International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 3 (September '2017)
Proceeding of - International Conference on Recent Trends in Computational Life Science and Information Technology,
Conference held at Tiruppur Kumaran College for Women,Tirupur,Tamilnadu, India.

ADAPTIVE LINEAR NEURON NETWORK IN LUNG CANCER PREDICTION SYSTEM

S.Karthigai,
Research Scholar,
Department of Computer Science,
Erode Arts and Science College,
Erode,Tamilnadu,India.

Dr.K.Meenakshi Sundaram,
Associate Professor,
Department of Computer Science,
Erode Arts and Science College
Erode,Tamilnadu,India.

Abstract: Both men and women Cancer is the most important cause of death. The early detection of cancer can be helpful in curing the disease completely. So the requirement of techniques to detect the occurrence of cancer nodule in early stage is increasing. A disease that is commonly misdiagnosed is lung cancer. Earlier diagnosis of Lung Cancer saves enormous lives, failing which may lead to other severe problems causing sudden fatal end. Its cure rate and prediction depends mainly on the early detection and diagnosis of the disease. One of the most common forms to analyze the cancer in early stage is the Knowledge discovery in data mining have applications in ADALINE is Adaptive Linear Neuron Network domain. The ability of a net to learn a new pattern equally well at any state of learning is called plasticity. Discovery of hidden patterns and relationships often goes unexploited. Using generic lung cancer symptoms such as age, sex, Wheezing, Shortness of breath, Pain in shoulder, chest, arm, it can predict the likelihood of patients getting a lung cancer disease. Aim of the paper is to propose a model for early detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient.

Keywords: Knowledge Discovery, ADALINE Network, plasticity

I. INTRODUCTION

Lung cancer is the one of the leading cause of cancer deaths in both women and men. Manifestation of Lung cancer in the body of the patient reveals through early symptoms in most of the cases. [1]. Treatment and prognosis depend on the histological type of cancer, the stage in the degree of spread, and the patient's performance status. Possible treatments include surgery, chemotherapy, and radiotherapy Survival depends on stage, overall health, and other factors, but overall only 14% of people diagnosed with lung cancer survive five years after the diagnosis. Symptoms that may suggest lung cancer include:

- Dyspnea (shortness of breath with activity), hemoptysis (coughing up blood),
- Chronic coughing or change in regular coughing pattern, wheezing,
- Chest pain or pain in the abdomen,
- Cachxia (weight loss, fatigue, and loss of appetite), dysphonia (hoarse voice),
- clubbing of the fingernails(uncommon), dysphasia(difficulty swallowing),
- Pain in shoulder ,chest , arm, Bronchitis or pneumonia,
- Decline in Health and unexplained weight loss.

Mortality and morbidity due to tobacco use is very high. Usually lung cancer develops within the wall or epithelium of the bronchial tree. But it can start anywhere in the lungs and affect any part of the respiratory system. Lung cancer mostly affects people between the ages of 55 and 65 and often takes many years to develop. There are two major types of lung cancer. They are Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) or oat cell cancer. Each type of lung cancer grows and spreads in different ways, and is treated differently. If the cancer has features of both types, it is called mixed small cell/large cell cancer. Non-small cell lung cancer is more common than SCLC and it generally grows and spreads more slowly. SCLC is almost related with smoking and grows more quickly and form large tumors that can spread widely through the body. They often start in the bronchi near the center of the chest. Lung cancer death rate is related to total amount of cigarette smoked . Smoking cessation, diet modification, and chemoprevention are primary prevention activities. Screening is a form of secondary prevention. Our method of finding the possible Lung cancer patients is based on the systematic study of symptoms and risk factors. Non-clinical symptoms and risk factors are some of the generic indicators of the cancer diseases. Environmental factors have an important role in human cancer. Many carcinogens are present in the air breathe, the food eat, and the water drink. The constant and sometimes unavoidable exposure to environmental carcinogens complicates the investigation of cancer causes in human beings. The complexity of human cancer causes is especially challenging for cancers with long latency, which are associated with exposure to ubiquitous environmental carcinogens.

