Journal of Engineering Sciences

Vol 11, Issue 5,May/2020
ISSN NO:0377-9254

A COMPARATIVE STUDY OF SUPERVISED AND UNSUPERVISED MACHINE LEARNING TECHNIQUES ON LUNG CANCER PREDICTION

M. Sheik Mansoor[1] and M. Mohamed Sathik[2]

[1]*Research Scholar r (Reg. No. 17221192161007), Sadakathullah Appa College, Affiliated to Manonmanium Sundaranar University, Tirunelveli, Tamilnadu, India*

[2]*Principal and Research Supervisor, Sadakathullah Appa College, Tirunelveli, Affiliated to Manonmanium Sundaranar University, Tirunelveli, Tamilnadu, India*

***Abstract*— Lung cancer is one of the most dangerous type of cancers which has the high spread rate. Lung cancer metastases spreads through fluid lymph nodes and bloodstreams to other organs like bone, glands and brains. Due to the air and industrial pollution the rate of people who has affected by the lung cancer is increasing enormously. According to the prediction reports of World Health Organization (WHO) the number of lung cancer deaths will reach 9.6 million in 2020, which is an alarming issue. Diagnosis the lung cancer at its earlier stage could help the physicians to treat the patients. Though the manual analysis of CT scan exists in the medical field, it is too hard for the medical advisors to predict the exact stage of the disease using the CT scan images. Hence, the medical informatics research community has created several machine learning model to predict the lung cancer and its type in the earlier stage. In this comparative research study, we have downloaded the lung cancer dataset from the Cancer Image Archive and given as the input to the two most accepted machine learning models such as, Artificial Neural Networks (ANN), Support Vector Machine (SVM) from supervised learning method and another unsupervised dataset as input for Apriori and K-means model from unsupervised learning to observe the changes. The final results and the performance metrics of the machine learning algorithms such as accuracy, precision and recall are compared with each other and tabulated.**

***Keywords*— Machine Learning; Lung Cancer Prediction; Supervised Learning; Cancer Diagnosis.**

1. INTRODUCTION

Lung cancer is a type cancer which starts in the cells of the lungs and spreads to the other parts of the human body [3]. Likewise, cancer cells such as breast, mouth and kidney can also spread to the lungs via lymph nodes or bloodstreams[2, 3]. The lung are is made up of sponge like structure in the chest of the human. The main objective of the lungs is to take oxygen into the body and release the carbon dioxide [1]. While breathing air passes through pipe like structure called trachea and propagates through bronchi nodes to enter lungs and come outs in the same path. The small sized holes in the bronchi nodes called alveoli passes the oxygen to the blood and takes out the carbon dioxide out from the blood [4,5].

At initial stage of lung cancer, DNA of the patient will change or damage and mutate the genes. Mutated genes will not work properly because they will not get any instruction from DNA properly or in a correct manner. This will cause the cells in the lung to divide and grow out of control in and around the lungs and causes the lung cancer [6].

As stated by Global Cancer Observatory (GCO), every 5.4 person has lung cancer among one million peoples in India. The alarming issue in the raise of lung cancer is, it has very low survival rate compare to any other cancer diseases. In India, 25% of cancer victims loses their life every year. Due to late stage diagnosis and fast outspread, deaths rate of lung cancer is too high compared to other prostate, colorectal, skin, kidney and breast cancers [7]. Accurately identify the lung cancer cells in its initial stage through manual analysis of CT scan is not possible. It makes difficult for medical advisors to predict the exact stage of the cancer using the CT scan images.

To overcome these issues and to identify the cancer type in early stage, Machine Learning techniques are used on the patient data. It helps the physicians

to acquire a clear cut knowledge about condition of the patients. Moreover, it helps physicians in identifying the type and vigorous of the cancer cells [8, 9].

Machine learning techniques can be classified into two major types based on its application and working nature. While considering the lung cancer prediction several research contributions and prediction methods were been introduced. In this research work, we have taken two supervised learning methods such as Artificial Neural Networks (ANN), Support Vector Machine (SVM) and two unsupervised learning methods Apriori and K-means for this comparative study. The datasets were downloaded from the open- source Cancer Imaging archives and given as the training set to these machine learning algorithm. The preprocessing, feature extraction and selection are kept same for all these four methods.

