JES Journal of Engineering Sciences

Vol 11, Issue 3, March/
2020 ISSN NO:0377-9254

Comparison of Supervised and Unsupervised Data Mining Techniques on Predicting Lung Cancer

N Sri Hari[1], Sk Nyamtulla[2], V Tarunsai[2], Sk Sameer[2], k nagaraju[2]

[1]*Asst.Prof, Dept. of Computer science and engineering,*
Vasireddy Venkatadri Institute of Technology, Nambur, India

[2]*Student, Vasireddy Venkatadri Institute of Technology, Nambur, India*

Abstract - Lung cancer is the prominent type of cancer being the second most diagnosed cancer around the world. Machine learning and data mining techniques can do great for the field of medicine in predicting diseases such as lung cancer by using data recorded from patients. But there are supervised techniques which can be used when the output is known and unsupervised techniques to find patterns in the data. Here we are predicting lung cancer using both supervised and unsupervised techniques and comparing their performances. KMeans is a clustering algorithm which works as unsupervised and for supervised learning Random forest is chosen for best results. Aim of the paper is to propose a model for early detection of the disease which will help the doctor in saving the life of the patient.

Key Words: Lung cancer, data mining, KMeans, Random Forest

1. INTRODUCTION

Early prediction of lung cancer helps to prevent the lung from being affected more by the dangerous cells. Any cancer identified in the early stages is very significant to provide better treatment and gives a huge positive progress to get rid of cancer. Lung cancer is caused due to smoking and pollution, so early detection can help the patients to stop from smoking or from other factors that cause the cancer.

1.1 Types of lung cancer There are broadly 2 types of Lung Cancers. Which are common in most of the lung cancer cases.

1.1.1 Small Cell Lung Cancer (SCLC): This type of lung cancer can occupy up to 10% to 15% of all lung cancers. It is very rare for someone who has never smoked. SCLC often starts in the bronchi near the center of the chest, and it tends to spread widely through the body, the highest value is selected for the node.

1.1.2 Non-Small Cell Lung Cancer (NSCLC): Around 80 to 85 out of 100 lung cancers (around 80 - 85%) in the UK are non-small cell lung cancer (NSCLC). The three main types are adenocarcinoma, squamous cell carcinoma and large cell carcinoma.They are grouped together because they behave in a similar way and respond to treatment in a similar way.

1.2 Symptoms of lung cancer There are symptoms like cough, shortness of breath, chest pain, feeling tired etc..[1] which are considered in the research.

- The lung cancer affected people have a new cough, it's not like a regular one but started recently and it's continues, this cough can produce blood.
- Pain in the chest, back and shoulders particularly while coughing.
- Shortness of breath at any moment

- Sudden weight loss, and loss of appetite
- Feeling tired in most of the times
- Pain in swallowing of food and pain in the neck
- Damages to the fingers called clubbing of fingers
- Pain in chest while taking breath.

2 RELATED WORKS

V.Krishnaiah , Dr.G.Narasimha , Dr.N.Subhash Chandra[2] had used WEKA tool to process the data by using algorithms like If Then rule, decision tree, Naive Bayes, Bayesian Classification, One Dependency Augmented Naïve Bayes classifier (ODANB) and naive credal classifier 2 (NCC2) and Artificial Neural networks. The Naive Bayes turned out to be the best classifier with the highest accuracy. And neural networks are clumsy and inefficient to predict lung cancer.

Dr.T.ChristopherP1P, J.Jamera banuP2 [3] Analyzed the lung cancer prediction using classification algorithms such as Naive Bayes, Bayesian network and J48 algorithm using WEKA tool. They concluded that Naive Bayes algorithm gives better solutions compared to other classification algorithms. The suggestion is to use clustering algorithms as well.

The article by G Vijaya, Dr.A.Suhasini [4] provides an overview of the available literature on the detection of lung cancer in the data mining framework. The conclusion is that the most effective model to predict patients with Lung Cancer appears to be Naïve Bayes, followed by Association Rule Mining, and Decision Trees.

