A Comparative Study of Different Machine Learning Algorithms for Disease Prediction

Anantvir Singh Romana
Department of Electronics Technology, Guru Nanak Dev University, Amritsar, Punjab, India

DOI: 10.23956/ijaresse/V7I7/0177

Abstract— Accurate diagnostic detection of the disease in a patient is critical and may alter the subsequent treatment and increase the chances of survival rate. Machine learning techniques have been instrumental in disease detection and are currently being used in various classification problems due to their accurate prediction performance. Various techniques may provide different desired accuracies and it is therefore imperative to use the most suitable method which provides the best desired results. This research seeks to provide comparative analysis of Support Vector Machine, Naïve bayes, J48 Decision Tree and neural network classifiers breast cancer and diabetes datasets.

Keywords— Artificial Intelligence, Machine Learning, WEKA, Classification, Disease Detection.

I. INTRODUCTION

The development of automated diagnostics was instigated by the need to aid the physician in decision making. There application in healthcare has spanned from the electrocardiograms to ultrasounds etc. The traditional setup for error detection and monitoring of disease progression heavily rest on the technicians within the healthcare. The increase in the number of patients within healthcare who require continuous assessment has led to the technical development of the automated systems. Transformations of the qualitative information to quantitative measures are at the forefront in solving classification problems.

This paper tries to provide the solution of the above problem using different machine learning algorithms and further based on the performance of these algorithms on two medical datasets i.e. breast cancer and diabetes, we conclude which machine learning algorithm is better for disease detection at early stage.

II. LITERATURE REVIEW

The increase in the number of deaths determined within the healthcare systems has led to the development of medical diagnostic support systems to aid the medical personnel’s in decision making process [1]. Various experts systems and machine learning algorithms have been utilized to provide supporting information based on the input knowledge. Some of the significant developments include 2D and 3D medical imaging, feature extraction, pattern analysis and classification have been used in providing solutions for edge detection and region growing among other problems [1]. According to Pena-Reyes and Sipper (1999) an effective diagnostic systems should be able to provide higher accuracy of disease identification as malignant or benign. In addition, the systems should also be able to determine with a degree of confidence indicating the accuracy of diagnosis with some levels. Another major important aspect is the systems interpretability which provides information on the steps followed resulting to the outcomes generated. The Artificial neural network on the other hand has been determined to be an effective tool in classification though the operations within the network structure are hidden. Classification problem seems to have generated interests among researchers. The classification approach is used in data analysis and pattern recognition problems. This approach involves classifier modeling which is used as a function that associates a class to different attributes. The concept of association based on similarities or trained performance has been embedded in various approaches such as neural networks, decision trees, decision graphs and etc [2]. The methodology of the neural networks can be performed in two phases i.e. training and testing. The training phase involves feature extraction and computation utilizing the classification rules. On the other hand, testing data is used for performance evaluation on the accuracy of the classification process determined by the training data [1]. Breast cancer diagnosis and prognosis has instigated the research interest and has been explored utilizing various artificial neural networks such as Radial Basis Function, Multilayer perceptrons, Backpropagation, and Learning Vector Quantization network. Other methods which have been utilized to determine the breast cancer diagnosis includes Fuzzy systems and Evolutionary algorithms. The fuzzy systems are used to represents different degrees of the disease (malignant or benign) a patient suffers from; on the other hand, the evolutionary algorithms are used to perform search to determine the most suitable fuzzy systems [3]. Isotonic separation which is a linear programming technique is based on the underlying assumption of maintaining same consistency in diagnosis. For example the Breast cancer dataset (Wisconsin) a patients being diagnosed with malignant tumor based on certain characteristics of the cell structures, for other patients showing similar symptoms with more damage to the cells would end up receiving the same diagnosis [4] and Rank nearest neighbor technique (k-RNN) [5]. The k-RNN has been determined as technique used in approximating the densities based on the evaluations of the nearest neighbors [5]. The aforementioned technique has been applied in univariate and multivariate data in examining various classifications problems including breast cancer. In order for a patient to receive the appropriate breast cancer treatment, it is necessary that accurate classification of the cells be
determined. This has lead researchers to combine and employ various machine learning techniques and select the one with the highest prediction accuracy [6]. The comparative analysis of the ANN ranges from two to six networks or more being evaluated to determine the most appropriate technique. Integration of different ANN networks has lead to improve performance measures. The RBF properties when applied to tuning the SVM has been determined to provide higher prediction accuracy for breast cancer data [7].