This comparative study paper is organized in such a manner such that, Section 2, describes the difference between the supervised learning and unsupervised learning. Section 3, explains the preprocessing, feature extraction and selection. Section 4, evaluates the performance of the ANN, SVM, Apriori and K-means and Section 5 concludes and discusses about the future work of the comparative study.

2. SUPERVISED LEARNING AND UNSUPERVISED LEARNING

In supervised learning, the machine learning system is trained with a well labeled information, which means that some data is already tagged with the correct answer. So it can be directly compared with learning process. A supervised learning algorithm learns from labeled training data, helps you to predict outcomes for unforeseen data. Where as in unsupervised machine learning technique, the dataset will not have a clear label. Instead it should be programmed in a manner, in which it should discover the information on its own. Unsupervised machine learning techniques can perform more complex processing tasks compared to the supervised learning algorithm but the results of the unsupervised machine learning are unpredictable compared to other deep learning and natural learning process.

A. Supervised learning algorithm: Artificial Neural Networks (ANN)

ANN maintains an interconnected nodes, called as neurons to gather information by identifying relationships and new pattern between the data. It has three layer such as, input neuron layer, hidden neuron layer and output neuron layer. Neurons in each layers will receive the input data, performs operations and forwards the data to the nearby connected neurons. Each neurons and the edge which connects the neurons has a particular weight. The weight will change on the neurons based on the learnings. ANN allows both forward and backward propagation for learning.

The final result of ANN are produced based on the maximum probability of neurons present in output layer. Even there exist several algorithms to predict the early state lung cancer, using ANN will produce an accurate result.

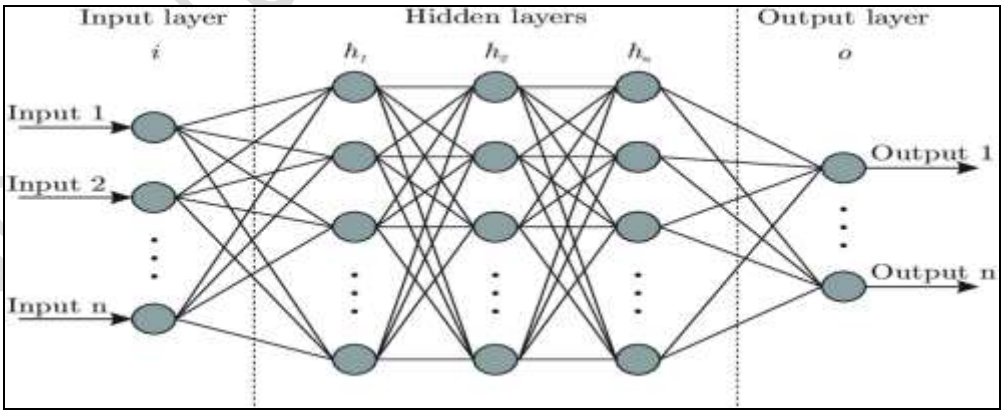

Fig. 1. Stucture of Neural Networks

B. Supervised learning algorithm : Support Vector Machine (SVM)

SVM is a supervised learning technique which performs classification and regression to identify the associations between the data in the given dataset. SVM is a discriminative classifier, which draws a hyper plane to differentiate the classes that are derived as outputs. The hyper planes are the decision boundaries. The maximum accuracy can

be attained only if the SVM draws the hyper plane separating all the objects to its classes correctly.

In here, support vectors are data that are very closer to the hyper plane and influence the position and orientation of the hyper plane. Using these support vectors, the margin of the classifier can be maximized to get the clear idea. Deleting the support vectors will change the position of the hyper plane. These are the points that help us build accurate SVM model.

In here, support vectors are data that are very closer to the hyper plane and influence the position and orientation of the hyper plane. Using these support vectors, the margin of the classifier can be maximized to get the clear idea. Deleting the support vectors will change the position of the hyper plane. These are the points that help us build accurate SVM model.

