

Effective Cancer Detection Using Soft Computing Technique

¹Bashetha. A, ²Dr. G.Umarani Srikanth
^{1,2}, (CSE, S.A Engineering college, India)

Abstract: Cancer research is rudimentary research which is done to identify causes and develop strategies for prevention, diagnosis, treatment and cure. An optimized solution for the better treatment of cancer and toxicity minimization on the cancer patient is performed by identifying the exact type of tumor. A clear cancer classification analysis system is required to get a clear picture on the insight of a problem. A systematic approach to analyze global gene expression is followed for identifying exact problem area. Molecular diagnostics provide a promising option of systematic human cancer classification. But these types of tests are not mostly applied because characteristics molecular markers have yet to be identified for most solid tumors. Recently, DNA micro-array based tumor gene expression profiles have been used for cancer diagnosis. In the proposed system, gene expressions are taken from multiple sources and an ontological store is created. Ant colony optimization technique is used to analyze the cluster of data with attribute match association rule for detecting cancer using the acquired knowledge.

Keywords: Gene expression, cancer cells, ontological store;

I. Introduction

Data mining is the widely used technique to obtain the knowledge data from the existing history of data. The large collection of data sets is stored in the database and using the concept of mining we can obtain the knowledge data. The large collection of data bases is known as warehouse. Data mining technique has its applications in the field of computer science, statistics, and artificial intelligence and in many other fields.

Data mining technique have several stages such as pre-processing, mining and validation stages. The data pre-processing is the first stage in which the data sets collected are arranged in the proper structure or format suitable for the mining process. The unwanted, unambiguous and redundant data are removed from the repository and proper structured data base is created. The removal of unwanted data is known as data cleaning.

The second stage is the mining stage in which the actual work is done. Various mining techniques are used such as clustering, pattern matching, regression, classification, etc. to obtain the knowledge data which are previously unknown data. The third and final stage is validation stage in which the results obtained from mining process is validated. The result produced by the mining process is not always prone to be correct therefore the results obtained should be validated.

In the study of human genetics, sequence mining helps address the important goal of understanding the mapping relationship between the inter-individual variations in human DNA sequence and the variability in disease prediction. In simple words, it aims to find out how the changes in an individual's DNA sequence affects the risks of developing common diseases such as cancer, which is of great importance to improve the methods of detecting, preventing, and handling these diseases.

II. System Analysis

2.1 Overview of cancer

The abnormal growth of cells in the body leads to development of tumors in any part of the body. Not all tumors lead to cancer, there are two major classifications of tumors, one is malignant tumor and the other is benign tumors. The benign tumors are non-cancerous or harmless tumors whereas the malignant tumor are harmful and leads to cancer. There are many reasons to cause of cancer based upon the genome we can classify it into two main categories such as genetic and epigenetic.

The cancer is one type of a genetic disease which passes throughout the generations through heredity. Cancer is a disease of tissue growth regulation failure. A normal cell to transform into a cancer cell, the genes which control cell growth and differentiation must be altered. Epigenetic alterations refer to functionally relevant modifications to the genome that do not involve a change in the nucleotide sequence.

The great majority of cancers, some 90–95% of cases, are due to environmental factors. The remaining 5–10% is due to inherited genetics. Environmental factor means any cause that is not inherited genetically, such as lifestyle, economic factor and behavioral factors, and not merely due to pollution. Common environmental factors that contribute to cancer death include tobacco (25–30%), diet and obesity (30–35%), infections (15–20%), radiation (both ionizing and non-ionizing, up to 10%), stress, lack of physical activity, and environmental pollutants.

It is difficult to find or detect cancer in the earlier stage; the symptoms of cancer depend on the type and location of the cancer. There are no symptoms in the earlier stages of cancer; they appear only when the cancer grows rigorously. The symptoms of cancer are most commonly occurring diseases such as cold, fever etc. Thus it is common that people diagnosed for cancer are given treatments for other commonly occurring diseases which are actually the symptoms caused by the cancer. Therefore it becomes late to identify cancer.

2.2 Early cancer detection

The early detection of cancer is very much important to give early treatments so that the probability of curing the disease increases. The pain and toxicity minimization can be given to the patient if the disease is found in the earlier stages. Screening is the method used to detect the disease in the earlier stages before the symptoms are shown. Not all people have to undergo screening test, the people who have high risk factor can undergo screening test. Risk factors are different for different cancer types. The two main types of risk factors are genetic and non-genetic which are discussed earlier.

