International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 11 | Nov 2020 www.irjet.net p-ISSN: 2395-0072

A STUDY OF CANCER PREDICTION USING NEURAL NETWORK

Austin Titty Philip[1], Shashank Pal[2], Aarohi Verma[3]

[1]UG Student- College of Engineering Munnar, Kerala, India
[2] UG Student- SRM Institute of Science and Technology, Chennai, India
[3]UG Student- SRM Institute of Science and Technology, Chennai, India

---

Abstract - *Recent advances in both the biological and computer sciences have spurred researchers to pay greater attention to the role of computational methods in the broad sphere of cancer research. Specific focus has been given to the demonstrated benefits of artificial intelligence (AI) and machine learning approaches when compared to current methods for the diagnosis and treatment of cancer. An artificial neural network is a form of AI based on algorithms that mimic human brain function. Neural networks are especially useful in the interpretation of nonlinear data, which is commonly encountered in biological research studies. Neural networking technologies may be used to diagnose cancer more easily and effectively than traditional methods as they decrease the need for invasive procedures and interpreting the results of imaging methods. Additionally, neural networks have been trained to analyze individual prognoses and treatment plans with an accuracy comparable to that of experienced physicians. Advances such as these aid both medical professionals and patients in making optimal health care decisions. As large-scale computing initiatives – such as the recent Microsoft project aiming to "solve" cancer with computer science – move forward, it has become increasingly apparent that the future of medical research will involve technologies such as neural networking and other forms of AI.*

Key Words*: *Machine Learning, AI, Neural Networks

1. INTRODUCTION

Cancer is the most serious health problem .With increase in the death rate due to cancer day by day. The possible treatments are surgery, chemotherapy and radiotherapy. The earlier detection gives the higher the chances of successful treatment. More people die due Lung cancer and other types of cancer such as breast, skin, colon, prostate cancers. There is significant proof indicating that the earlier detection of cancer will decrease the mortality rate. The latest estimate according to the recent survey provided by world health organization indicates that around 7.6 million dead‟s in whole world. There are many techniques to diagnose cancer, like just radiography (x-ray), computed tomography(CT) ,magnetic resonance imaging (MRI Scan).But ,most of these techniques are costly and time consuming and most of these techniques are detecting the cancer in its advanced stages. Hence, there is a great need of new technology to diagnose cancer in its early stages. Deep learning technique is used to detect cancer in its early stage. Deep learning is a part of broader family of machine learning methods based on learning data representation as opposed to task specific algorithms. Use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation .Each successive layer uses the output from the previous layer as input. Neural network offers a completely different approach to problem solving and they are sometimes called the sixth generation computing. To apply neural network and their associated technique, Deep Learning to health care, Specifically management of cancer patients.

2. LITERATURE SURVEY

2.1 TITLE

Nikhil R. Pal, "A fuzzy rule based approach to identify biomarkers for diagnostic classification of cancers", 2007

A method for simultaneous feature/gene selection and fuzzy rule extraction is introduced. Since the feature selection method is integrated into the rule base tuning, it can account for possible subtle nonlinear interaction between features as well as that between features and the tool, and hence can identify a useful set of features for the task at hand. For this data set, a neural network is used to reduce the dimension of the data and then applied the method to find biomarkers and rules. The system could find only eight genes including a novel gene that can do the diagnostic prediction task with a high accuracy. The system can be extended to non-classification applications also. Unlike other feature selection methods used in connection with fuzzy rules, the method can account for interaction between features as well as that between features and the tool, and hence can find a small set of important features. The method is applied to identify a small set of biomarkers for diagnostic prediction of the SRBCT group of childhood cancer. But depending on the learning parameters and the initial rule base different trials of the algorithm may result in different sets of important features because gradient descent search usually settles at a local minima. To decide on the number of useful features, a threshold for the modulator values has to be chosen. Further investigation is needed to come up with useful guidelines for this. Also investigations need to be done to analyze the performance of the algorithm when the dimension of the input is very high, say a few thousand

2.2. TITLE

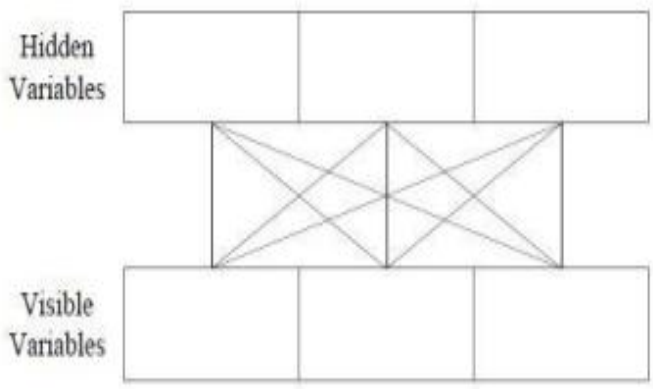
Yuchun Tang, Yan-Qing Zhang, Zhen Huang, Xiaohua Hu, Yichuan Zha,“ Recursive fuzzy granulation for gene subsets extraction and Cancer Classification”,2008

