Lung Cancer Prediction Using ML

Doaa Haider

Dr.Ammar Mohamed
Cairo Uni.

***Abstract*—lung cancer is one of the most popular deasis and the most killing cancer's types .About every two and a half minutes, someone in the U.S. is diagnosed with lung cancer, and 4 in 5 of those diagnosed will ultimately die from the disease. Yet, more Americans than ever are surviving lung cancer. While the disease remains the leading cause of cancer deaths among both women and men, over the past five years, the survival rate has increased by a dramatic 13machine-learning act a big role to detect the disease most of the related papers worked on one or two according to data type but here we will try and test all classification model and make a fine tuning of their hyper parameters and also use different kinds of metricise to judge perfectly on the algorithms our big challenge is to reach the best performance of the model bu make comparison between different algorithms and choose the highly performance also tune it's parameter to reach the maximum optimisation of this parameters**

1. Introduction

artificial intelligence and machine learning helps efficiently in the medical field as they seeking of extract the best result from data set which help to detect disease and reduce number of death so how to detect if the patient' lung has cancer or not it's very critical as you can save his life . based on The (American Cancer Society) report there are most common symptoms of lung cancer based on presences of those symptoms we will use the machine learning to detect if the patient may has lung cancer or not by using numbers of most common classifiers (KNN , SVM , LogisticRegression ,Random-Forest , Decision-Tree) and optimise their performance most of papers solved this problem(detect lung cancer) by using deep-learning which achieve a very good accuracy but at this paper we need to achieve high performance of using machine learning as an early detection and avoid using CT scan only if the o in this paper we try to achieve maximum performance of classification algorithms through best handle of their parameters to detect if the patient has lung cancer or not based on most common Symptoms of the disease and we can use the other techniques only on the positive cases to avoid the danger of CT scan

2. Related Work

here let's present many related from kaggle worked on this topic using the same data but applied different models the first one used two predictor:

1- Ensemble Technique using a Rondom Forest classifier with ...

$$(n_estimators = 300, job = -1, maxdepthinrange(3, 10, 2)$$

which result 91 percentage Accuracy 2- neural networks Technique with 265 batch size and 50 epoch which result 88 percentage the second one using XG-BOOST classifier with ...

$$(alph = 2, base_score = 0.5, booster = gbtree)$$

3. Data

in this paper we have a data set from kaggle of 309 patient and 15 features extracted divided into two classes YES/NO represented as 1 for NO /2 for YES only Age has continuous values the minimum and maximum values are between 21 and 87 this attribute has a mean of 62.67 and median of 62.0 with a standard deviation of 8.21 , this tell us that the Age distribution is skewed towards older people although there may be some very Young people have lung cancer the target is Yes has lung cancer or No has not lung cancer the percentage between positive and negative case is balanced All the remaining numerical attributes,'Smoking' through Chest Pain' have minimum and maximum values of 1 and 2 respectively . if the data was perfectly balanced or split between these values, each of these features would have mean 1.5 , attributes with mean values above 1.5 exhibit more data points with a value of 1 this hints at the possibility of an unbalanced initial data set that could be further refined with careful sampling after check the data set it found 33 values duplicated so we should remove them as duplicated values because

Duplicated values can be problematic as they can overweight certain possibilities due to that combination of attributes being over represented when developing any kind of trend for data it's important to ensure that data points or attribute vectors are seen a realistic number of times they can artificially enhance particular cases so they should be removed in some cases they are a natural part of the sampling method employed this where domain knowledge and a deeper understanding how data collection was carried out can assist in making a decision when numerous attributes are involved and they are all sampled independently it is quite rare to have a duplicate entry in those cases it is easier to identify and drop them.

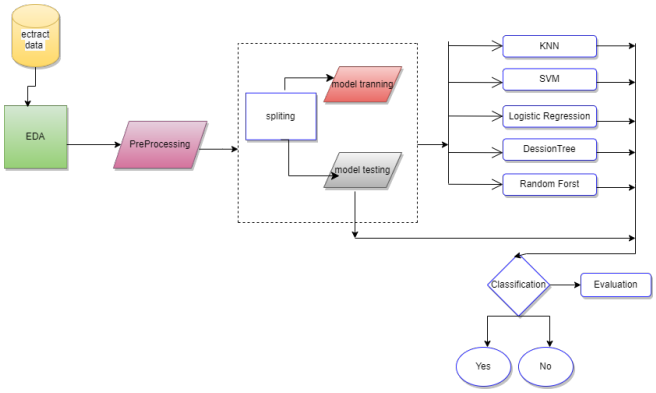
3.1. our methodology is

: 1- collect the data
2- EDA and pre-processing
3- choose the model
4- train and test the model
5- evaluate the model
6- tune the parameters
detecting lung cancer is a very challenge

