CS 445/545: Machine Learning

TUBERCULOSIS DETECTION FROM CHEST X-RAY USING MACHINE LEARNING ALGORITHMS

Abstract:

Tuberculosis is a disease caused by Mycro-bacterium, which can affect all the organs along with the Lungs. India has the world's highest burden of Tuberculosis with millions of cases estimated per year. Studies suggest that active tuberculosis increases the progression of Human Immunodeficiency Virus (HIV) infection. It is much more likely to be a fatal disease among HIV-infected persons than persons without HIV infection. Diagnosis of pulmonary tuberculosis has always been a problem. Classification of medical data is an important task in the prediction of any disease. In this paper, we propose a machine learning approach to compare the performance of both K Nearest Neighbors and the ensemble of classifiers on Tuberculosis data. The classification models were trained using real data. The prediction accuracy of the classifiers was evaluated using k -fold Cross-Validation and the results have been compared to obtain the best prediction accuracy. The results indicate that the ensemble Boosting Algorithm performs well among basic KNN, Local KNN, and Weighted KNN with an accuracy of 69.2%.

Introduction:

There is an explosive growth of biomedical data, ranging from those collected in pharmaceutical studies and cancer therapy investigations to those identified in genomics and proteomics research. The rapid progress in data mining research has led to the development of efficient and scalable methods to discover knowledge from these data. Medical data mining is an active research area under data mining since medical databases have accumulated large quantities of information about patients and their clinical conditions. Relationships and patterns hidden in this data can provide new medical knowledge that has been proved in many medical data mining applications.

The data classification process using knowledge obtained from known historical data has been one of the most intensively studied subjects in statistics, decision science, and computer science. Machine Learning techniques have been applied to medical services in several areas, including prediction of the effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data. In order to help the clinicians in

diagnosing the type of disease, computerized data mining, and decision support tools are used which are able to help clinicians to process a huge amount of data available from solving previous cases and suggest the probable diagnosis based on the values of several important attributes. There have been numerous comparisons of the different classification and prediction methods, and the matter remains a research topic. No single method has been found to be superior over all others for all data sets.

In this dataset before applying the algorithms, we have applied the preprocessing techniques like Trimming, inverting, and Compressing methods in order to get the data into the same size of 1024*1024 to get more accurate results and then apply the algorithms.

The information of our total patients is given below:

Total Patients	155
Patients with TB	50%
Patients without TB	50%
COLUMNS	2(Study ID and TB findings)
Total images	155
Variable Image Size	50-70 KB

Methods:

K-Nearest Neighbor:

The k-nearest Neighbors algorithm (k-NN) is a method for classifying objects based on the closest training feature space. KNN is a type of instance-based learning, or lazy learning where the function is only approximated locally, and all computation is deferred until classification. Here an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive, typically small).

Local mean based Kmeans:

A method called local mean-based k-nearest neighbor (LMKNN) is being proposed. This method is a simple, effective, and resilient nonparametric classification. This LMKNN has been proven to improve classification performance and also reduce the effectiveness of existing outliers, especially in small data sizes. The LMKNN process could be described as follows:

1. Determination of Value k
2. Compute of the distance between test data to each all training data using the Euclidean distance using the equation:

   $$(x,y)=||x-y|| =\sqrt{\sum N|x-y|2}$$

3. Sort distance of data from the smallest to the largest as much as k for each data class
4. Calculate local mean vector of each class with the equation

   $$mk=i\sum kyNN$$

5. Define the test data class by calculating the closest distance to the local mean vector of each data class using the equation

   $$w = argmind(x,mkcwjwj), = 1,2,...,M$$

Distance Weighted KNN:

A method called distance weight k-nearest neighbor (DWKNN) is being proposed. This method specifies a new data class based on the weight value obtained from the distance between data so that misclassification occurs due to ignoring the proximity between data can be overcome. This weighting method had a good performance because it can reduce the influence of outliers and the distribution of unbalanced data sets. The DWKNN process could be described as follows:

How to determine the k value?

Calculate the test data distance for each data in each class using the Euclidean distance

1. Sort the distance between data from the smallest to the largest according to the number of *k*.
2. Calculate the weights from the distances between the ordered data.

3. In solutions to count weight based on the distance between data, one of which may be use equation:-

$$wi = 1\ (4)\ (xq, xi)$$

Gradient Boosting:

Boosting is an algorithm for constructing a “strong” classifier as a linear combination of a “simple” “weak” classifier. Instead of re-sampling, Each training sample uses a weight to determine the probability of being selected for a training set. The final classification is based on the weighted vote of weak classifiers. It is sensitive to noisy data and outliers. Decision trees are usually used when doing gradient boosting.