C. Unsupervised learning algorithm – Apriori Algorithm

The Apriori algorithm is a classical frequent item sets generation algorithm and a milestone in the development of data mining. It is used for finding frequent item in a dataset for Boolean association rule. Apriori algorithm uses prior knowledge of frequent item properties. An iterative approach or level-wise search where k-frequent item are used to find k+1 item.

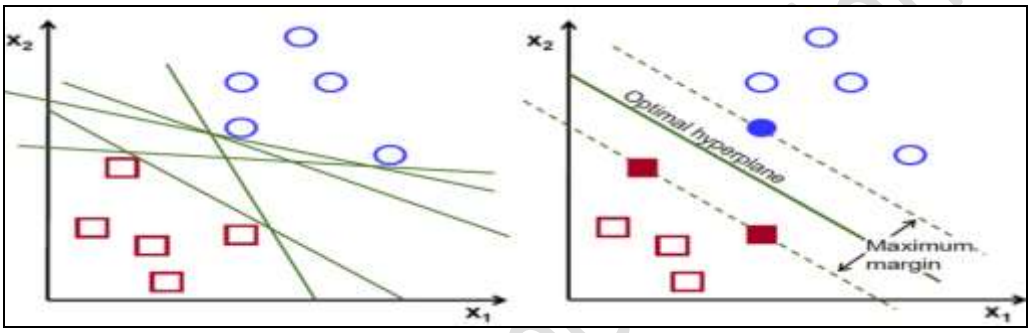

Fig. 2. Hyperplane of SVM

To improve the efficiency of level-wise generation of frequent item, an important property is used called Apriori property which helps by reducing the search space. Apriori property states that, all non-empty subset of frequent item set must be frequent.

D. Unsupervised learning algorithm – K Means

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

3. HANDLING DATA

A. Pre-processing

The pre-processing is an important task that is used for transforming the raw data into a useful and efficient data. The pre- processing include several steps such as data cleaning, transformation and reduction. Data cleaning is a process in which the missing values are replaced or removed and involves in removing the noise in the data through several methods such as regression, clustering or binning method. The transformation of data includes normalization of data, selection of attributes, transferring of continuous dataset to discrete through discretization and generation of hierarchy. The data reduction includes several actions such as aggregation of data as per the need, subset selection section in the particular attribute, replacement of original data into a data representation through parametric or non-parametric numerosity reduction and reduction of dimensionality.

B. Feature Selection

Feature selection is a dimension reduction method which is used to select the relevant feature for constructing the model. It includes four important approaches such as wrapper, filter, embedded and hybrid approaches for selecting the features. Wrapper approach is an approach which is highly complex computation. It selects the feature through classification and uses a learning algorithm for calculating the accuracy of the classification. The filter approaches select the subset of the feature without using any leaners. The database with higher dimension can use this type of feature selection approach. The embedded approach selects the feature during the training of the data and it uses applied learning algorithms for deriving the specificity of the approach. The hybrid approach is another approach where the filter and wrapper approaches are used in combination for selecting the feature. The feature is selected through the filter approach and are tested with wrapper approach. Thus it uses both the advantages for feature selection.

C. Feature Extraction

Feature extraction is another dimensionality reduction method through which the raw data will be transformed into a group of manageable data for further processing. It plays an important role in image processing as multiple parameters are needed to process the images. It includes low level extraction, edge level extraction, curvature extraction, shape detection, motion detection and so on. Here the low level processing of images includes several detection such as detection of edges, detection of corners, blob detection for detecting the regions in the images, ridge detection for extracting the thin line which is brighter than the nearby regions and feature transform through difference in scales of images. The curvature extraction intends to extract the direction of edges. It also identifies the change in intensity of images and the autocorrelation. The shape detection involves in finding the threshold of the images, region extraction and template matching. It also includes hough transformation which involves in extract the imperfect features of the objects by comparing it within the class through voting procedure. The motion detection model involves in extracting the motion of images and the optical flow by admiring the area of the images.

