CS 445/545: Machine Learning

TUBERCULOSIS DETECTION FROM CHEST X-RAY

[Using machine learning algorithms]

Anvitha Pasumarthi Faiyazthulla Shaik Chaitanya Boyapati Sandeep
Singamaneni Lahari Katepalli

Abstract:

Tuberculosis is a bacterial infection caused by Mycrobacterium that can damage any organ, including the lungs. With millions of cases reported each year, India has the world's greatest burden of tuberculosis. Active TB appears to hasten the course of Human Immunodeficiency Virus (HIV) infection, according to research. It is far more likely to be a deadly illness in HIV-positive people than in healthy people. Pneumococcal TB diagnosis has always been difficult. In the prognosis of any disease, classification of medical data is critical. On Tuberculosis data, we offer a machine learning technique to compare the performance of K Nearest Neighbors and an ensemble of classifiers. Real data was used to train the categorization models. The classifiers' prediction accuracy was assessed using k-fold Cross-Validation, and the results were compared to determine the best prediction accuracy. The ensemble Boosting Algorithm outperforms basic KNN, Local KNN, and Weighted KNN with an accuracy of 69.2 percent, according to the results.

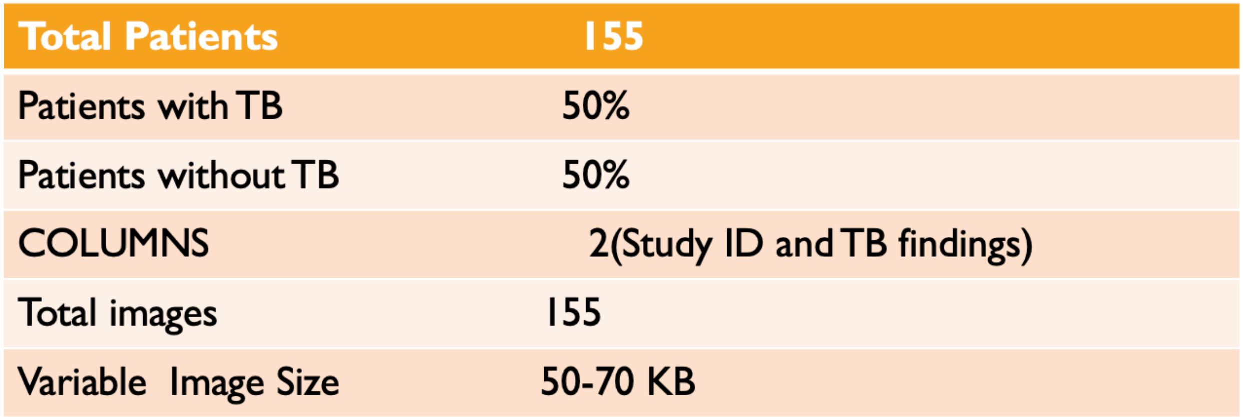
Introduction:

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.

One of the most intensely researched issues in statistics, decision science, and computer science has been the data categorization process employing information derived from known historical data. Machine Learning techniques have been used in medical services to predict the success of surgical treatments, medical testing, medicines, and the finding of correlations between clinical and diagnostic data, among other things. Computerized data mining and decision support systems are used to assist physicians in detecting the kind of disease. These tools may help clinicians analyse a large quantity of data gathered from previous cases and offer a likely diagnosis based on the values of many key features.

Numerous comparisons of alternative categorization and prediction systems have been conducted, and the subject remains a study issue. For all data sets, no single strategy has been determined to be better to the others. Before applying the algorithms to this dataset, we used preprocessing techniques such as trimming, inverting, and compressing to reduce the data to the same size of 1024*1024 in order to obtain more accurate results, and then we used 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

Individual Focus Areas:

Anvitha Pasumarthi and Faiyazthulla Shaik: Pre-Processing/Scaling the data:

➔ Downloaded the data

➔ Imported the necessary libraries

➔ Scaled the data using Trim

➔ Then compressed the data

➔ At last inverted the images into the desired format.

Chaitanya Boyapati: Applied local K-nearest neighbor(KNN) algorithm

➔ Using the scaled data , divided into test and train.

➔ On the training data, applied K-means

➔ Printed out the performance report : precision, recall, f1 score, support and accuracy.

Sandeep Singamaneni: Applied distance weight K-nearest neighbor(KNN) algorithm

➔ Using train dats, applied weighted KNN

➔ Printed out the performance report: precision, recall, f1 score, support and accuracy.

➔ Plot the graph showing the accuracy from both the classifiers (local KNN and distance weighted KNN).

