to make a single decision. random forest is used to reduce the overall variance.

5.2.7 Decision Tree:

- ecision tree is a classifier used to predict whether an event will occur or not
- A dataset has may attribute
- Divide the attribute into subsets

- Smaller the tree, accuracy is more
- It is a divide and conquer algorithm
- The aim is to divide the attributes into pure subsets where it cannot be divided further.

Result:

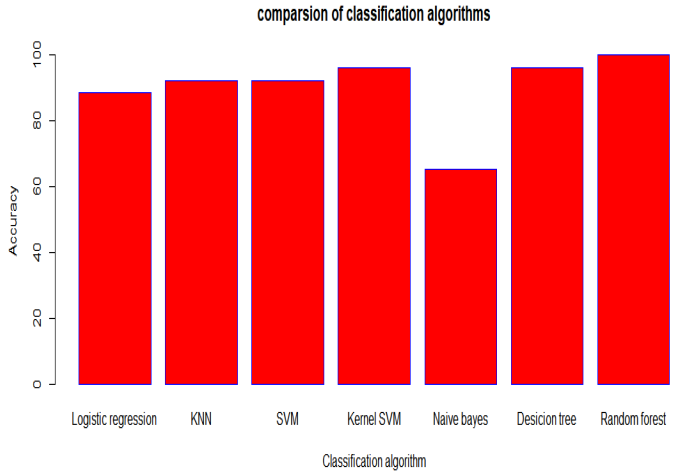

Chart 1: Comparison of Classification Algorithms

Conclusion and Future Work

In this paper, we have used machine learning based classification algorithms for a label medical dataset. Also the system is trained the dataset with ML classification algorithms and the accuracy of each algorithm is given in table 1.

Table1. Accuracy of Classification Algorithms

Algorithms	Accuracy(in percentage)
Logistic regression	88.46
K- nearest neighbor	92.31
Support vector machine	92.31
Kernel support vector machine	96.15
Naïve Bayes	65.38
Decision tree	96.15
Random forest	100

Random forest achieved 100% accuracy when compared to other classification algorithms. Therefore, it is better to use Random Forest algorithm for medical dataset. In the future work, it is aimed to apply the ML algorithms with other types of datasets and also try to apply and identify other types of classification algorithms for finding the most suitable one for the datasets.

References

Journals

[1]. International Journal of Emerging Technologies and Innovative Research, ISSN:2349-5162, Vol.6, Issue 6, page no. pp17-26, June 2019

[2]. Masashi Sekiya, Sho Sakaino, Toshiaki Tsuji," Linear logistic regression for estimation of lower limb muscle activations" journal of LATEX class file, VOL 14, no 8, august 2015.

[3]. Hinton, G.E., Krizhevsky, A., Srivastava, N., Sutskever, I., & Salakhutdinov, R, "Dropout: a simple way to prevent neural networks from overfitting", Journal of Machine Learning Research, 15, 1929-1958. 2014

[4]. Amorim, D.G., Barro, S., Cernadas, E., & Delgado" Do we need hundreds of classifiers to solve real world classification problems", Journal of Machine Learning Research, 2014.

Conference Proceedings

[5]. J Ifthikar ahmad, Mohammad basheri, Muhammad javed equal and Aneel Rahim, "Performance comparison of support vector machine, random forest & Extreme learning

machine for intrusion detection “, IEEEAccess, May 30, 2018.

[6]. Krishnaveni K.S, Rohit R Pai, Vignesh Iyar, “Faulty rating system based on student feedbacks using sentimental analysis”, International Conference on Advances in Computing, Communications and Informatics (ICACCI) 2017.

[7]. Fei-Fei, L., Karpathy, A., Leung, T., Shetty, S., Sukthankar, R., & Toderici, G, “Large-Scale Video Classification with Convolutional Neural Networks”**,** IEEE Conference on Computer Vision and Pattern Recognitiom, 2014.

[8]. R. Salakhutdinov and A. Mnih, “Bayesian probabilistic matrix factorization using Markov chain Monte Carlo”, In Proceedings of the 25th International Conference on Machine Learning. ACM, 2008.

[9]. Sermanet, S. Chintala, and Y. LeCun, “Convolutional neural networks applied to house numbers digit classification”, . In International Conference on Pattern Recognition (ICPR 2012), 2012.

[10]. L van der Maaten, M. Chen, S. Tyree, and K. Q. Weinberger, “Learning with marginalized corrupted features”, In Proceedings of the 30th International Conference on Machine Learning, pages 410–418. ACM, 2013.