4. Performance evaluation of ANN, SVM, Apriori and K- means

A. Performance comparison of ANN and SVM

The ANN and SVM machine learning experiments are carried out on the Tensor Flow software, which is a free open-source software developed by Google Inc., The dataset used for the implementation is taken from Cancer Imaging Archives. The chosen dataset consist of CT scan data of 1019 patients with different cancers. Initially, information about the patients, who has affected by the NSCLC cancer is taken out from the given dataset. Around 419 patient records are extracted. Later these, NSCLC cancer data is separated into training dataset and test dataset with the ratio of 70:30. The training dataset are fed as an input to ANN and SVM, simultaneously. They are trained and computed simultaneously for best prediction results.

		True/Actual		
		Type ‘T’	Type ‘M’	Type ‘N’
Predicted	Cancer Type ‘T’	96	8	4
	Cancer Type ‘M’	5	89	5
	Cancer Type ‘N’	4	5	104

Table. 1. Prediction of cancer type using ANN method

		True/Actual		
		Type ‘T’	Type ‘M’	Type ‘N’
Predicted	Cancer Type ‘T’	101	6	2
	Cancer Type ‘M’	7	95	8
	Cancer Type ‘N’	8	5	88

Table. 2. Prediction of cancer type using linear SVM

Table 1 and Table 2, represent the prediction made through ANN and SVM respectively. Accuracy of ANN model is 90.2%. In 320 total predicted value, ANN has correctly predicted 290 values. However, accuracy of SVM algorithm is 88%, where 284 predictions are made correctly.

The second important performance metric of ML algorithm is precision. It is the fraction of relevant information retrieved (i.e.) in lung cancer type prediction, what fraction of patients belong to a particular cancer type.

In predicting lung cancer type, precision value of type 'x' cancer is found by the fraction of correctly predicted type 'x' cancer from the total prediction. Precision is calculated by the formula specified below,

Precision (Type 'x') = (No. of correctly predicted Type 'x') / (Total predicted Type 'x' cancer)

Precision values (in percentage)		
	ANN	**SVM**
Cancer Type 'T'	88.8%	92.6%
Cancer Type 'M'	90.8%	86.3%
Cancer Type 'N'	92.6%	87.1%

Table. 3. Precision values of ANN and SVM algorithm for given dataset.

The recall can be derived through the below mentioned formula,

Recall = (No. of correctly predicted type 'x' cancer) / (No. of actual type x cancer patients)

Table. 4. Precision values of ANN and SVM algorithm for given dataset.

On comparing both the algorithms we can observe that the ANN is more effective than SVM in many of the cases. On observing the precision value ANN works better than SVM but in few cases such as in detecting cancer type T the precision value is good for SVM. On comparing the recall value we can observe that ANN results good than SVM but in few cases such as predicting cancer type M, the SVM computes better than ANN as the correctness of result is good.

B. Comparison of Apriori and K-Means

This research work used a dataset which is needed to extract to achieve useful information about the effect of k-means algorithm to apriori algorithm from computation time and rule achieved. The dataset used consists of 8243 disease diagnose data. Medical data variables consist of disease diagnosis, age group, gender, the status of care. The partial data used can be seen in Table 5.

In the first approach, directly apply the apriori algorithm in the dataset to 4 input variables, namely disease diagnosis, age group, gender, the status of care in order to obtain confidence values, rules and computational time on apriori algorithms. The test results obtained from the Apriori algorithm can be seen in Table 6.

This rule information obtained in the Large Itemset 4 results in two rules, namely the diagnosis of another allergic rhinitis with the age group of female children and outpatient status. Then, the diagnosis of postoperative disease with the adult age group gender male and outpatient status with each confidence value of 69%. From these results, it can be seen that the information obtained from the Apriori algorithm is still lacking.

Recall values (in percentage)		
	ANN	**SVM**
Cancer Type 'T'	91.4%	87%
Cancer Type 'M'	86.4%	89.2%
Cancer Type 'N'	91.2%	89.7%

As shown in Table 7 above, the combination of the K-Means algorithm and the Apriori algorithm produces more complete and detailed information compared to the results obtained by the application of a priori algorithm only.

