

Effective Cancer Detection Using Data Mining and Ant Colony Optimization

Dr. G. Umarani Srikanth

Professor and Head of the Department

Department of CSE

S. A. Engineering College

gmurani@saec.ac.in

A. BASHETHA

Department of CSE

S. A. Engineering College

Abstract

Cancer is a disease caused by an uncontrolled division of abnormal cells in a part of the body. Cancer research is one of the major research fields in the current trend. Identifying the cancer in the patient using the gene expression profile is the proposed concept. The gene expressions are obtained by the method of 2-d micro array technology. The proposed system uses two important algorithms such as supervised multi attribute clustering algorithm and ant colony optimization algorithm. It is imperative to identify the cancer in early stages so that the measures can be taken to cure the disease. The increase in the death rate of cancer is due to the identification of cancer in the later stages. Using the proposed concept the disease can be identified in the earlier stage.

Keywords: Gene expression, cancer, Ant Colony Optimization(ACO), decision trees.

Introduction

There are several data mining techniques that have been developed and used in many projects. The widely used techniques are clustering, classification, sequential patterns, prediction and decision tree. One of the best known data mining techniques is the association technique. In association, a pattern is discovered based on a relationship between items in the same transaction. Hence the association technique is known as relation technique.

The next data mining technique is the classification technique which is based on machine learning. This technique is used to classify each item in a set of data into one

of predefined set of classes or groups. This method uses mathematical techniques such as linear programming, neural networks, decision trees and statistics.

Clustering is a data mining technique which is used to form meaningful or useful cluster of objects which have similar characteristics using automatic technique [1], [3]. This technique defines the classes and puts objects in each class, whereas in the classification techniques, objects are assigned into predefined classes.

The prediction is one of the data mining techniques that identify relationship between independent variables and relationship between dependent and independent variables. Sequential patterns analysis is one of data mining techniques that seek to discover or identify similar patterns, recurring events or trends in transaction data over a business period.

Decision tree is one of the most used data mining techniques as its model is easy to understand for users. The root of the decision tree is a simple question or condition that has multiple outcomes. The outcomes obtained leads to a set of questions or conditions that help us determine the data so that the final decision is made based on it.

System Analysis

Overview of Cancer

Cancer is a class of diseases characterized by uncontrollable growth of cell [11]. There are more than hundred different types of cancer which is classified by the type of cell that it is initially affected. Cancer harms the body when damaged cells divide uncontrollably to form lumps or masses of tissue called tumors. Tumors can grow and intervene with the digestive, circulatory systems and nervous, and they can produce hormones that changes the body function.

Cancer is ultimately the result of cells that uncontrollably grow and do not die. Normal cells in the body follow a procedural path of growth, division, and death. Programmed cell death is called apoptosis, and when this process stops working, cancer begins to form. Unlike regular cells, cancer cells do not undergo programmatic death and instead continue to grow and divide. This leads to a mass of abnormal cells that grows out of control.

Cells undergo uncontrolled growth if there are damages or mutations to DNA, and causes damages to the genes involved in cell division. The four key types of genes which are responsible for the cell division process are oncogenes which tells the cells when to divide, tumor suppressor genes inform cells when not to divide, suicide genes control apoptosis instructs the cell to kill itself if something goes wrong, and DNA-repair genes informs a cell to repair damaged DNA.

Cancer occurs when a cell's gene mutations make the cell unable to correct DNA damage and unable to commit suicide. Cancer is a result of mutations that hinders oncogene and tumor suppressor gene function which leads to uncontrollable growth of cell.

Early Cancer Detection

If the cancer is diagnosed earlier it can be treated and there is a better chance of it being cured. Few types of cancer such as those of the skin, mouth, breast, prostate, testicles, and rectum can be detected by regular self-exam or other screening measures before the symptoms become serious. In more cases the cancer is detected and diagnosed after a tumor is formed. Very rarely cancer is diagnosed incidentally as a result of evaluating or treating other medical conditions.