GENE expression microarrays (including the cDNA microarray and the GeneChip) have been developed as a powerful technology for functional genetics studies that simultaneously measures the mRNA expression levels of thousands to tens of thousands of genes. A typical microarray expression experiment monitors the expression level of each gene multiple times under different conditions or in different phenotypes. To select multiple highly informative gene subsets for cancer classification and diagnosis, a new *Fuzzy Granular Support Vector Machine—Recursive Feature Elimination* algorithm (FGSVM-RFE) is designed in the paper. As a hybrid algorithm of statistical learning, fuzzy clustering, and granular computing, the FGSVM-RFE separately eliminates irrelevant, redundant, or noisy genes in different granules at different stages and selects highly informative genes with potentially different biological functions in balance. Moreover, the FGSVM-RFE can extract multiple gene subsets on each of which a classifier can be modeled with 100% accuracy. Specifically, the independent testing accuracy for the prostate cancer dataset is significantly improved. The previous best result is 86% with 16 genes and in this case, best result is 100% with only eight genes. The identified genes are annotated by Onto-Express to be biologically meaningful. First, the FGSVM-RFE utilizes RI metrics for gene prefiltering to remove most of the irrelevant genes. Second, the FGSVMRFE explicitly groups genes with similar expression patterns into function granules by recursively clustering with the FCM. An SVM-based ranking is thereafter carried out in each granule to safely remove lower-ranked genes as redundant genes because higher informative genes with similar functions still survive. Finally, the FGSVM-RFE deals with complex correlations among genes by assigning a gene into several granules with different fuzzy membership values so that a really informative gene can achieve more than one opportunity to be extracted. The newly identified genes by the algorithm need to be further confirmed biologically, which may generate more insights for cancer mechanism, treatment, and study. Nevertheless, the extraction of these cancer-related gene subsets may help to stimulate and guide detailed cancer studies on the gene functions.

2.3 TITLE

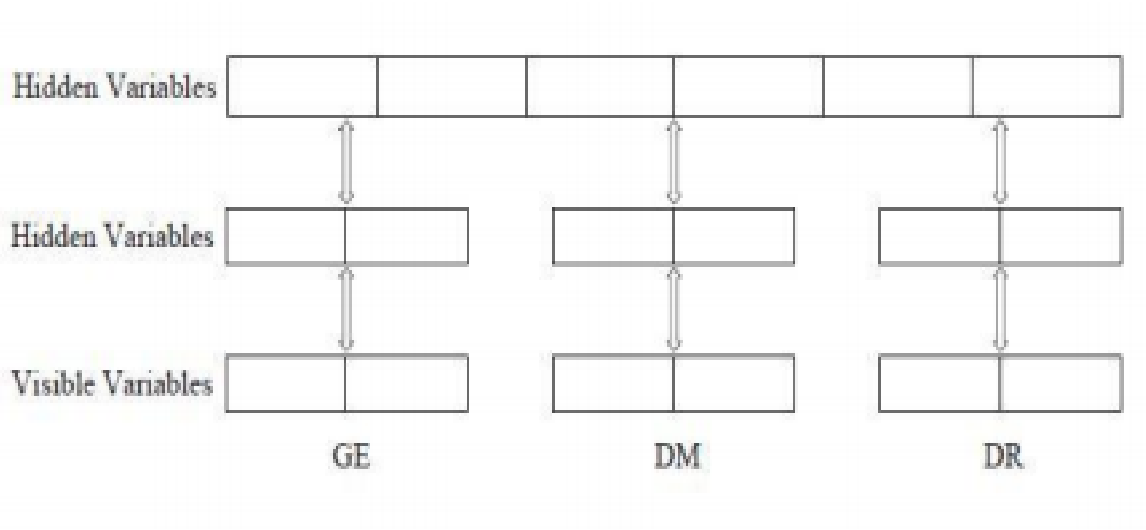
Muxuan Liang, Zhizhong Li, Ting Chen and Jianyang Zeng,”Integrative Data Analysis of Multi-platform Cancer Data with a Multimodal Deep Learning Approach”, 2014

Multimodal deep belief network (DBN) is used to cluster cancer patients from multi-platform observation data. In the integrative clustering framework, relationships among inherent features of each single modality are first encoded into multiple layers of hidden variables. Then a joint latent model is employed to fuse common features derived from multiple input modalities. A practical learning algorithm, called contrastive divergence (CD), is applied to infer the parameters of our multimodal DBN model in an unsupervised manner. Tests on two available cancer datasets show that the integrative data analysis approach can effectively extract a unified representation of latent features to capture both intra- and cross-modality correlations, and identify meaningful disease subtypes from multi-platform cancer data. In addition, the approach can identify key genes and mi RNAs that may play distinct roles in pathogenesis of different cancer subtypes. Methods:



(1) Restricted Boltzmann Machines

A restricted Boltzmann machine (RBM) is an undirected graphical model which consists a layer of visible variables vi, i = 1, ...,m, and a layer of hidden variables hj , j = 1, ..., g, where m is the number of visible variables and g is the number of hidden variables. Here, the number of visible variables is equal to the number of genomic profiles per patient. In an RBM model, each visible variable is connected to every hidden variable, but no connection is allowed between any two variables within the same layer (Fig. 1).