Predictive genetic testing is used to determine if an individual has a genetic that may predispose him/her to developing cancer. This test will reveal a mutation caused in the gene, but cannot guarantee that a person will develop cancer. It can identify the cause for cancer in a ménage and, as a result, could help to identify ménage who are at a high risk for developing cancer. Therefore they can take preventive measures and can undergo more frequent screening procedures to detect cancers at early stages when they are most treatable. Also this test could identify that a person is not at an increased risk for developing cancer and, as a result, removes uncertainty or anxiety in that person.

2.3 Cancer and cell types

Body is made up of millions of cells. These cells are organized together to form tissues and organs. Each type of cells has their own unique functionalities. Some of the cells are more common and found in all parts of the body. But some are very peculiar and found in only one part of the body. The examples of this type are brain cells.

Cancer can also be classified based upon the type of cell that it has started in. The main types of cells are epithelial cell, cells of the blood and lymphatic system and connective tissue cells. Cancer that developed in different type of cells has different names. Cancer that develops in epithelial cell is called carcinomas.

The different types of epithelial cells are squamous cells, adeno cells, urothelial cells and basal cells. Carcinomas cancer can develop in any one of these cells. The most occurring type of cancer is carcinomas. Nearly 80-90% cancer caused is carcinomas. Fig. 1. depicts the percentage of various cancer types. Cancer that develops in blood or bone marrow is known as leukaemias and one that develops in lymphatic system is known as lymphomas.

Cancer that starts in connective tissue cells are known as sarcomas. Cancer also develops in other type of cells but they are very rare. The common cancer that develops in other cell is brain cancer. The cancer can also be classified based upon the site where it first developed. The most common type of cancers is skin, lung, and breast, prostate.

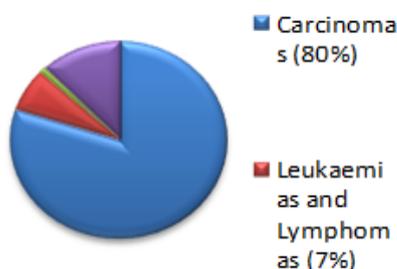


Fig. 1. A pie chart depicting percentage of various types of cancer affected based upon the cell that it gets started.

III. Proposed System

Predicting cancer by analyzing the gene expression is the proposed concept. The data mining techniques are used to predict cancer by comparing the gene expressions samples taken from the patient with the expert's documental data. The cancer gene expression patterns are designed and these patterns are compared with the sample gene expressions to find out the affected gene expression patterns. Then the clustering technique is used to form the clusters of related gene patterns. Then by analyzing this cluster the final prediction of cancer is done.

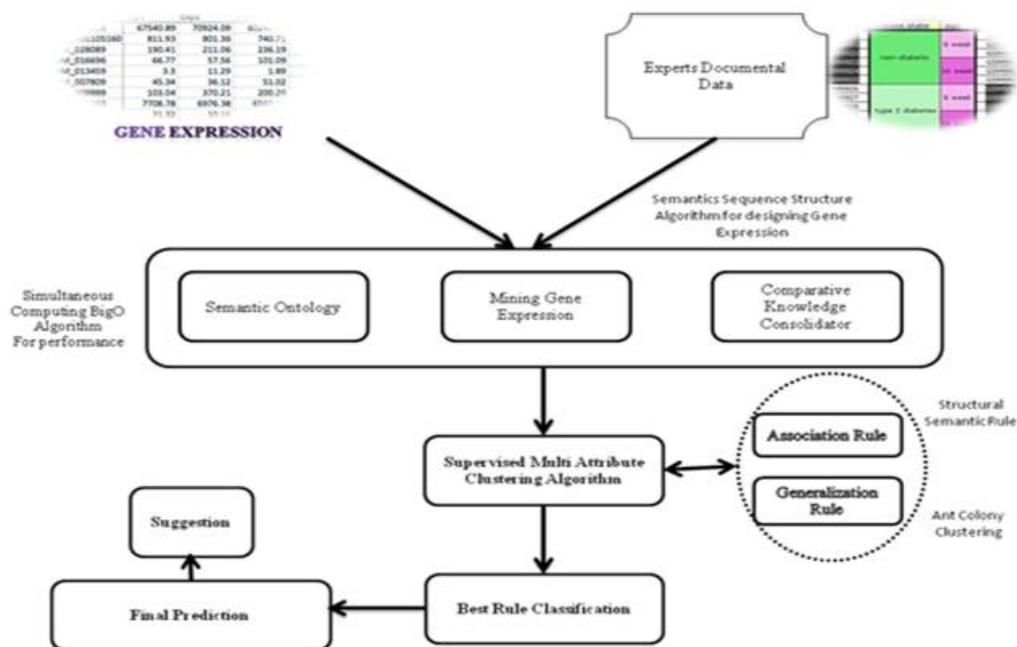


Fig. 2. System Architecture diagram depicting the concept of proposed system.