S. No.	Disease Diagnosis	Age Cluster	Gender	Status of Care
1	Observation of Febris	Baby	Male	Outpatient
2	Observation of Febris	Baby	Female	Outpatient
3	Observation of Febris	Baby	Male	Outpatient
4	Observation of Febris	Baby	Female	Outpatient
5	Paronychia	Adult	Female	Outpatient
6	Hnp Lumbalis	Adult	Male	Outpatient
:	:	:	:	:
8243	Disputes with the counselor	Toddlers	Female	Outpatient

Table 5. Sample patient diagnosis data in 2016

Using Apriori							
Full Data							
Disease Diagnose	Cataracts not Specified	Cataracts not Specified	Cataracts not Specified	Cataracts not Specified	Another Allergic Rhinitis	Another Allergic Rhinitis	Post Operation
Age Cluster	--	Elder	Elder	Elder	--	Child	Adult
Gender	Male	--	Female	--	Female	Female	Male
Status of Care	Out	Out	Out	Out	Out	Out	Out
Confidence (%)	69	76	60	66	69	69	69

Table 6. Data processing using Apriori algorithm

K-Means + Apriori				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Disease Diagnose	Cataracts not Specified	Cataracts not Specified	Another allergic rhinitis	Post Operation
Age Cluster	Elder	Elder	Child	Adult
Gender	Male	Female	Female	Male
Status of Care	Outpatient	Outpatient	Outpatient	Outpatient
Confidence (%)	66	66	92	93

Table 7. Data processing using K- Means and Aprior

Meanwhile, the computation time of K-Means and Apriori algorithms combinations are faster than the Apriori algorithm, where the total time from K-Means algorithm and Apriori algorithms combinations are 17.41 minutes while the total time of the Apriori algorithm is 21.93 minutes.

5. CONCLUSION

In this comparative research study, we have downloaded the lung cancer dataset from the Cancer Imaging archives and given as the input to the two most accepted machine learning models such as, Artificial Neural Networks (ANN), Support Vector Machine (SVM) and another patient dataset is given as input for unsupervised learning method such as Apriori and K-means model for observing the performance difference. The final results and the performance metrics of the machine learning algorithms such as accuracy, precision and recall are compared with each other and tabulated. Thus the comparison of unsupervised and supervised algorithms are compared.

Reference

1. K. Kancherla and S. Mukkamala," Feature Selection for Lung Cancer Detection Using SVM Based Recursive Feature Elimination Method", Lecture Notes in Computer Science, Vol. 7256, pp. 168-176,2012.
2. S.K. Lakshmanaprabu, S.N. Mohanty, K. Shankar, N Arunkumar, and G. Ramirez, "Future Generation Computer System, Vol. 92, pp.374-382, 2018
3. A. Trivedi and P. Shukla, "Lung Cancer Diagnosis by Hybrid Support Vector Machine", Communications in Computer and Information Science, Vol. 628, pp. 177-187, 2016.
4. T. Nadira and Z. Rustama, "Classification of Cancer Data Using Support Vector Machines with Features Selection Method Based on Global Artificial Bee Colony", Vol. 2023(1), pp. 1-7, 2018.
5. Arora, P., Boyne, D., Slater, J. J., Gupta, A., Brenner, D. R., and Druzdzel, "Bayesian Networks for Risk Prediction Using Real-World Data: A Tool for Precision Medicine, Value in Health",Vol. 22, pp.437-445, 2019.

6. M. B. Sesen, T. Kadir, R. B. Alcantara, J. Fox, and M. Brady," Survival Prediction and Treatment Recommendation with Bayesian Techniques in Lung Cancer", AMIA Annual Symposium, pp. 838-847.
7. K. Jayasurya, G. Fung, S. Yu, C. Dehing-Oberije, D. De Ruysscher, A. Hope, W. De Neve, Y. Lievens, P. Lambin , and A. L. A. J. Dekker, "Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy" Medical Physics, Vol. 37, pp. 1401-1407, 2010.
8. Dharshinni N P, Mawengkang H and Nasution M K M 2018. Mapping of medicine data with k-means and apriori combinations based on patient diagnosis. International Conference on Computing and Applied Informatics.Vol 2 (978).
9. E. Adetiba and O. Olugbara, "Lung Cancer Prediction Using Neural Network Ensemble with Histogram of Oriented Gradient Genomic Features" The Scientific World Journal", Vol. 2015, pp. 1-17, 2015.