(2) Multimodal Deep Belief Networks

A multimodal DBN is a network of stacked RBMs, in which the separate RBMs at the bottom level take multimodal data as input, and the top level RBMs contain hidden variables that represent the common features across different modalities from multi-platform data. To further illustrate a multimodal DBN framework, we use a specific example (see Fig. 2) of cancer data analysis, in which input data involve three modalities, including gene expression, DNA methylation and drug response. In this example, gene expression and DNA methylation are both represented as real-valued vectors, while drug response is represented as a binary vector.

2.4 TITLE

Rasool Fakoor, Faisal Ladhak, Azade Nazi, Manfred Huber, "Using deep learning to enhance cancer diagnosis and classication", 2014

In this paper, we show that how unsupervised feature learning can be used for cancer detection and cancer type analysis from gene expression data. The main advantage of the proposed method over previous cancer detection approaches is the possibility of applying data from various types of cancer to automatically form features which help to enhance the detection and diagnosis of a specific one. The technique is here applied to the detection and classification of cancer types based on gene expression data. In our method, we first reduce the dimensionality of the feature space using PCA, and then apply the result of PCA as a compressed feature representation which still encodes the data available in the sample set, along with some randomly selected original gene expressions (i.e. original raw features) as a more compact feature space to either a one or a multi-layered sparse auto-encoder to find a sparse representation for data that will then be used for the classification. This overall approach to building and training a system to detect and classify cancer from gene expression data is shown in Figure 1. As shown in the figure, the approach proposed here consists of two parts, the feature learning phase and the classifier learning phase.

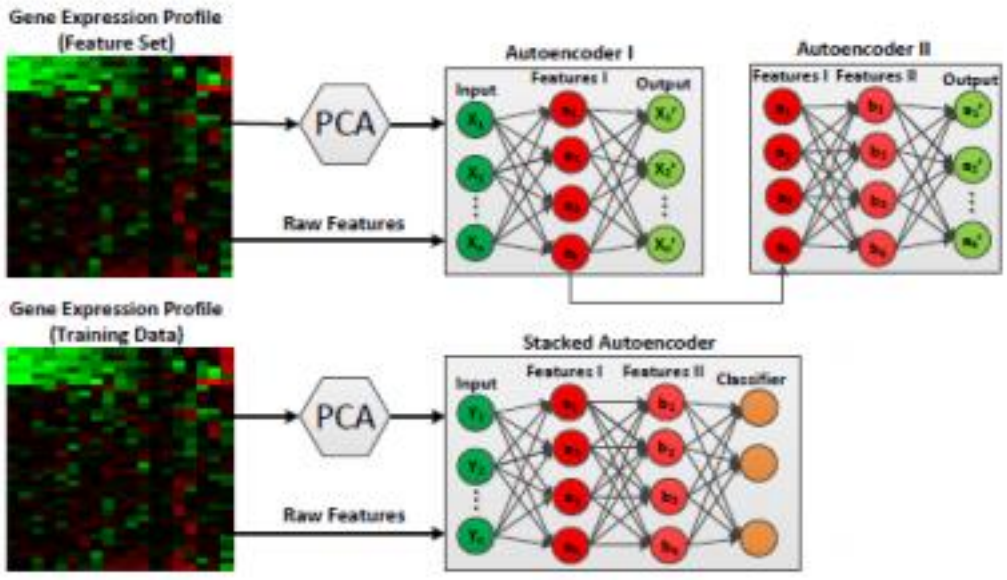


Figure 1. Overall Approach.

2.5 TITLE

Vishwanath S. Mahalle, Ms. Vidhi L. Chawda,"Learning to recommend descriptive tags for Health Seekers using Deep Learning",2017

In this work we first analyze and categorize needs of health seekers and ask for manifested symptoms for disease prediction. Then user will search for their query. Then the query get processed to give prediction of disease to the user or health seekers. Here the concept of hidden layers are being used. First medical signatures mines from raw features. These features and signatures deems as input node in one layer and hidden nodes in subsequent layer.

We proposed and develop the scheme who studies the user information and health related data. In our application, the user request is compared with the different dataset. The datasets are automatically created using the discussion of doctors. The doctors will discuss on various medical concepts. The system will save the medical information in the form of datasets. So one thing here we can say that, according to the doctors conversation the datasets inside the database get automatically updated. In the existing system, the admin has to manually update the information or we can say that he/she has to update the information about the new diseases, their symptoms, causes, precautions, etc. The main Objectives of this paper are:

A) To provide more accurate information to the userfor their query:

In proposed system, every user will get more accurate information for their query. When compared with existing system, the proposed system gets more perfect result for users query. This will help the user for their disease prediction and the user can move to further procedure immediately as they get accurate disease inference.

B) To infer the possible diseases, given the questions of health seekers: Proposed system identifies discriminant feature for specific disease. This means that there is no confusion in disease prediction according to users query. So that user can get accurate answer for their query. Distinguished answers will obtain to the user for the query. This means that for disease prediction proper query must be fired by health seekers. This will help the system to give proper prediction for user queries.