3.1 System architecture

The system architecture of the proposed system is given in Fig. 2. Architecture diagram clearly explains the concept of proposed system. The expert’s documental data is stored in the data base, the documental data is nothing but the history of gene expressions obtained from the cancer affected patient previously. The sample gene expression from the patient for whom the cancer diagnosis has to be done is given as input. A semantic ontology is created using the above two data sets.

Next the data mining technique is used to compare the two data sets and the characteristic of each gene is extracted or is identified. The knowledge extracted about the gene expression data set is collectively known as knowledge consolidator. This knowledge consolidator mainly focuses on the performance of each individual gene expression data’s.

The next step is ontological mapping. The ontological mapping is the mapping of two different gene expression data’s to identify the difference in gene characteristics. This ontological mapping plays vital role in analysis for providing solution. The next step is to form the clusters of genes with similar characteristics. The multi attribute clustering algorithm is used to form clusters based upon various association and generalization rules. Finally, the ant colony optimization algorithm is used to find out the clusters which have the cancer gene expression characteristics and thus the final prediction of cancer is done.

IV. Algorithms Used

The two important algorithms used in the proposed system are supervised multi attribute clustering algorithm and ant colony optimization algorithm.

4.1 Clustering techniques

Clustering is a technique in which the objects which have similar characteristics are grouped into cluster. It is a technique in which the logically related objects are physically stored in the data base. There are many clustering methods available each produces the different data clusters. The choosing of clustering method is based upon the desired outcome required. Based upon the structure of the cluster the clustering can be divided into two type hierarchical and non-hierarchical methods. The non-hierarchical methods divide a set of K objects into N clusters with or without overlap. This may sometimes come under partition method in which clusters are mutually exclusive and in which overlapping is allowed.

The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster is formed. The hierarchical methods can be classified into agglomerative or divisive methods. In this agglomerative method, the hierarchy is build up in a series of K-1 agglomerations or fusion of pairs of objects, beginning with the data set which is un-clustered. Some of the divisive methods begin with all objects in a single cluster and at each of K-1 steps divide some clusters into two smaller clusters, until each object present in its own cluster.

The partitioning methods generally result in a set of N clusters where each object belongs to one cluster. Each cluster is identified by a centroid or a cluster representative, which represents all the objects contained in a cluster. A simple method of partitioning method is the single pass method. In this method the first object is made as the centroid of the first cluster, for the next object calculate the similarity S with the existing cluster centroid with some similarity function.

If the calculated S is greater than the specified threshold value then it adds the object to the corresponding cluster and the centroid of the cluster is re-determined. As the name implies this method requires only one pass through the data set. The time requirements are typically of order O (NlogN) for order O (logN) clusters. Therefore it is a very efficient clustering method for a serial processor. A drawback is that the clusters formed are not independent of the order in which the data set are processed, the first cluster formed is usually larger than those created later in the clustering run.

The hierarchical agglomerative clustering methods are most commonly used. The construction of a hierarchical agglomerative classification can be achieved by the following general algorithm. In this method the clusters are formed using the following method. The two nearest objects are compared if they have similar characteristics they are grouped into a cluster. Find and merge the next two closest points, where the point is either an individual object or a cluster of objects. If more than one cluster is remained then again repeat the same steps until the common cluster is formed.

4.2 Ant colony optimization algorithm

The ant colony optimization algorithm (ACO) is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. The ant colony algorithm is an algorithm for finding optimal paths that is based on the behaviour of ants searching for food. At first, the ants walk randomly in search of food. When the ant finds food source, it walks back to the colony leaving "markers" (pheromones) that shows which path has food. When other ants come near to the markers, they follow the path with a certain probability. Then they populate the path with their own markers as they bring the food back. When more ants find the path, it gets stronger until there are a couple streams of ants travelling to various food sources near the colony. It iteratively constructs a solution for the problem. The solutions found in the intermediate stages are referred as solution states.

```
procedure ACO_MetaHeuristic
While (not_termination)
Generate Solutions ()
Daemon Actions ()
Pheromone Update ()
End while
End procedure
```

Fig. 3. Ant Colony Optimization Algorithm outline

V. Related Work

5.1 DNA methylation

The modification of a strand of DNA after it is replicated, in which a methyl (CH₃) group is added to any cytosine molecule that stands directly before a guanine molecule in the same chain. The methylation of cytosines in particular regions of a gene can cause that gene to suppress, therefore it is one of the methods used to regulate the expression of genes. The chemical reaction places a methyl group (a combination of one carbon atom and three hydrogen atoms) at a particular spot on DNA during organism development. The effect of this process is probably to "turn off" various genes during the process of cellular distinction, causing the cell to develop into a specific type.