Cancer diagnosis begins with a thorough physical exam. Laboratory studies of blood, urine, and stool can detect uncommonness that may indicate cancer. When a tumor is identified, imaging tests such as X-rays, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), ultrasound, and fiber-optic endoscopy examinations help doctors determine the cancer's exact location and its size. Confirmation of the diagnosis of most cancers is done by biopsy [9]. A biopsy needs to be performed in which a tissue sample is removed from the suspected tumor and studied under a microscope to check for cancer cells.

Classification of different tumors in cancer detection and drug identification is very important task. Earlier, the cancer classification was based on clinical information that has a limited ability for detection and debugging. Now, micro array technology has enabled monitoring the description of thousands of genes simultaneously [2]. The gene expressions are obtained using micro array technology and the cancer can be predicted at earlier stage.

Cancer Types

Cancers are often described by the body part that they originated in. Some body parts contain multiple types of tissue, so for greater accuracy, cancers are classified by the type of cell that the tumor cells got originated. The types of cancer are carcinoma, sarcoma, lymphoma and leukemia, germ cell tumor and blastoma. Carcinoma is the cancers derived from epithelial cells. This group includes most common cancers, especially in older adults. Nearly all cancers developing in the breast, lung, prostate, colon and pancreas are carcinomas.

Sarcoma is the cancers arising from connective tissue such as bone, cartilage, fat, nerve and each of which develop from cells originating in mesenchyme cells outside the bone marrow. Lymphoma and leukemia are the two classes of cancer arise from cells that form blood. Leukemia is the most common type of cancer in children accounting for about 30%. Majority of adults develop lymphoma and leukemia. Germ cell tumor is the cancers derived from pluripotent cells, most often present in the testicle or the ovary (seminoma and dysgerminoma, respectively). Blastoma is the cancer derived from immature "precursor" cells or embryonic tissue. This type of cancer is more common in children than in older adults.

In some cases types of cancer are named for the size and shape of the cells under a microscope, such as giant cell carcinoma, spindle cell carcinoma, and small-cell carcinoma.

Related Work

DNA Methylation

The modification of a strand of DNA after it is replicated, in which an ethyl (CH_3) group is added to any cytosine molecule that stands directly before a guanine molecule in the same chain [2]. The methylation of cytosine in particular regions of a gene can cause suppression of gene, so it is one of the techniques used for the regulation of gene expression. The chemical reaction places an ethyl 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.

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 genes in large number-in numerous cases and the entire genome of an organism can be researched in a single experiment. This experiment can generate large amount of data, with which the researchers can be able to estimate the overall state of a cell or organism. The large quantity 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 the methods for detection [2], [11]. All the cells that are present 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 [2].

Gene Therapy

Gene therapy is the use of nucleic acid polymers as a drug to treat disease by therapeutic delivery into the patient cells, in which they are either revealed as proteins or possibly even correct genetic mutations. The generalized gene therapy involves DNA that conceals 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 in the body. In gene therapy, DNA must be modulated to the cells of the patient that need repair, enter the cell, and demonstrates protein in a useful way [11]. 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 are also found, which are used in the process of vaccine development.

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 gene expression patterns for cancer 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 examining this cluster the final prediction of cancer is done.

System Architecture

The system architecture of the proposed system is given in Fig. 1. Architecture diagram clearly explains the concept of proposed system. The documental data obtained from expert 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. The knowledge extracted from the gene expression data set is collectively known as knowledge consolidator. This mainly focuses on the performance of each individual gene expression data. The next step is ontological mapping. The ontological mapping is the mapping of two different gene expression data to identify the difference in gene characteristics. This ontological mapping plays important role in analysis for obtaining solution. The next step is to generate 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, ACO algorithm is used to find out the clusters which have the cancer gene expression characteristics and thus the final prediction of cancer is done.