C) The system can remove the obstacles such as the vocabulary gap, incomplete information, correlated medical concepts:

There are various obstacles in the existing system like Vocabulary gap, incomplete information, correlated medical concepts. These obstacles are overcome in our proposed system. This will help the user to find the spellings of the diseases that they don‟t know. When we talk about the incomplete information, it means that only limited information is there in the system. Sometimes, this irritates the user and indirectly will affect on the reviews of the system.

D) System can able to find out the proper solution for users query:

All these objectives are key points in this proposed system, to get better results to the user so that every user gets satisfied results according to the query that user fired. So our system can be able to find out proper solution on the users query. This paper aims to build a disease inference scheme that is able to automatically infer the possible diseases of the given questions in community-based health services.

2.6 TITLE

Naresh Gowda M, Abdul Imran Rasheed," Hardware Implementation of Hybrid Classifier to Detect Cancer Cells", 2017

Image classification plays an important role in the field of medical applications. Classifiers are important in the field of medical science to detect cancer cells. The early detection of disease by giving proper medical treatment might save the life of the patient from danger. The image classification is necessary step for content analysis of multi-media. There are many technologies have been used for image classification such as clustering, decision tree method and neural network .Convolutional Neural Network and Support Vector Machine are the common machine learning algorithms for classification. Both the classification techniques are fused to develop a hybrid classifier. The hardware implementation of hybrid classifier can improve the performance of the system and reduce the power consumption for real time applications. The main aim of this study is early detection of cancer cell using hybrid classifier implemented on a low cost handheld device. A hardware design is proposed to implement hybrid classifier on FPGA. The proposed system is implemented on Zync SoC FPGA platform and it gives the high performance, low power consumption and low hardware utilization**.** The contribution of proposed paper is on hardware implementation of hybrid classifier to detect leukemia and melanoma cancer cell. The proposed system is implemented on Zync SoC platform and it gives the high performance, low power consumption and low hardware utilization. Compare to other implementations in the literature our hybrid classifier method implemented on Zync SoC platform is considered to be a low power system. The proposed system achieving an efficient hybrid classification technique to detection of cancer cell with high performance, low power consumption, cost and area. For future work, the proposed hybrid classifier can use for different image classification applications. In addition, the proposed hardware implementation of hybrid classifier on Zync SoC platform might be useful for other hardware designer to develop their classification system on hardware.

2.7. TITLE

Seema Singh, Sunita Saini, Mandeep Singh, "Cancer Detection Using adaptive Neural Network", 2017

Computing systems whose central theme is borrowed from the analogy of biological neural networks. Artificial neural networks are also referred to as ‚neural nets‘„ artificial neural systems‘, ‚parallel distributed processing systems‘, and ‚connectionist systems‘. We presented a system to detect that whether patient‟s cancer stage is malignant or benign. Neural networks are used to perform this task. Most cancer detection techniques rely on historical data. Experiments suggest that a cancer case currently under observation is malignant if it has the same or similar properties, as a malignant case that has been detected as malignant earlier in the same environment. Therefore, historical information helps us to detect whether a case is malignant or benign. In neural networks historical data which is obtained by regression analysis or by any other means are fed to network and the network gets trained accordingly that data is called training data. When we input operational data then network respond accordingly the trained data. So we used Adaptive Resonance Neural Networks (ARNN) which clusters already known modules which are faulty and fault-free. As it works under unsupervised learning so there is no distinction between training data and operational data. We used data from UCI machine learning data repository titled as „Wisconsin Breast Cancer Database‟ which contains 9 attributes of 699 patients under survey. On the basis of these 9 attributes cases are termed as benign or malignant. The ARNN clustrifies cases which are benign in one cluster and malignant in another cluster. We had applied ART2 model of adaptive resonance theory to cluster the dataset as it can be applied on continuous dataset. Clustering of the ARNN expressed reality interpretable knowledge about the dataset, and may be usable by the practitioners in a variety of ways. The result of this model is better the other models in case of accuracy. The ARNN can be tested on other datasets and variations in the neural networks itself need to be explored. In a developed software system, it would be desirable to improve the accuracy, precision and recall.

2.8 TITLE

David AH," A Deep learning approach for cancer detection and relevant gene identification", 2017