Lahari Katepalli : Applied gradient Boosting algorithm

➔ With the training data and learning rate for the range (0.05,1)

➔ Applied gradient boosting algorithm and calculated the accuracy.

➔ Compared the accuracies from the three algorithms and chose that gradient boosting algorithm has higher accuracy and best fits our problem.

Methods:

K-Nearest Neighbor:

The k-nearest Neighbors algorithm (k-NN) is a classification approach that uses the closest training feature space to categorize objects. KNN is a kind of instance-based learning (also known as lazy learning), in which the function is only estimated locally and all computation is postponed until classification. An item is categorised here by a majority vote of its neighbors, with the object being allocated to the class that is most prevalent among its k closest neighbors (k is a positive, often small number).

Local mean based Kmeans:

Local mean-based k-nearest neighbor (LMKNN) is a technique that has been proposed. This approach is a nonparametric categorization that is simple, effective, and durable. This LMKNN has been shown to boost classification performance while also reducing the effectiveness of existing outliers in small data sets. The following is a description of the LMKNN process:

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:

$$(x,y)=\|x-y\| =\surd\Sigma N|x-y|2$$

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

$$mk=i\Sigma kyNN$$

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 distance weight k-nearest neighbor (DWKNN) approach has been presented. This function creates a new data class based on the weight value derived from the distance between data, allowing misclassification caused by disregarding data closeness to be avoided. Because it may lessen the impact of outliers and the distribution of imbalanced data sets, this weighting strategy performed well. The following is a description of the DWKNN process:

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 a method of creating a "strong" classifier from a "simple" "weak" classifier by combining them linearly. Rather of resampling, each training sample is given a weight that determines the likelihood of being chosen for a training set. The weighted vote of weak classifiers determines the final categorization. It's susceptible to outliers and noisy data. When doing gradient boosting, decision trees are generally employed.

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 type of loss function utilized is determined on the issue being addressed. It must be differentiable, however there are numerous common loss functions available, as well as the ability to define your own.

2. Weak Learner: In gradient boosting, decision trees are utilized as the weak learner.

Regression trees are utilized specifically because they provide actual values for splits and can be combined together, allowing future model outputs to be added and "correct" the residuals in the predictions.

3. Additive Model: Existing trees in the model are not modified, and new trees are inserted one at a time. When adding trees, a gradient descent approach is utilized to minimize the loss. Gradient descent has traditionally been used to minimize a collection of parameters, such as regression coefficients or weights in a neural network. The weights are changed once the mistake or loss has been calculated in order to reduce the error.

Results:

When compared to KNN algorithms, the results suggest that certain algorithms perform better at boosting ensemble performance. These factors will be the most crucial for the classifier to examine when determining the optimal method for a specific category. The assessment measures used to estimate the greatest accuracy for various classification algorithms are listed below.

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

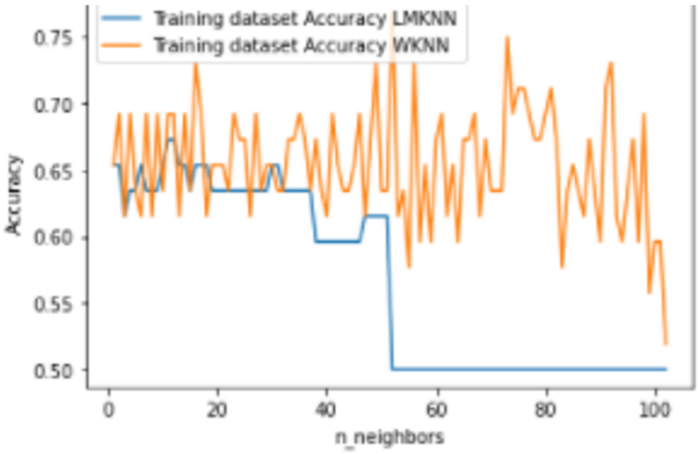
Learning rate: 0.75

Accuracy score (validation): 0.538

Learning rate: 1

Accuracy score (validation): 0.577

Predicting the accuracy of the Classifiers:



Conclusion:

Tuberculosis is a major public health problem since it is linked to a variety of other illnesses. Retrospective TB studies imply that active tuberculosis speeds up HIV infection progression. For categorization challenges, intelligent approaches such as Artificial Neural Networks (ANN) have recently been widely utilised. We suggested machine learning algorithms to classify TB in this study, employing both basic and ensemble classifiers. Finally, two algorithm selection models are given, both of which show considerable potential in terms of performance improvement. Boosting is the best approach among the algorithms analyzed because of its great 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, issue16, 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. Annie Luetkemeyer" Tuberculosis and HIV" HIV Insite Knowledge Base Chapter January
   2013.61.