The classification techniques of supervised learning are analyzed by many of the authors but not the clustering of unsupervised learning which can be used to draw patterns in the available data if class labels are not present, in the case of need to group lung cancer patients into clusters such as Low, medium, High according to their symptoms

3 REVIEW OF LITERATURE

In Data Mining there are two major categories of learning they are supervised and unsupervised learning.

In Supervised learning[5] algorithm builds a mathematical model from a set of data that contains both the input and desired outputs. These supervised models are trained using labeled data i.e. input and desired outputs are known. In this learning, the algorithm receives a set of inputs along with corresponding correct outputs. Algorithm learns by comparing its predicted output with actual outputs to find out errors. Then, the model is modified accordingly.

In Unsupervised learning [5] a mathematical model is to be built from a set of data which contains only inputs. Desired output labels are not present in this type of learning. Unsupervised learning is used against that data which doesn't consist of historical labels

A clustering algorithm KMeans is used in this research from unsupervised and Random Forest from supervised learning models.

3.1 Random Forest

Constructing every tree using a one of a kind bootstrap sample of the information, random forests trade how the classification or regression trees are constructed. In the case of standard trees, every node is splitting up by using the best amongst a subset of predictors randomly chosen at any node [6]. This method turns out to perform very well as compared to many different classifiers, such as discriminant analysis, support vector machines and neural networks.

Algorithm:

1. If there are N variables or N features in the input data set then we have to select a subset of m (m<n) features randomly out of N features, and observations or data instances should be picked randomly.

2. Using the best split method on the m features to calculate the number of nodes 'd'.

3. Keep on splitting the nodes to reach child nodes until the tree is grown to maximum possible extent.

4. Select a different subset from the training data with replacement and try it on another decision tree following steps 1 to 3 repeat this process to build and train 'n' decision trees.

5. Final class assignment is performed on the basis of the majority votes from n trees.

3.2 K Means

K-means clustering is a type of unsupervised learning algorithm developed by J. McQueen in 1967, J.A Hartigan and M.A Wong. [7, 8]

Algorithm:

1. We have the data objects ('n') which are classified into a 'k' number of clusters in which each observation belongs to the cluster having the nearest mean.

2. It defines 'k' sets initially, one for each cluster $k \leq n$. The clusters are formed away from each other.

3. Then, it will organize the data in suitable data sets and associate them to the nearest set. In case If there is no data pending, in this case it will perform early grouping. It is necessary to re-calculate 'k' new set as barycenter of the clusters from the previous step.

4. After having these 'k' new sets, the same data set points and nearest new sets are bound together.

5. Finally, a loop is generated. As a result of this loop, the 'k' sets change their location step by step until no more changes are made

4 DATA SET DESCRIPTION

The data set contains the record of 1000 patients and the major symptoms such as Alcohol use, smoking, passive smoker, Shortness of breath, Dry cough, Frequent cold are taken as the features to study. The class label is termed as Level which is categorized into three types low, medium, high. The python modules such as sklearn, numpy, pandas are used to perform data preprocessing and for implementing clustering and classification algorithms.

Table -1: Features in the data set

Features
Smoking
Passive Smoker
Shortness of breath
Alcohol use
Dry cough
Frequent cold

5 PERFORMANCE ANALYSES

In this research both classification and clustering algorithms such as Random Forest classification and KMeans clustering are implemented in python by using predefined libraries available such as cluster

and ensemble model RandomForestClassifier from Sklearn and to plot the graphs matplotlib library is used. The results of both algorithms are compared to find the best method to predict lung cancer. The performance metrics such as Accuracy, Precision, Recall and F1 score are calculated by using a Sklearn library called Metrics.

Random forest is also used for classification, which is best suitable for problems like prediction. As it takes the majority of voting from a number of trees, its results are quite impressive when compared with clustering technique. The model is trained and tested with train and test data. The confusion matrix and accuracies of both algorithms are shown in the Table-6

Table -2: Confusion matrix of Random Forest

Level	Low	Medium	High
Low	140	15	0
Medium	5	164	0
High	0	0	176

Table-2 shows the confusion matrix of Random Forest model tested with test data of 500 records. And the Precision, recall values of the model are shown in the Table 3

Table -3: Precision, recall and F1 score of Random Forest

	Precision	Recall	F1 Score
High	1.000	1.000	1.000
Low	0.932	0.936	0.929
Medium	0.934	0.940	0.937

Here KMeans algorithm is used here to create three clusters. patterns in the same cluster are alike and patterns belonging to two different clusters are different.[9] As there are three class labels in the data low, medium, high, three clusters are created and we can see that each cluster holds a particular class labeled data in majority.There is no need to divide dataset into train and test data as KMeans is a unsupervised learning and we can give whole data to form clusters and calculate accuracy by taking the positives from each cluster, the confusion matrix is shown in Table-4.