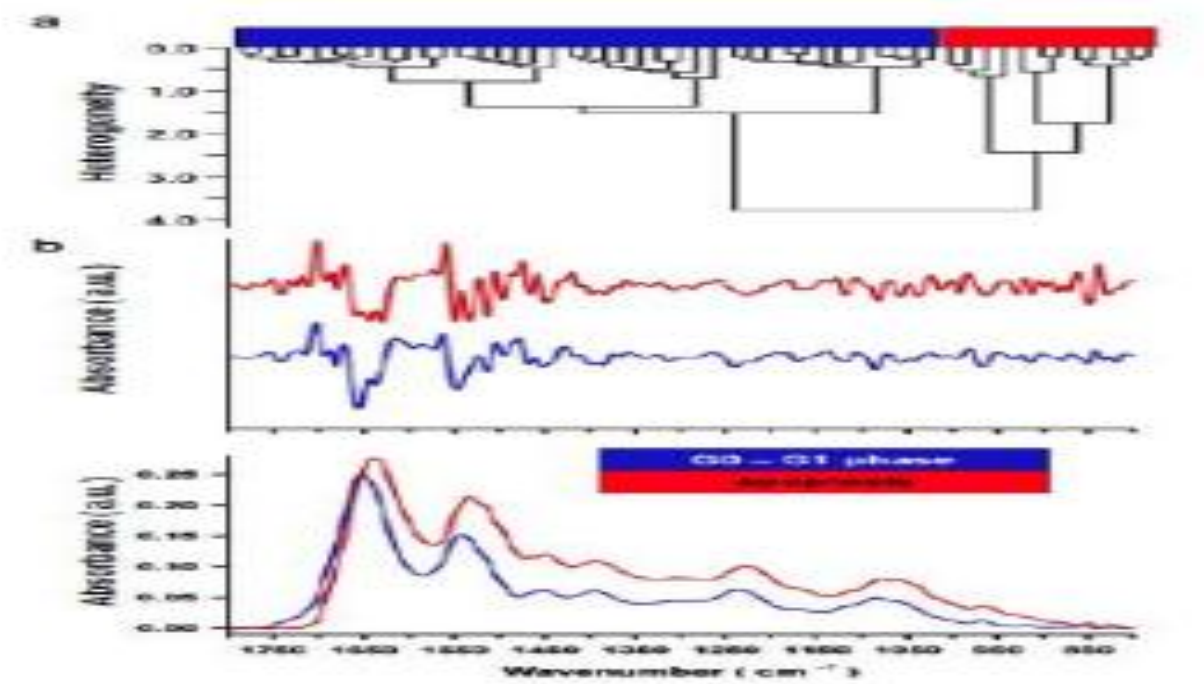
The analysis of gene expression data has the potential to lead to significant biological discoveries. Much of the work on the identification of differentially expressed genes has focused on the most significant changes, and may not allow recognition of more subtle patterns in the data. Machine learning methods for dimensionality reduction and classification of gene expression data have achieved some success, but there are limitations in the interpretation of the most significant signals for classification purposes. After decades of research there is still uncertainty in the clinical diagnosis of cancer and the identification of tumor-specific markers. Here we present a deep learning approach to cancer detection, and to the identification of genes critical for the diagnosis of breast cancer. First, we used Stacked „Denoising Auto encoder" (SDAE) to deeply extract functional features from high dimensional gene expression profiles. We were able to use the weights of this model to extract genes that were also useful for cancer prediction, and have potential as biomarkers or therapeutic targets. Next, we evaluated the performance of the extracted representation through supervised classification models to verify the usefulness of the new features in cancer detection. Lastly, we identified a set of highly interactive genes by analyzing the SDAE connectivity matrices. One limitation of deep learning approaches is the requirement for large data sets, which may not be available for cancer tissues. We expect that as more gene expression data becomes available, this model will improve in performance and reveal more useful patterns. Accordingly, deep learning models are highly scalable to large input data. Future work is needed to analyze different types of cancer to identify cancer-specific biomarkers. In addition, there is potential to identify cross-cancer biomarkers through the analysis of aggregated heterogeneous cancer data.

2.9 TITLE

SeungJin Lim,"FT-IR spectra analysis towards cancer detection", 2014

Since the middle of 20(th) century infrared (IR) spectroscopy coupled to microscopy (IR microspectroscopy) has been recognized as a non destructive, label free, highly sensitive and specific analytical method with many potential useful applications in different fields of biomedical research and in particular cancer research and diagnosis. Although many technological improvements have been made to facilitate biomedical applications of this powerful analytical technique, it has not yet properly come into the scientific background of many potential end users. Therefore, to achieve those fundamental objectives an interdisciplinary approach is needed with basic scientists, spectroscopists, biologists and clinicians who must effectively communicate and understand each other's requirements and challenges. In this review we aim at illustrating some principles of Fourier transform (FT) Infrared (IR) vibrational spectroscopy and microscopy (microFT-IR) as a useful method to interrogate molecules in specimen by mid-IR radiation. Penetrating into basics of molecular vibrations might help us to understand whether, when and how complementary information obtained by microFT-IR could become useful in our research and/or diagnostic activities. MicroFT-IR techniques allowing to acquire information about the molecular composition and structure of a sample within a micrometric scale in a matter of seconds will be illustrated as well as some limitations will be discussed. How biochemical, structural, and dynamical information about the systems can be obtained by bench top microFT-IR instrumentation will be also presented together with some methods to treat and interpret IR spectral data and applicative examples. The mid-IR absorbance spectrum is one of the most information-rich and concise way to represent the whole "..omics" of a cell and, as such, fits all the characteristics for the development of a clinically useful biomarker.

Infrared Spectroscopy is the analysis of infrared light interacting with a molecule. This can be analyzed in three ways by measuring absorption, emission and reflection. The main use of this technique is in organic and inorganic chemistry. It is used by chemists to determine functional groups in molecules. IR Spectroscopy measures the vibrations of atoms, and based on this it is possible to determine the functional groups.5 Generally, stronger bonds and light atoms will vibrate at a high stretching frequency (wavenumber).