Pre-diagnosis techniques

Pre-diagnosis helps to identify or narrow down the possibility of screening for lung cancer disease. Symptoms and risk factors (smoking, alcohol consumption, obesity, and insulin resistance) had a statistically significant effect in pre-diagnosis stage. The lung cancer diagnostic and prognostic problems are mainly in the scope of the widely

discussed classification problems. These problems have attracted many researchers in computational intelligence, data mining, and statistics fields. Cancer research is generally clinical and/or biological in nature, data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications. As the use of computers powered with automated tools, large volumes of medical data are being collected and made available to the medical research groups. As a result, Knowledge Discovery in Databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict the outcome of a disease using the historical cases stored within datasets. The objective of this study is to summarize various review and technical articles on diagnosis of lung cancer. It gives an overview of the current research being carried out on various lung cancer datasets using the data mining techniques to enhance the lung cancer diagnosis.

II. LITERATURE FOR LUNG CANCER

The approach that is being followed here for the prediction technique is based on systematic study of the statistical factors, symptoms and risk factors associated with Lung cancer. Non-clinical symptoms and risk factors are some of the generic indicators of the cancer diseases. Initially the parameters for the pre-diagnosis are collected by interacting with the pathological, clinical and medical oncologists (Domain experts).

Statistical Incidence Factors:

- Age-adjusted rate (ARR)
- Primary histology
- Area-related incidence chance
- Crude incidence rate

Lung cancer symptoms:

- A cough that does not go away and gets worse overtime
- Coughing up blood (heamoptysis) or bloody mucus.
- Chest, shoulder, or back pain that doesn't go away and often is made worse by deep Hoarseness
- Weight loss and loss of appetite
- Increase in volume of sputum
- Wheezing
- Shortness of breath
- Repeated respiratory infections, such as bronchitis or pneumonia
- Repeated problems with pneumonia or bronchitis
- Fatigue and weakness
- Puffiness of face
- Loss of appetite
- Nausea and vomiting

Lung cancer risk factors:

Smoking:

- Beedi
- Cigarette
- Hukka

- Second-hand smoke
- High dose of ionizing radiation
- Radon exposure
- Occupational exposure to mustard gas chloromethyl ether, inorganic arsenic, chromium, nickel, vinyl chloride, radon asbestos
- Air pollution
- Insufficient consumption of fruits & vegetables
- Suffering with other types of malignancy

III. KNOWLEDGE DISCOVERY AND DATA MINING

Data Mining has the practice of examining large pre-existing databases in order to generate new information. The various analysis tasks that the discovery process and research areas that are promising in solving these analysis tasks.

3.1. Knowledge Discovery Process

The Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While the data mining and KDD are often treated as equivalent words but in the real data mining is an important step in the KDD process.

The Knowledge Discovery in Databases process has the steps in leading the raw data collections of new knowledge .The iterative process consists of the following steps:

- Data cleaning has noise data and irrelevant data that are removed from the collection.
- Data integration which has the multiple data sources, heterogeneous and combined in a source.
- Data transformation is a phase in which the selected data is transformed into appropriate mining procedure.
- Data mining has the crucial step for clever techniques that are applied to extract patterns.
- Pattern evaluation based on strictly interesting patterns representing knowledge on given measures.
- Knowledge representation is the final phase in which the discovered knowledge is visually to the user.

3.2. Data Mining Process

In the KDD process, the algorithms are used for extracting the data from the database. The patterns that can be discovered depend upon the data mining tasks applied. Generally, there are two types of data mining tasks. The tasks that describe the general properties of the existing data and predictive tasks that attempt to do predictions based on available data. Data mining can be done on data which are in quantitative, textual, or multimedia forms.

Data mining applications can use different kind of parameters to examine the data. They include association of patterns where one event is connected to another event), sequence or path analysis the patterns where one event leads to another event, classification to identification of new patterns with predefined targets and clustering grouping of identical or similar objects. Data mining involves some of the following key steps :

- Problem definition is based on the correct series of tools can be applied to build the model.

