International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Cancer Gene Prediction and Computer Aided Diagnosis for Cancer

Tejal V. Gudadhe[1], Namrata D. Ghuse[2]

[1]M.E. I[st] year (CSE), P. R. Pote COET, Amravati, Maharashtra, India

[2]Assistant professor, Computer Science and Engineering, P. R. Pote COET, Amravati, Maharashtra, India

Abstract: ***Cancer has been known as one of the death dealing diseases Curing from this spiteful disease is mainly based upon early and exact diagnosis. But early and accurate detection of cancer is critical to the well being of patients. Cancer usually comes from the formation of a tumor. Computer aided diagnosis can be a very good helper in this area. Cancer computer aided diagnosis can be based on gene expression, microscopic images of extracted cell specimens, or radiogram of a whole organ, a survey around these different diagnosis ways has been presented in this paper to be a basis for revolutionary research. Also a new process for prediction, diagnosis and curing from cancer has been suggested by combining multiple old ways for prediction and diagnosis with the new nanotechnology way.***

Keywords: Cancer diagnosis; computer aided diagnosis; CAD; gene expression; nanotechnology

1. Introduction

Cell lifecycle as follows; cell grows and divides in a controlled way to produce more cells as they are needed to keep the body healthy. When cells become old or damaged, they die and are replaced with new cells. [6]

Sometimes this orderly process goes wrong. The genetic material (DNA or gene expression) of a cell can become damaged or changed, that is to say, that new cells are formed when they are not needed, and they group together to form a tumor. Cancer usually comes from the formation of a tumor. Tumors form in the body when cells are produced unnecessarily. The tumor can be benign, which means that it is non-cancerous, or it can be malignant, which means that it is cancerous. If cells break away from a malignant tumor, they will enter the bloodstream, and spread throughout the body, damaging other parts of the body.

Data mining is an essential step of knowledge discovery. Knowledge discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. In particulars, data mining may accomplish class description, association, classification, clustering, prediction and time series analysis. Data mining in contrast to traditional data analysis is discovery driven.

Accuracy of diagnosis decisions of malicious diseases represents an important factor for curing. Recently, a number of computer aided decision support systems have been created to address this emerging need. Two decisions need to be addressed, first whether the patient has cancer or not then the second is the cancer degree. It has been shown that the use of ANN including SVM in computer aided diagnosis in general and especially for cancer and cancer degree classification is very powerful and has a lot of advantages like [2]:

- Ease of optimization, resulting in cost-effective and flexible non-linear modeling of large data sets.
- Accuracy for predictive inference, with potential to support clinical decision making.
- These models can make knowledge dissemination easier by providing explanation for instance, using rule extraction or sensitivity analysis.

1.1 Prediction for Cancer

A. KDD Process

The idea of automatic knowledge discovery in large databases is first presented informally, by describing some practical needs of users of modern database systems. Several important concepts are then formally defined and the typical context and resources for KDD are discussed. KDD methods often make possible to use domain knowledge to guide and control the process and to help evaluate the patterns.

Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. The principal resource of KDD process is a database containing a large amount of data to be searched for possible patterns. KDD is never done over the entire database, but over a representative target data set, generated from the large database.

Phases of KDD process: Selection, Preprocessing, Transformation, DM, Interpretation/Evaluation.

1.2 Diagnosis of cancer

Decision in the different systems has been based on 3 different directions:

A. Gene Expression Microarray

One of these cancer diagnosis bases has been the gene expressions microarray. Cancer as mentioned before is a change of the cell DNA or gene expression of a given cell. Based on this, cells can be classified as normal or cancerous based on their gene expression.

B. Microscopic cell characteristics

The cancerous cell nuclei changes in characteristics for example it increases in dimensions, its chromatin density changes etc., because of the genetic changes. All these features changes can be used for classifying the cell in one of 2 classes either cancerous or non-cancerous.

C. Radiogram Regions characteristics

Cancerous tumors are shown in radiograms as regions in the organ with special characteristics and shapes different from the normal organ tissue. Identifying these regions and classifying them whether they are cancerous or benign is another way which is used in cancer diagnosis and it is generally the most common way. There are 3 computerized ways of cancer diagnosis: Gene Expression, Cell Characteristics, and radiogram regions characteristics respectively.