2.10 TITLE

Li Shen and Eng Chong Tan, "Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data", 2005

The use of penalized logistic regression for cancer classification using microarray expression data is presented. Two dimension reduction methods are respectively combined with the penalized logistic regression so that both the classification accuracy and computational speed are enhanced. Two other machine-learning methods, support vector machines and least squares regression, have been chosen for comparison. It is shown that our methods have achieved at least equal or better results. They also have the advantage that the output probability can be explicitly given and the regression coefficients are easier to interpret. Several other aspects, such as the selection of penalty parameters and components, pertinent to the application of our methods for cancer classification are also discussed. When you have multiple variables in your logistic regression model, it might be useful to find a reduced set of variables resulting to an optimal performing model.

Penalized logistic regression imposes a penalty to the logistic model for having too many variables. This results in shrinking the coefficients of the less contributive variables toward zero. This is also known as regularization. Regression problems with many potential candidate predictor variables occur in a wide variety of scientific fields and business applications. These problems require you to perform statistical model selection to find an optimal model, one that is as simple as possible while still providing good predictive performance. Traditional stepwise selection methods, such as forward and backward selection, suffer from high variability and low prediction accuracy, especially when there are many predictor variables or correlated predictor variables (or both). In the last decade, the higher prediction accuracy and computational efficiency of penalized regression methods have made them an attractive alternative to traditional selection methods. This paper first provides a brief review of the LASSO, adaptive LASSO, and elastic net penalized model selection methods. Then it explains how to perform model selection by applying these techniques with the GLMSELECT procedure, which includes extensive customization options and powerful graphs for steering statistical model selection.

2.11 TITLE

Mr. S.Sivakumar, D.Viji , "Cancer Prediction Using Mining Gene Expression Data",2017

Cancer is a major cause of all natural mortalities and morbidities throughout the world. Pointed out the exact tumour types provides an optimized solution for the better treatment and toxicity minimization due to medicines on the patients. To get a clear picture on the insight of a problem, a clear cancer classification analysis system needs to be pictured followed by a systematic approach to analyse global gene expression which provides an optimized solution for the identified problem area. Molecular diagnostics provides a promising option of systematic human cancer classification, but these tests are not widely

applied because characteristic molecular markers for most solid tumor save yet to be identified. Recently, DNA microarray-based tumor gene expression profiles have been used for cancer diagnosis. Existing system focused in ranging from old nearest neighbour analysis to support vector machine manipulation for the learning portion of the classification model. We don‟t have a clear picture of supervised classifier (Supervised Multi Attribute Clustering Algorithm) which can manage knowledge attributes coming two different knowledge streams. Our proposed system takes the input from multiple source, create an ontological store, cluster the data with attribute match association rule and followed by classification with the knowledge acquired

2.12 TITLE

B Satheesh kumar,P Sharanya,"A Survey on Feature Selection of Cancer Disease Using Data Mining Techniques ",2016

This paper presents a survey on medical image feature selection using data mining techniques. In medical field there are different kinds of problem in medical imaging like classification, segmentation, extraction and selection. Medical datasets are often categorized by huge amount of disease measurements and comparatively small amount of patient records. These measurements (feature selection) are not relevant, where this irrelevant and redundancy features are difficult to evaluate. On the other hand, the large number of features may cause the problem of memory storage in order to represent the data set. Different kinds of data mining techniques (or) algorithms can convenient with imprecision and uncertainty in data analysis and can effectively remove noisy and redundant information. Data Mining is the process of discovering interesting knowledge from large amounts of data stored in database. It is an essential process where intelligent methods are applied in order to extract data patterns. Simply stated, data mining refers to "extracting" or "mining" knowledge from large number of data.[1] The major reason for using data mining has attracted a great deal of concentration in the information industry. Growth of the information technology has a huge amount of data and the imminent need for revolving such data into useful information and knowledge. These kinds of information gathered from business management, engineering design, science exploration and medical fieldImage Mining plays vital role among the researchers in the field of data mining. The image mining mainly focuses the extraction of pattern from large collection of images. Now a day, the rapid development of digital medical device, medical information data bases have included not only the structural information of patients, it also have non - structural medical image information. Even medical images give a small picture of organ in the body. It may take more time to extract the useful information. Because the medical researchers consider effectiveness of the treatment depends upon the information in the medical data base.

2.13 TITLE

Joseph A. Cruz and David S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis", 2006

Machine learning is not new to cancer research. Artificial neural networks (ANNs) and decision trees (DTs) have been used in cancer detection and diagnosis for nearly 20 years (Simes 1985; Maclin et al. 1991; Ciccheti 1992). Today machine learning methods are being used in a wide range of applications ranging from detecting and classifying tumors via X-ray and CRT images (Petricoin and Liotta 2004; Bocchi et al. 2004) to the classification of malignancies from proteomic and genomic (microarray) assays (Zhou et al. 2004; Dettling 2004; Wang et al. 2005). According to the latest PubMed statistics, more than 1500 papers have been published on the subject of machine learning and cancer. However, the vast majority of these papers are concerned with using machine learning methods to identify, classify, detect, or distinguish tumors and other malignancies. In other words machine learning has been used primarily as an aid to cancer diagnosis and detection. It has only been relatively recently that cancer researchers have attempted to apply machine learning towards cancer prediction and prognosis. As a consequence the body of literature in the field of machine learning and cancer prediction/prognosis is relatively small .