III. DIFFERENT MACHINE LEARNING ALGORITHMS

This section discusses the different machine learning algorithms that are used by us in our experimentations.

A. Support Vector Machine (SVM): In machine learning, SVM are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

B. Naive Bayes: In machine learning, Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s, and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

C. J48 Decision Tree: It is an algorithm used to generate a decision tree developed by Ross Quinlan. It is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by it can be used for classification, and for this reason, it is often referred to as a statistical classifier.

D. Neural Network: They are computing systems inspired by the biological neural networks that constitute animal brains. Such systems learn (progressively improve performance) to do tasks by considering examples, generally without task-specific programming. For example, in image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually labelled as "cat" or "no cat" and using the analytic results to identify cats in other images. They have found most use in applications difficult to express in a traditional computer algorithm using rule-based programming.

IV. EXPERIMENTATION AND RESULTS

The aim of this paper is to find the best among the machine learning algorithms discussed in previous section for disease detection. Using these algorithms we have built four different classifiers. For this task we have used WEKA 3.8 [8]. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Further for entire experimentation we have used two medical datasets i.e. breast cancer and diabetes. These datasets have been used for training and testing purposes.

We have applied 10-fold cross validation in order to check the results of every classifier for unknown instances. The results of entire experimentation are shown in Table 1. While Figure 1, shows the pictorial representation of efficiency of these algorithms for the two datasets.

Table 1. Comparison of Different Machine Learning Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Breast Cancer Dataset</th>
<th>Diabetes Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Efficiency (%)</td>
<td>Time Taken (Milliseconds)</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>69.58</td>
<td>310</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>71.67</td>
<td>24</td>
</tr>
<tr>
<td>J48 Decision Tree</td>
<td>75.52</td>
<td>40</td>
</tr>
<tr>
<td>Neural Network</td>
<td>71.32</td>
<td>590</td>
</tr>
</tbody>
</table>

[Note: MAE stands for Mean Absolute Error]
Figure 1. Efficiency of different Machine learning Algorithms

From above results it is clear that none of the algorithms can be considered as best. As all algorithms behave differently for different datasets. For breast cancer dataset SVM was the best efficient algorithm while J48 Decision Tree was the worst. Similarly J48 Decision Tree was best efficient algorithm for diabetes dataset, while SVM was worst. One algorithm best for one dataset and worst for another. Figure 2, shows the results for time taken by algorithm to build the classifier.

Figure 2. Time Taken by different Machine learning Algorithms

Unlike the efficiency, the time take by all algorithms for both datasets give us a clear picture that J48 Decision tree is most time efficient algorithm while neural network is the worst.

V. CONCLUSIONS

Machine Learning Algorithms are extensively used in medical line so that disease detection can be made easy and at early stages, so that with proper treatment the suffering patient can be cured. The aim of this paper was to apply different machine learning algorithms on two different machine learning algorithms in order to determine which among the lot is best for disease detection. Our results showed that none of the algorithm can be termed as most efficient as all algorithms performed differently for 2 different datasets. However we can say that J48 Decision tree is the most time efficient algorithm among the lot.
I will like to thank Mr. Prab Simran Singh (Research Fellow), for helping me in carrying out this research. Further I will like to thank Department of Electronics Technology, Guru Nanak Dev University, Amritsar for providing me necessary infrastructure to carry out this research.

REFERENCES