Gradient boosting involves three elements:

1. A loss function to be optimized.
2. A weak learner to make predictions.
3. An additive model

1. Loss Function: The loss function used depends on the type of problem being solved. It must be differentiable, but many standard loss functions are supported and you can define your own.
2. Weak Learner: Decision trees are used as the weak learner in gradient boosting. Specifically, regression trees are used that output real values for splits and whose output can be added together, allowing subsequent models outputs to be added and “correct” the residuals in the predictions.
3. Additive Model: Trees are added one at a time, and existing trees in the model are not changed.A gradient descent procedure is used to minimize the loss when adding trees.Traditionally, gradient descent is used to minimize a set of parameters, such as the coefficients in a regression equation or weights in a neural network. After calculating error or loss, the weights are updated to minimize that error.

Results:

Results show that certain algorithms demonstrate that boosting ensemble performance is best when compared to the KNN algorithms. These measures will be the most important criteria for

the classifier to consider as the best algorithm for the given category. Below is the evaluation measures used for various classification algorithms to predict the best accuracy

Local KNN:

	precision	recall	f1-score	support
0	0.63	0.85	0.72	26
1	0.76	0.50	0.60	26
accuracy			0.67	52
macro avg	0.70	0.67	0.66	52
weighted avg	0.70	0.67	0.66	52

Weighted KNN:

	precision	recall	f1-score	support
0	0.66	0.96	0.78	26
1	0.93	0.50	0.65	26
accuracy			0.73	52
macro avg	0.79	0.73	0.72	52
weighted avg	0.79	0.73	0.72	52

Gradient Boosting:

Learning rate: 0.05
Accuracy score (validation): 0.577

Learning rate: 0.075
Accuracy score (validation): 0.635

Learning rate: 0.1
Accuracy score (validation): 0.615

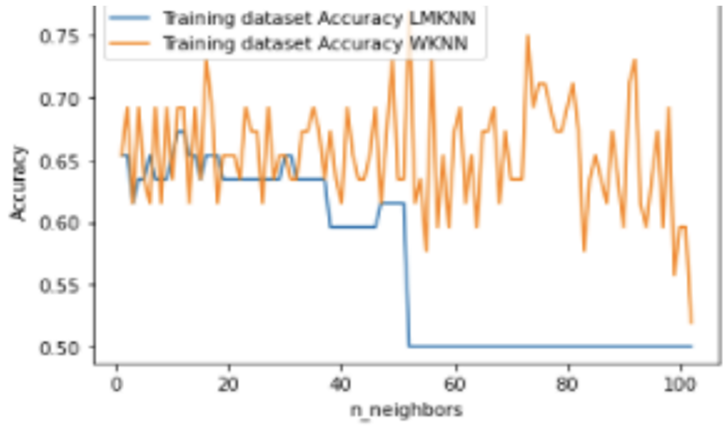
Learning rate: 0.25
Accuracy score (validation): 0.692

Learning rate: 0.5
Accuracy score (validation): 0.673

Learning rate: 0.75
Accuracy score (validation): 0.538

Learning rate: 1
Accuracy score (validation): 0.577

Predicting the accuracy of the Classifiers:



Conclusion:

Tuberculosis is an important health concern as it is also associated with other diseases. Retrospective studies of tuberculosis suggest that active tuberculosis accelerates the progression of HIV infection. Recently, intelligent methods such as Artificial Neural Networks(ANN) have been intensively used for classification tasks. In this article, we have proposed Machine learning approaches to classify tuberculosis using both basic and ensemble classifiers. Finally, two models for algorithm selection are proposed with great promise for performance improvement. Among the algorithms evaluated, Boosting is the best method as it has high accuracy.

References:

1. Sejong Yoon and Saejoon Kim, "Mutual information-based SVM-RFE for diagnostic Classification of digitized mammograms", Pattern Recognition Letters, Elsevier, volume 30, issue 16, pp 1489–1495, December 2009.
2. Rethabile Khutlang, Sriram Krishnan, Ronald Dendere, Andrew Whitelaw, Konstantinos Veropoulos, Genevieve Learmonth, and Tania S. Douglas, "Classification of Mycobacterium tuberculosis in Images of ZN-Stained Sputum Smears", IEEE Transactions On Information Technology In Biomedicine, VOL. 14, NO. 4, JULY 2010.
3. Erol S. Kavvas, Edward Catoiu, Nathan Mih, James T. Yurkovich , Yara Seif , Nicholas Dillon, David Heckmann, Amitesh Anand, Laurence Yang, Victor Nizet ,Jonathan M. Monk& Bernhard O. Palsson. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance, Nature Communications. 2018.
4. Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC, Dye C"The Growing Burden of Tuberculosis- Global trends and interactions with the HIV epidemics", Arch Intern Med. 2003 May 12;163(9):1009-21
5. Annie Luetkemeyer" Tuberculosis and HIV" HIV Insite Knowledge Base Chapter January 2013.61.