2. Literature Review

Data mining is a technique which provides automatic pattern recognition and tries to uncover patterns in data that are difficult to detect with traditional methods. Data mining techniques form a group of heterogeneous tools and techniques and are used for different purposes. These techniques and methods are based on statistical techniques, visualization, machine learning, etc. Data mining algorithms try to fit a model closest to the characteristics of data under consideration. These Models can be descriptive or predictive. Descriptive models are used to identify patterns in data, clustering, association rules, and visualization are some are the tasks of descriptive modelling.

2.1 Micro Array Technology

Micro array technology provides a tool for estimating expressions of thousands of genes simultaneously. Some steps as follows: Firstly, three-fourth of the samples of data is used to train the Classifier. secondly, the trained classifier is used to predict or test the one-fourth of the samples. The goals of classification are to identify the differentially expressed genes that may be used to predict class membership for new samples.

2.2 Noval Method

The noval method for mutational disease prediction using bioinformatics tools and datasets for diagnosis the malignant mutations with powerful Artificial Neural Network (Backpropagation Network) for classifying these malignant mutations are related to gene(s). This noval method didn't take in consideration just like adopted for dealing, analysing and treat the gene sequences for extracting useful information from the sequence, also exceeded the environment factors which play important roles in deciding and calculating some of genes features in order to view its functional parts and relations to diseases. This method proposing optimal and more accurate system for classification and dealing with specific disorder using back propagation with mean square rate 0.000000001

3. Methods

3.1 Back propagation Algorithm

The Back Propagation Algorithm is a multi-layered Neural Networks for learning rules [4], credited to Rumelhart and McClelland. It produces a prescription for adjusting the initially randomized set of synaptic weights such that to maximize the difference between the neural network's output of each input fact and the output with which the given input is known (or desired) to be associated. Back propagation is a supervised learning algorithm and is mainly used by Multi-Layer- perceptron to change the weights connected to the net's hidden neuron layer(s). The back propagation algorithm uses a computed output error to change the weight values in backward direction [12]. To get this net error, a forward propagation phase must have been done before. The neurons are being activated using the sigmoid activation function while propagating in forward direction.

3.2 Gene Expression Microarray

To use the gene expression microarray based diagnosis support system 3 steps are to be done [3]:

- Gene Selection; which is used to select the genes to be used when comparing the suspected cancerous cells genes this is done to reduce parameters dimensions and avoid overfitting. Without gene selection (If you use all the genes for comparison) decision overfitting may occur; which is creating diagnostic models that may not generalize well to new data.
- Choice of best suited classification algorithm, there are many neural networks and AI based algorithms which may be used for classification of cancerous cells like Nearest Neighbor NN, KNN, Multicategory Support Vector Machines (MC-SVMs), Decision Tree, Weighted Voting, to name a few. System training, to find the parameters which when passed to the newly developed system best suites the cancer diagnosis. Now system will be ready to be used for cancer diagnosis.

3.3 Cell Characteristics

The cell characteristics can show the DNA changes which occurred to the cell and so classify the cell as either cancerous or non-cancerous. Cell characteristics can be calculated through the integral optical density of the cell nuclei and then by comparing this to the reference integral optical density of the normal cells microscopic images, the cell can be classified as normal or cancerous through any classifier program. The above described methodology has been used for classifying the cancer and reference cells extracted from oral mucosa in.

Algorithm

To avoid overfitting, you first choose the relevant features in the optical microscopic image of the cell nuclei and then specify the normal values of the selected features.

When a new cell is to be checked whether it is cancerous or not, the Euclidean distance is measured between the selected features array and its counterpart in the normal cells then a

classifier like KNN (K Nearest Neighbor), fuzzy KNN, or SVM is used to finally classify the cell as cancerous or not cancerous as shown in figure 1 [9]. The classification accuracy result when this experiment was done is lying between95.5 – 95.7, and the best classifier with lowest computational cost was the SVM based on Radial Basis Function (RBF) using a set of 2 features only.