4. Problem Formulation

the world looking for early detect of lung cancer and reduce number of death in this paper based on some of symptoms by using the suitable classifier and optimise it's result the model could predict if the patient has lung cancer or not

5. Model

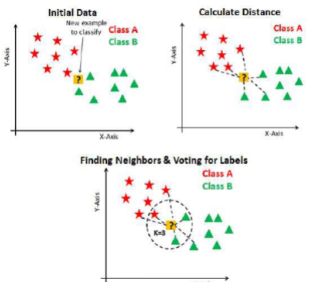


there are many types of classification algorithms let's introduce them briefly

5.0.1. KNN (KeyNearstNeghbours). one of the simpler classification technique there is no model to be fit it's an instance base model.

it's main idea is :

to get labelled data set and choose "K" number of neighbours find the nearest neighbours to the new data point using any distance measure then count the number of points in each category . assign the new data to major catergory .



it's main application :

- Customer Churn
- Speech Recognition
- Bio-metric- Data

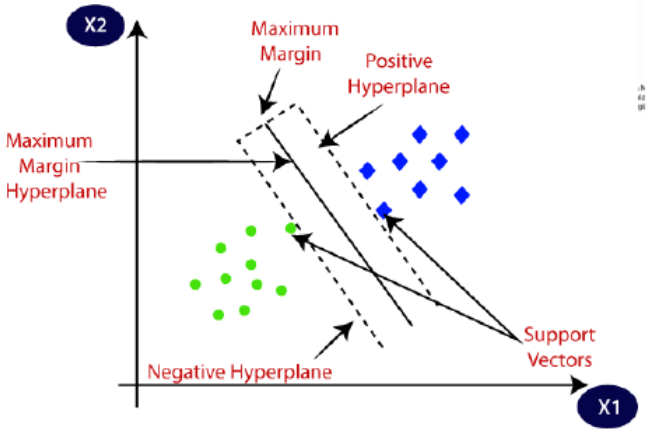
it's main Advantages and dis Advantages :

- it's very simple and easy .
- has a few parameters to handle.
- it's very slow and choosing "K' and distance measure will affect final result.
- not suitable for high dimension data.
- the result based on the quality of the data.
- high memory usage as we need to load all the data in every training time

5.1. SVM (support vector machine

one of the best classifier ever was the dominate before ANN(artificial neural networks) SVM works with linear and non linear classification also used for regression tasks and outlier detection it's main application is:

(Face Recognition and image classification ,text categorisation , Bio-informatics) the main idea is choosing the best hyperplane that maximise the margin support vector : the closest point that identify the best hyperplane Hyperplane : we can choose the hyperplane depends on what the percentage of miss-classification allowed - maximal hyper classifier it's hard margin - support vector classifier it's soft margin - support vector machine it's non linear



in this algorithm the hyper-parameter is called (kernal trick) a kernal trick means a kernal function transform the data into higher dimension space to make it possible to perform liner separation on the data there are many types of this kernal function : - polynomial kernal

- radial bases function
- segmoid function
- linear function

the choice of the kernal and the kernal specific parameter affect the final result in our data this algorithm result hands on machin learnning (Aurelien Geron ch5(p153-p174

No	Kernel function	Formula	Optimization parameter
1	Dot -product	$K(x_n, x_i) = (x_n, x_i)$	C
2	RBF	$K(x_n, x_i) = \exp(-\gamma \parallel x_n - x_i \parallel^2 + C$	C and γ
3	Sigmoid	$K(x_n, x_i) = \tanh(\gamma(x_n, x_i) + r)$	C, γ, and r
4	Poly-nomial	$K(x_n, x_i) = (\gamma(x_n, x_i) + r)^d$	C, γ, r, d