- Data exploration is the quality of data for an accurate model and storage for analysis all data
- Data preparation is the clean and transform the data for missing and invalid values in robust analysis.
- Modeling based on the data and the desired outcomes of algorithms is selected for analysis.
- Evaluation and Deployment analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

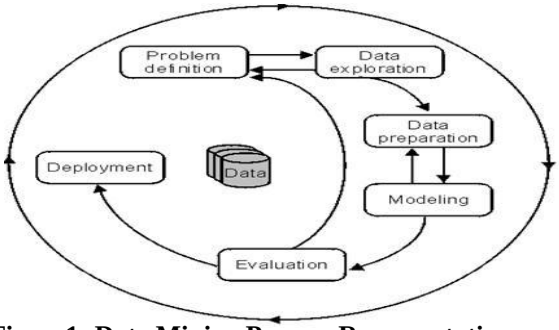

Figure1: Data Mining Process Representation

IV. LUNGS CANCER DETECTION USING ADALINE NETWORK

ADALINE networks are designed to allow the user to control the degree of similarity of patterns placed on the same cluster. However, since input patterns may differ in their level of detail, the relative similarity of input patterns placed on the same cluster can be determined by the introduction of weight vector for the cluster unit rather than the absolute difference between the input vectors. As the net is trained, each training pattern may be presented several times. Pattern may be placed in one cluster unit the first time it is presented and then placed on a different cluster when it is presented later. A stable net will not return the pattern to a pervious cluster. Some nets achieve stability by gradually reducing the learning rate as the same set of training patterns is presented many times. However, this does not allow the net to learn readily a new pattern that is presented for the first time after a number of training epochs have already taken place. The ability of a net to learn a new pattern equally well at any state of learning is called plasticity. ADALINE nets are designed to be both stable and plastic.One of the simple tasks that neural nets can be trained to perform is pattern classification. In pattern classification problems, each input vector belonging or does not belonging to a particular class or category is determined. For neural net approach that have a set of training patterns for which correct classification is known. In the simplest case considering the membership in a single class the output unit represents membership in the class having a response of 1or a response of 0 is used to indicate that a pattern is not a member of class. This method of pattern of classification done by ADALINE network is discussed below.

4.1 ADALINE Classification Operation

The ADALINE classification process consists of three major phases- recognition, comparison, and search which are explained as below.

4.1.1Recognition Phase

Initially no input vector is applied and all the components of input vector X are zero. This sets F2 to zero, thereby disabling all recognition layer neurons and causing their outputs to be zero. Because all the recognition layer neurons start out in the same state, all have an equal chance to win the subsequent competition. The pattern to be classified is now applied for each neuron in the recognition layer. A dot product is formed between its associated weight vector and the input vector. The neuron with largest dot product has weights that best match the input vector. It wins the competition and fires, inhibiting all other outputs from this layer i ,e. the ADALINE network stores a set of patterns in the weights associated with the recognition layer neurons one for each classification category.

4.1.2Comparison Phase

The single network firing in the recognition layer passed a 1 back to the comparison layer on its output signal. In accordance with the two-third rule, the only comparison layer neurons that will fire are those that receive simultaneously from the input vector X and the vector P (the comparison layer excitation vector). If there is a substantial mismatch between the X and P, few neurons in the comparison layer will fire while X contains over. This indicates that pattern P being feedback is not the one sought and the neuron firing in the recognition layer should be inhibited. A comparison of the input vector X to the inner layer vector is done and if degree of similarity is less than the vigilance parameter, the network causes reset. The effect of the reset is to force the output of the firing neuron in the recognition layer to zero, disabling it for the duration of the current classification.

4.1.3.Search Phase

If there is no rest signal generated, the match is adequate and the classification is finished. Otherwise, the other stored pattern must be researched to seek a correct match. It is seen previously in the description of ART that once a pattern is found that has reset ratio less than vigilance parameter, that pattern is inhibited and a new pattern is fed to the comparison layer. This process repeats, until one of two events occurs.