This technique is classifying any cell in one of 2 classes cancerous or normal. The best application of this technique is for the choice of the cells to be used for further analysis on the cancer degree of the patient, which is currently a manual process. Based on the cancer degree the corresponding treatment is chosen.

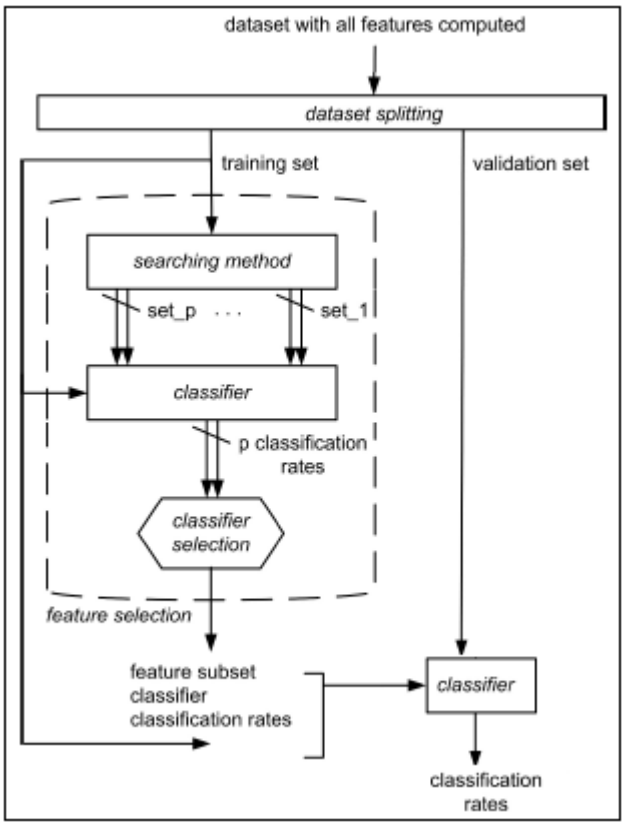

Figure 1: Algorithm for cell classification between cancerous and noncancerous. [9]

3.4 Radiogram Characteristics

If we can segment the tumors regions from normal organ tissue and then classify cancerous tumors from benign ones through ANN classifiers while deducing the cancer degree of the cancerous tumors through ANN classifier too. This can be a defined as a process for computer aided cancer diagnosis from radiogram characteristics. This has been used in lung cancer, prostate cancer, breast cancer ...etc.

Cancerous lung nodules and their degree have been detected using this process when applied to a chest X-ray using the fact that the nodules are relatively low-contrast white circular objects within the lung field and the cancerous nodules are different in intensity, uniformity, roughness, regularity, directionality, coarseness, smoothness and granulation.

For lung nodules detection too , and the paper suggests applying multi-scale 2D filter on continuous CT slices of lung suspected nodules and then apply an AND operator between the different results of the filter to finally lead for an efficient non-time consuming detection of lung nodules. While in, 3 SVM classifiers were combined using logical gates with Hippocrates-mst system to detect breast cancer from micro calcifications appearing in a mammogram.

They extracted from the FDG-PET image of the whole body the positive Gauss curvature and the negative mean curvature as the cancerous regions candidate then based on the Size, existence position, maximum intensity of the interesting peak surface, intensity difference between the interesting peak surface region and its circumference region they classify the region as really cancerous or not. This is a way to diagnose cancerous regions in the whole body rather than one organ.

4. Conclusion

This paper presents a primary research on predicting survival times for cancer patients based on clinical data, blood test results, and weight-loss assessment. We combined all data set and use feature selection to extract useful attributes and survey on the different methods which can be used for computer aided cancer diagnosis and suggested a new futuristic way for curing from this deadly disease by combining cancer regions and cells classifiers together with nanotechnology.

5. Future Work

Obvious feature work includes discovering a more accurate and appropriate model and determining whether our model can be better than what is practicing in the hospital. Also, it will be better if we can gather more available data for analysis, which will definitely, improve both the consistency of our data set and the usability of our results. For future works, we are looking to try different kind of feature selection, e.g., using Lasso algorithm. The drawback with lasso is it cannot reduce number of features to less than number of data points. However, this does not apply tour data set as we have more data points comparing to number of features. The other future work is applying neural network for the prediction similar to the approach discussed in [3]. However, setting the parameters in the neural network is always really hard. Finally, it worth to try to use semisupervised learning algorithm instead of removing the unlabeled data from the data set.