5.2. Decision Tree

: a decision tree uses the tree structure to represent several possible decision paths. it's like a flow chart that mapping out the outcomes - it's built by recursive partitioning algorithm - the data is repeatedly partitioned using predicted value that do the best job of separating the data into relatively homogeneous partitions decision trees are widely used in classification and regression tasks it's main application is (Risk assessment - fraud prevention) a tree is composed of nodes and those nodes are chosen looking for the optimum split of the features . for that purpose different criteria exist there are :

Gini impurity :

it's measure the frequency at which any element of the data sets will be miss labelled when it's randomly labelled

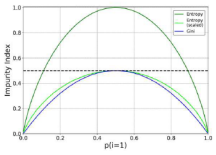
entropy :

is a measure of information that indicates the disorder of the features with the target the Gini criterion is much faster because it's less computationally expensive on the other hand the obtained results using the entropy criterion are slightly better.

$$Gini = 1 - \sum_{i=1}^{n} p^2(c_i)$$

$$Entropy = \sum_{i=1}^{n} -p(c_i)log_2(p(c_i))$$

where $p(c_i)$ is the probability/percentage of class c_i in a node.



to build a tree . the algorithm searches over all possible decisions and find the one that is most informative about the Target variable

your tree a better chance of finding features that are discriminative. consider performing dimensional reduction (PCA) beforehand to give your tree better chance of finding features that are discriminate

use max-depth=3 as an initial tree depth to get a feel for how the tree is fitting to your data , and then increase the depth use maximum tree to control the size of the tree to prevent the over-fitting .

5.3. Ensemble learning (Random forest)

Random forests are an example of an ensemble method, a method that relies on aggregating the results of an ensemble of simpler estimator.

Random forest used to avoid over-fitting caused by decision tree thy constructed bu building and combining multiple decision trees we can use it for classification through (majority vote) and for regression through (average prediction)

their main advantages and disadvantages:

- both training and prediction are very fast because of the simplicity of under-lying decision trees - they are very power-full often work well without heavy tuning - they are suitable for large data but not in high-dimension also the result are not easily interpret-able

it's main application is : - Credit Risk assessment - Chronic disease prediction - Mechanical parts breakdown prediction

5.4. Logistic-Regression

:

Logistic regression despite it's name is a linear model for classification rather than regression.

In this model, the probabilities describing the possible outcomes of a single trial are mode-led using a logistic function. it uses in Binary and Multi-nomial classification it's main application : - Email filtering - Tumor detection it's main Advantages and Disadvantages : - one of the simple algorithm provides great efficiency - can be updated easily using stochastic gradient descent - doesn't handle large categorical variables well - it requires transformation for nonlinear features

6. Results

in this paper we applied many classification algorithms to the data-set and got blow result:

6.1. Result for LogisticRegression :

```
Classification Report :

              precision    recall  f1-score   support

         0.0       0.92      1.00      0.96        44
         1.0       1.00      0.67      0.80        12

    accuracy                           0.93        56
   macro avg       0.96      0.83      0.88        56
weighted avg       0.93      0.93      0.92        56


The Accuracy of Logistic Regression is 92.86 %
```

6.2. Result for SVM :

here we use SupportVectors Classifier with hyper parameters (kernal = rbf , c=100 ,degree = 2)

```
Classification Report :

              precision    recall  f1-score   support

         0.0       1.00      0.75      0.86        12
         1.0       0.94      1.00      0.97        44

    accuracy                           0.95        56
   macro avg       0.97      0.88      0.91        56
weighted avg       0.95      0.95      0.94        56


The Accuracy of Support Vector Machine is 94.64 %
```

6.3. RodomForest :

```
Classification Report :

              precision    recall  f1-score   support

         0.0       0.86      1.00      0.93        44
         1.0       1.00      0.42      0.59        12

    accuracy                           0.88        56
   macro avg       0.93      0.71      0.76        56
weighted avg       0.89      0.88      0.85        56


The Accuracy of Random Forest Classifier is 87.5 %
```

6.4. Result for KNN :

```
Classification Report :

              precision    recall  f1-score   support

         0.0       0.83      1.00      0.91        44
         1.0       1.00      0.25      0.40        12

    accuracy                           0.84        56
   macro avg       0.92      0.62      0.65        56
weighted avg       0.87      0.84      0.80        56


The Accuracy of K Nearest Neighbors Classifier is 83.93 %
```

7. Conclusion :

all classifier provided good results but the best one is the SupportVectorClassifier as it's accuracy is 94 percentage