The fundamental goals of cancer prediction and prognosis are distinct from the goals of cancer detection and diagnosis. In cancer prediction/prognosis one is concerned with three predictive foci: 1) the prediction of cancer susceptibility (i.e. risk assessment); 2) the prediction of cancer recurrence and 3) the prediction of cancer survivability. In the first case, one is trying to predict the likelihood of developing a type of cancer prior to the occurrence of the disease. In the second case one is trying to predict the likelihood of redeveloping cancer after to the apparent resolution of the disease. In the third case one is trying to predict an outcome (life expectancy, survivability, progression, tumor-drug sensitivity) after the diagnosis of the disease. In the latter two situations the success of the prognostic prediction is obviously dependent, in part, on the success or quality of the diagnosis. However a disease prognosis can only come after a medical diagnosis and a prognostic prediction must take into account more than just a simple diagnosis (Hagerty et al. 2005).

2.14 TITLE

P.Ramachandran, N.Girija , T.Bhuvaneswari , "Early Detection and Prevention of Cancer using Data Mining Techniques", 2017

Cancer is one of the most common diseases in the world that results in majority of death. Cancer is caused by uncontrolled growth of cells in any of the tissues or parts of the body. Cancer may occur in any part of the body and may spread to several other parts. Only early detection of cancer at the benign stage and prevention from spreading to other parts in malignant stage could save a person‟s life. There are several factors that could affect a person‟s predisposition for cancer. Education is an important indicator of socioeconomic status through its association with occupation and life-style factors. A number of studies in developed countries have shown that cancer incidence varies between people with different levels of education. A high incidence of breast cancer has been found among those with high levels of education whereas an inverse association has been found for the incidence of cancers of the stomach, lung and uterine cervix. Such differences in cancer risks associated with education also reflect in the differences in life-style factors and exposure to both environmental and work related carcinogens. This study describes the association between cancer incidence pattern and risk levels of various factors by devising a risk prediction system for different types of cancer which helps in prognosis. Data mining technique involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods in early detection of cancer. In classification learning, the learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples. In association learning, any association among features is sought, not just ones that predict a particular class value. In clustering, groups of examples that belong together are sought. In numeric prediction, the outcome to be predicted is not a discrete class but a numeric quantity. In this study, to classify the data and to mine frequent patterns in data set Decision Tree algorithm is used. A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label. The top most node is the root node. The attribute value of the data is tested against a decision tree. A path is traced from root to leaf node, which holds the class prediction for that data. Decision trees can be easily converted into classification rules. This decision tree is used to generate frequent patterns in the dataset. The data and item sets that occur frequently in the data base are known as frequent patterns. The frequent patterns that is most significantly related to specific cancer types and are helpful in predicting the cancer and its type is known as Significant frequent pattern. Using this significant patterns generated by decision tree the data set is clustered accordingly and risk scores are given. Clustering is a process of separating dataset into subgroups according to their unique features. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. In K- means clustering, the number of clusters needed is found out and then an algorithm is used to successively associate or dissociate instances with clusters until associations stabilize around k clusters. In this research all the above mentioned Data Mining techniques are implemented together to create a novel method to diagnose the existence of cancer for a particular patient. When beginning to work on a data mining problem, it is first necessary to bring all the data together into a set of instances. Integrating data from different sources usually presents many challenges. The data must be assembled, integrated, and cleaned up. Only then it can be used for processing through machine learning techniques. This developed system can be used by physicians and patients alike to easily know a person‟s cancer status and severity without screening them for testing cancer. Also it is useful to record and save large volumes of sensitive information which can be used to gain knowledge about the disease and its treatment.

2.15 TITLE

Seema Singh, Sunita Saini , Mandeep Singh , "Cancer Detection Using adaptive Neural Network",2017

Computing systems whose central theme is borrowed from the analogy of biological neural networks. Artificial neural networks are also referred to as ‚neural nets‘„artificial neural systems‘, ‚parallel distributed processing systems‘, and ‚connectionist systems‘. We presented a system to detect that whether patient‟s cancer stage is malignant or benign. Neural networks are used to perform this task. Most cancer detection techniques rely on historical data. Experiments suggest that a cancer case currently under observation is malignant if it has the same or similar properties, as a malignant case that has been detected as malignant earlier in the same environment. Therefore, historical information helps us to detect whether a case is malignant or benign. In neural networks historical data which is obtained by regression analysis or by any other means are fed to network and the network gets trained accordingly that data is called training data. When we input operational data then network respond accordingly the trained data. So we used Adaptive Resonance Neural Networks (ARNN) which clusters already known modules which are faulty and fault-free. As it works under unsupervised learning so there is no distinction between training data and operational data. We used data from UCI machine learning data repository titled as „Wisconsin Breast Cancer Database‟ which contains 9 attributes of 699 patients under survey.