1. A stored pattern is found that matches X above the vigilance parameter.
2. All stored patterns have been tried, found to mismatch the input vector and all recognition layer neurons are inhibited.
3. The accurate detection and identification of friend or foe in battlefield environment. This is difficult because of the large amount of signal noise inherent in the signals returned from instruments line radar.
4. Motor Vehicle control, including automated string (characterizing visual input), gear selection and so on.
5. The Boeing Aeroplane company uses an ADALINE neural network system for the identification and retrieval of 2-D and 3-D representation of engineering designs. This avoids redesigning of existing parts and tools. There is an existing production database having more than 55,000 entries and 2-D database that contains more than 95,000 entries.

All the above discussed applications come under the Category of pattern classification.

4.2. Data representation Schemes

The original data is present in the form of analog values with values ranging from 0-10. The given data sets are converted to their equivalent digital form. Scaling has the advantage of mapping the desired range of variables ranging between minimum and maximum range of network input.

4.2.1Digital Conversion

Here the conversion of the given data sets into binary form is done based on certain ranges, which are defined for each attribute. The following explanation shows how the number of bits and range is taken for each attributes.There are total 10 attributes (1 class category and 9 numeric features). The 9 numerical attributes are in the analog form scaled in the range between 0 and 1. First from the given range of inputs, the minimum and maximum value is picked up and this scaling is done by the following formula.

New value (after scaling) = current value-Min value) / (Max value-Min value)

The new values obtained after truncating are converted into binary form by the following scaling. Grouping is done on the basis of range, the values in the range of 0 to 5 are assigned as '0' , and values in the range 5 to 10 are assigned as '1'. These attributes can be fed to the neural network for training and testing.

V. CONCLUSION

A prototype lung cancer disease prediction system is developed using data mining classification techniques. The system extracts hidden knowledge from a historical lung cancer disease database. The most effective model to predict patients with Lung cancer disease appears to be followed by Simulation of ADALINE network has achieved the objective of classification of group in the database of Lungs cancer according to a high level of accuracy based on different train – test ratios. Thus for unsupervised input pattern, the output is obtained with a good level of accuracy.

VI. FEATURE WORK

Further development in the simulation can be obtained by increasing the level of accuracy by training more patterns. Lung cancer prediction system can be further enhanced and expanded. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases. Another challenge would be to integrate data mining and text mining .The simulated for pattern classification of breast cancer data using the ADALINE learning and testing algorithm, and about 92% accuracy is obtained. It can be safely concluded that ADALINE is the best-suited network for pattern classification as of today.

VII.REFERENCES

[1] Sang Min Park, Min Kyung Lim, Soon Ae Shin & Young Ho Yun 2006. Impact of prediagnosis smoking, Alcohol, Obesity and Insulin resistance on survival in Male cancer Patients: National Health Insurance corporation study. Journal of clinical Oncology, Vol 24 Number 31 November 2006.

[2] Osmar R. Zaïane, Principles of Knowledge Discovery in Databases. [Online]. Available:webdocs.cs.ualberta.ca/~zaiane/courses/cmput690 /notes/Chapter1/ch1.pdf.

[3] Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers.

[4] Maria-Luiza Antonie, Osmar R. Za¨ıane, Alexandru Coman Application of Data Mining Techniques for Medical Image Classification.Page 97

[5] Bayesian Networks. 2004, London: Pearson Prentice Hall.

[6] www.webmd.com/lung-cancer/tc/lung-cancer-references

[7] *https://www.crcpress.com/Textbook-of-Lung-Cancer-Second.../book/978041538510.*

[8] https://www.elsevier.com/journals/lung-cancer/0169-5002/guide-for-authors

[9] *https://sebastianraschka.com/faq/docs/diff-perceptron-adaline-neuralnet.html*

[10] Krishnapuram, B., et al., A Bayesian approach to joint feature selection and classifier design. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004. 6(9): p. 1105-1111.