5.2 Microarray analysis

Microarray analysis techniques are used in interpreting the data generated from experiments on DNA, RNA, and protein microarrays. This allows researchers to carry out research on the expression state of a large number of genes. In many cases, an organism's the entire genome of an organism can be researched in a single experiment. This experiment can generate large quantity of data, allowing researchers to estimate the overall state of a cell or organism. The large amount of data produced is difficult to analyze, when there is absence of good gene annotation.

A microarray is a multiplex lab-on-a-chip. It is a 2D array on a solid substrate (generally a glass slide or silicon thin-film cell) that examines large amounts of biological material using high-throughput screening, multiplexed and parallel processing and detection methods. All of the cells in the human body contain same kind of genetic material; the same genes are not active in the entire cell. Examining which genes are active and

which are inactive in different cell types helps scientists to understand both how these cells function normally and how they are affected when various genes do not perform properly. On development of DNA microarray technology, scientists can now inspect how active thousands of genes are at any given time.

5.3 Gene therapy

Gene therapy is the use of nucleic acid polymers as a drug to treat disease by therapeutic delivery into the cells of the patient, in which they are either expressed as proteins or possibly even correct genetic mutations. The general gene therapy involves DNA that encodes a mutated gene with the functional therapeutic gene. In gene therapy, the nucleic acid molecule is wrapped within a "vector". This vector is used to get the molecule inside cells within the body. In gene therapy, DNA must be regulated to the cells of the patient that need repair, enter the cell, and demonstrate protein in a medically useful way. In general the DNA is integrated into an engineered virus that serves as a vector to the DNA through the cells of the blood stream, and integrated into a chromosome. Naked DNA approaches have also been found, which are used in the context of vaccine development.

VI. Conclusion And Future Work

Predicting Cancer by analysing gene and converting the gene expression is the proposed concept of the project, which leads to identifying and analysing the cancer result set. Controlling gene activity from gene to functional protein & phenotype has also been analysed in order to identify the cancer cells. In the proposed methodology, the expert's documental DNA methylation (Gene expression segments) is a kind of binding site for proteins which make DNA inaccessible to be in alive state. Semantic ontology based mining gene expression analysis tends to compare the gene expression values by using the comparative Knowledge Consolidator. Ant colony optimization technique has been used to find the Best Rule Classification in the gene expression to find the final prediction of cancer disease.

Several machine learning and data mining techniques are presently applied for identifying cancer using gene expression data. The dissimilarity in efficiency exist, none of the entrenched approaches is uniformly superior to others. The standard of algorithm is important, but it is not itself an assurance of the quality of a specific data analysis.

References

- [1]. G.-M. Elizabeth and P. Giovanni, "Clustering and classification for gene expression data analysis". Johns Hopkins Univ.,Dept. Of Biostatist Working Paper 70.
- [2]. E. Shay,(2003,Jan.). "Microarray cluster analysis and applications". Available: <http://www.science.co.il/enuka/Essays/Microarray-Review.pdf>.
- [3]. D. Jiang, C. Tang, and A.Zhang, "Cluster analysis for gene expression data: A Survey", IEEE Trans. Knowl. Data Eng., vol.16, no.11, pp. 1370-1386, Nov.2004.
- [4]. D.A. Roff and R. Preziosi, "The estimation of the genetic correlation: The use of the jack knife," Heredity, vol. 73, pp.544-548, 1994.
- [5]. T. Scharl and F. Leisch, "Jack knife distances for clustering time course gene expression data," in proc. ASA biometrics, p.8, 2006.
- [6]. N. Pasquier, C. Pasquier, L. Brisson, and M. Collard, "Mining gene expression data using domain knowledge," Int. J. Softw. Informat, vol. 2, pp. 215-231, 2008.
- [7]. B. Collard, "An ontology driven data mining process" Inst. TELECOM, TELECOM Betagne, 2008.
- [8]. Y. Su, T. M. Murali, V. Pavlovic, M. Schaffer, and S. Kasif, "RankGene: Identification of diagnostic genes based on expression data," Bioinformatics, vol. 19, no. 12, pp. 1578-1579, 2003.
- [9]. S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumours using gene expression data," J. Amer. Statist. Assoc., vol. 97, no. 457, pp. 77-87, March. 2002.
- [10]. K. Raza and A.Mishra, "A novel anticlustering filtering algorithm for the prediction of genes as a drug target," Amer. J. Biomed. Eng., vol.2, no. 5, pp. 206-211, 2012.