Table -4: Confusion Matrix of KMeans

Level	Cluster 0	Cluster 1	Cluster 2
Low	295	0	70
Medium	10	293	0
High	30	141	161

Table -5: Precision, recall and F1 score of K Means

	Precision	Recall	F1 Score
Cluster 0	0.881	0.808	0.843
Cluster 1	0.675	0.967	0.795
Cluster 2	0.697	0.485	0.572

Fig.1 shows the clusters formed, the Principal Component Analysis [10] is used to reduce the 6 dimensional data into 2 dimensions to get better understanding of scattering of data.

Fig.1: Clusters formed by Kmeans algorithm.

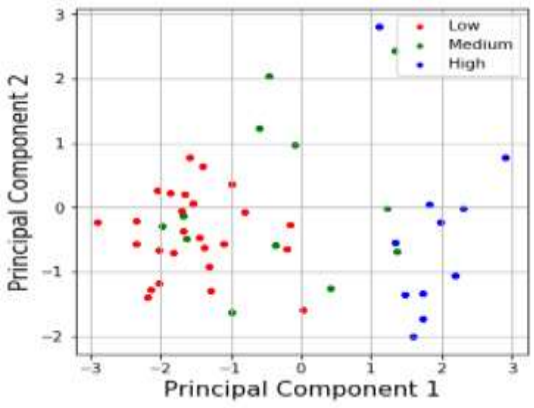

Table -6: Accuracy Comparison table

Algorithm	**Accuracy**
KMeans	74.90%
Random Forest	95.80%

6 CONCLUSIONS

Here, we did the comparison of clustering and classification techniques to predict lung cancer by taking KMeans and Random Forest into account and found that Random Forest is best suitable for lung cancer prediction with high accuracy compared to KMeans clustering. But, clustering techniques also have the capability to produce good results and can also be used in case of unlabeled data.

REFERENCES

[1] American Cancer Society. Cancer Facts & Figures 2012.

[2] V. Krishnaiah, Dr.G.Narasimha, Dr.N.Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques"/ (IJCSIT) Vol. 4 (1) , 2013, 39 - 45

[3] Dr.T.ChristopherP1P, J.Jamera banuP2, "Study of Classification Algorithm for Lung Cancer Prediction", IJISET - Vol. 3 Issue 2, February 2016

[4] G Vijaya, Dr.A.Suhasini "Early Detection of Lung Cancer using Data Mining Techniques: A Survey", Proceedings of International Conference "ICSEM' 13.

[5] Ayushi Chahal, Preeti Gulia, "Machine Learning and Deep Learning", (IJITEE) ISSN: 2278-3075, Volume-8 Issue-12, October 2019.

[6] Vrushali Y Kulkarni, Pradeep K Sinha, "Effective Learning and Classification using Random Forest Algorithm", (IJEIT) Volume 3, Issue 11, May 2014.

[7] Dr. S.P.Singh, Ms Asmit Yadav, "Study of K-Means and Enhanced K-Means Clustering Algorithm", ijarcs Volume 4, No. 10, Sep-Oct 2013.

[8] Oyelade, Oladipupo, Obagbuwa, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", ijcsis, Vol. 7, o. 1, 2010.

[9]. Sudhir Singh and Nasib Singh Gill, "Analysis And Study Of K-Means Clustering Algorithm", (IJERT) Vol. 2 Issue 7, July - 2013.

[10] B.Firdaus Begam, J.Rajeswari, "Visualization of Chemical Space using Kernel Based Principal Component Research", (IJITEE) ISSN: 2278-3075, Volume-8, Issue-11S, September 2019.