References

[1] Bellaachia, A., and Guven, E. Predicting Breast Cancer Survivability Using Data Mining Techniques.
[2] Bittern, R., Dolgobrodov, D., Marshall, R., Moore, P., Steele, R., And Cuschieri, A. Artificial Neural Networks In Cancer Management. E-Science All Hands Meeting 19 (2007),
[3] Djebbari, A., Liu, Z., Phan, S., And Famili, F. International Journal Of Computational Biology And Drug Design (Ijcbdd). 21st Annual Conference on Neural Information Processing Systems (2008).

[4] Figueiredo, M., And Jain, A. Unsupervised Learning Of finite Mixture Models. Ieee Transactions On Pattern Analysis And Machine Intelligence 24 (2002),
[5] Zupan, B., Demsar, J., Kattan, M., Beck, R., and Bratko, I. Joint European Conference On Artificial Intelligence In Medicine And Medical Decision Making. 21st Annual Conference On Neural Information Processing Systems 1620 (1999),
[6] "What Is Cancer? ", National Cancer Institute US National Institute of Health. http://www.cancer.gov/cancertopics/what-is-cancer, 2008-11-25.
[7] P. J. Lisboa, A. F. G. Taktak, "The use of artificial neural networks in decision support in cancer: A systematic review", http://pcwww.liv.ac.uk/~afgt/Lisboa&Taktak.pdf,
[8] data: A comprehensive evaluation to inform decision support system development", IOS Press, IMIA, Amsterdam, 2004.
[9] T. E. Schneider , "Automated classification of analysis and reference cells for cancer diagnostics in microscopic images of epithelial cells from the oral mucosa ", Institute of Imaging and Computer Vision, RWTH Aachen University, Sommerfeldstrasse 24, 52074 Aachen, Germany, Acta Politechnica Vol. 47 No. 4-5/2007, 2007.
[10] H. K. Nehemiah, A. Kannan, "An intelligent system for lung cancer diagnosis from chest radiographs", International Joural of Soft Computing 1(2), pp 133-136, @Medwell Online 2006, Chennai India, 2006.
[11] S. Takemura, X. Han, Y. Chen, K. Ito, I. Nishikwa, M. Ito, "Enhancement and detection of lung nodules with multiscale filters in CT images", IIHMSP, pp 717-720, 2008
[12] I. Andreadis, G. Spyrou, A. Antaraki, G. Zografos, D. Kouloheri, G. Giannakopoulou, K. Nikita & P. Ligomenides, "Combining SVM and Rule-Based classifiers for optimal classification in breast cancer diagnosis", proceedings of HERCMA 2007, Athens Greece, 2007.
[13] T. Tozaki, M. Senda, S. Sakamoto, K. Matsumoto,"Computer assisted diagnosis method of whole body cancer using FDG-PET images", ©IEEE, ICIP, 2003.
[14] N. Halas, "Is nanotechnology the key to curing cancer?",www.cnn.com, http://edition.cnn.com/2007/TECH/science/06/11/halas. vision/, March 11, 2008
[15] "Treating tumors Golden slingshot, the next generation of cancer treatments may be delivered by nanoparticles", http://www.nanostart.de/en/m_1071, Nov 6th, 2008, From the Economist print edition. 1265

Author Profile

Tejal V. Gudadhe received her B.E.(CSE) From P. R. Pote COET, Amravati Affiliated to Sant Gadge Baba Amravati University, Amravati, Maharashtra, India in 2013. Currently pursuing her M.E. in computer science and engineering at P. R. Pote COET, Amravati Affiliated to Sant Gadge Baba Amravati University, Amravati, Maharashtra, India.

Namrata D. Ghuse Working as Assistant Professor in the Department of Computer Science and Engineering, P. R. Pote COET, Amravati. She completed her M.E.(CSE) in 2014 from PRMIT Badnera and guided several PG and B.E. Projects.