On the basis of these 9 attributes cases are termed as benign or malignant. The ARNN clustrifies cases which are benign in one cluster and malignant in another cluster. We had applied ART2 model of adaptive resonance theory to cluster the dataset as it can be applied on continuous dataset. Clustering of the ARNN expressed reality interpretable knowledge about the dataset, and may be usable by the practitioners in a variety of ways. The result of this model is better the other models in case of accuracy. The ARNN can be tested on other datasets and variations in the neural networks itself need to be explored. In a developed software system, it would be desirable to improve the accuracy, precision and recall. Adaptive resonance theory (ART) models are a neural network that performs clustering, and can allow the number of clusters to vary with the size of the problem. The major differences between ART and other clustering methods is that ART allows the user to control the degree of similarity between members of the same cluster by means of a user-defined constant called the vigilance parameter. It uses a MAXNET. A ‚maxnet' is a recurrent one-layer network that conducts a competition to determine which node has the highest initial input value. Apart from neural plausibility and the desire to perform every task using neural networks, it can be argued that the maxnet allows for greater parallelism in execution than the algebraic solution.

3. PROBLEM STATEMENT

To accurately assess cancer risk in average- and high-risk individuals and determining cancer prognosis in patients by inputting scan reports, daily routines, food habits etc.

4. METHODOLOGY

Artificial Neural Networks (ANNs) are computer systems that operate by using algorithms based on human brain functions to interpret nonlinear data, which does not follow a sequential pattern. ANNs are composed of many smaller units called neurons, which are organized into multiple layers between the input of data and the output of results. In the same fashion as biological neurons, the connections and patterns between these units determine the behavior of the network, and this behavior may be learned through a process called backpropagation (Drew & Monson, 2000). In backpropagation, a data set for which the correct output is already known is input into the network. The output of the ANN is then continuously compared to the known output, and adjustments are made to the pattern of network components to minimize the least mean square difference over the entire set. After a sufficient number of repetitions, the network adjusts to a higher level of accuracy, and can perform complex tasks without requiring the as many computing resources as traditional methods (Drew & Monson, 2000). Researchers are now discovering that biomedical systems must increasingly be represented by nonlinear systems, making ANNs a valuable computing resource for biological research. ANNs have been applied to various aspects of cancer medicine for decades. Yet, more recent research efforts in this field have been accompanied by new knowledge about the biological aspects of cancer (Cruz & Wishart, 2006). As a result, computational methods are becoming significantly more effective than ever before.

5. CONCLUSION

Neural Network Algorithm is implemented using open source and its performance is compared to other classification algorithms. It shows the best results with providing highest TP Rate and lowest FP Rate and in case of correctly classification, it gives the 96.04% result as compare to other classifiers. Neural network model is a diagnostic system that performs at an accuracy level is constructed. In this process, the performance of neural network structure was investigated for cancer diagnosis problem. People can be checked for cancer disease quickly and painlessly and thus detecting the disease at an early stage. This indicates that neural network can be effectively used for cancer diagnosis to help oncologists. The prediction could help doctor to plan for a better medication and provide the patient with early diagnosis. The million order dataset can be selected and image classification can be done on larger dataset. With improved size of dataset various issues such as uploading data, managing feature set, increased execution time of classification algorithms etc. could be considered. More image features can be extracted for better classification. Various combinations of previous features can be used to correctly classify medical data.

REFERENCES

1. Nikhil R. Pal, “A fuzzy rule based approach to identify biomarkers for diagnostic classification of cancers”,2007
2. Yuchun Tang, Yan-Qing Zhang, Zhen Huang, Xiaohua Hu, Yichuan Zha,“ Recursive fuzzy granulation for gene subsets extraction and Cancer Classification”,2008
3. Muxuan Liang, Zhizhong Li, Ting Chen and Jianyang Zeng,”Integrative Data Analysis of Multi-platform Cancer Data with a Multimodal Deep Learning Approach”,2014
4. Rasool Fakoor ,Faisal Ladhak ,Azade Nazi, Manfred Huber,”Using deep learning to enhance cancer diagnosis and classication”,2014

5. Vishwanath S. Mahalle, Ms. Vidhi L. Chawda,"Learning to recommend descriptive tags for Health Seekers using Deep Learning",2017
6. Naresh Gowda M, Abdul Imran Rasheed," Hardware Implementation of Hybrid Classifier to Detect Cancer Cells",2017
7. David AH," A Deep learning approach for cancer detection and relevant gene identification ", 2017
8. Seema Singh, Sunita Saini , Mandeep Singh , "Cancer Detection Using adaptive Neural Network",2017
9. SeungJin Lim," FT-IR spectra analysis towards cancer detection", 2014
10. Li Shen and Eng Chong Tan, "Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data",2005
11. Mr. S.Sivakumar, D.Viji , "Cancer Prediction Using Mining Gene Expression Data",2017
12. B Satheesh kumar,P Sharanya,"A Survey on Feature Selection of Cancer Disease Using Data Mining Techniques ",2016
13. Joseph A. Cruz and David S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis", 2006
14. P.Ramachandran , N.Girija , T.Bhuvaneswari , "Early Detection and Prevention of Cancer using Data Mining Techniques ", 2017
15. Seema Singh, Sunita Saini , Mandeep Singh , "Cancer Detection Using adaptive Neural Network",2015