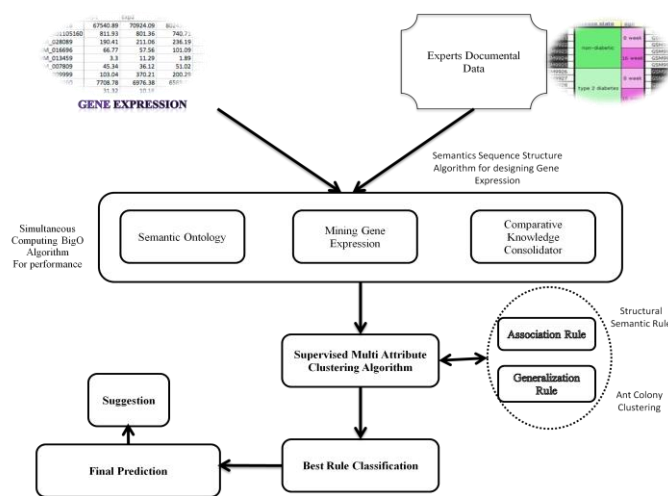


Figure 1: System Architecture diagram depicting the concept of proposed system

Algorithms Used

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

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 [1], [3]. The choosing of clustering method is based upon the desired outcome needed. Depending upon the structure of the cluster the clustering can be divided into two type hierarchical and non-hierarchical methods. The method which divides a set of K objects into N clusters with or without overlap is called non-hierarchical method. This method 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 and divisive method. In the agglomerative method, the hierarchy is build up in a series of K-1 agglomerations or fusion of pairs of objects, starting with the data set which is un-clustered[11]. 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 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. An easy method of partitioning is the single pass method. In this method the first object is made as the centroid of the first cluster, for the remaining objects 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 determined again. This method requires only one pass through the data set. The time required are typically of order $O(N \log N)$ for order $O(\log N)$ clusters. Therefore it is a very efficient clustering method for a serial processor[10]. 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 most commonly used clustering method is hierarchical agglomerative clustering method. 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 neighboring objects are compared if they have similar characteristics they are grouped into a cluster. Then 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 remains then repeat the same steps until the common cluster is formed.

Ant Colony Optimization Algorithm

The ACO is a probabilistic technique for solving computational problems through graphs which can be reduced to finding good paths. The ant colony algorithm is used for finding optimal paths that is based on the behavior of ants searching for food. At first, the ants walk randomly in search of food[11]. Once it finds food source, it walks back to the colony leaving "markers" (pheromones) that shows which path has food. When remaining ants come near to the markers, they follow the path with a certain probability. Then they colonize 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. By this process it iteratively constructs a solution for the problem [11]. 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
    
```

Figure 2: Ant Colony Algorithm outline

Results and Discussion

The results obtained by implementing the proposed concept is been discussed in this section.

ID	Gene title	Gene symbol	Gene ID	UniGene title	UniGene symbol	UniGene ID	Nucleotide Title
1007_s_at	discoidin domain ...	DDR1	780				Human receptor t...
1053_at	replication factor ...	RFC2	5982				Human replicatio...
117_at	heat shock 70kD...	HSPA6	3310				Human heat-sho...
121_at	paired box 8	PAX8	7849				H.sapiens Pax8 ...
1255_g_at	guanylate cyclas...	GUCA1A	2978				Homo sapiens gu...
1294_at	ubiquitin-like modi...	UBA7	7318				Homo sapiens ub...
1316_at	thyroid hormone r...	THRA	7067				Homo sapiens m...
1320_at	protein tyrosine p...	PTPN21	11099				H.sapiens mRNA...
1405_i_at	chemokine (C-C ...	CCL5	6352				Human T cell-spe...
1431_at	cytochrome P450...	CYP2E1	1571				Human cytochro...

Figure 3: Experts documental data

	ID_REF	IDENTIFIER	GSM627133	GSM627216	GSM627134	GSM627151	GSM627115	GSM627087
▶	231224_x_at	PRKAG2	35.8955	30.7351	32.9837	32.7659	37.144	38.6663
	240882_at	R85522	11.6371	10.5217	10.8537	11.8015	11.9516	13.0749
	1561849_at	PKD1L2	8.45343	8.46326	7.39276	7.33646	7.37014	7.52145
	1565746_at	LOC100132815	10.8116	11.1259	9.72156	8.65285	12.1602	10.7603
	1560853_x_at	ZNF826P	16.0361	16.6499	14.5228	15.519	17.4264	16.0635
	230660_at	SERTAD4	12.4425	14.2759	12.7199	13.7253	12.7716	15.6567
	229708_at	TOR4A	13.3051	13.8157	12.2562	19.5204	14.6117	15.015
	244781_x_at	R37682	8.12955	10.2744	8.59731	10.2062	9.19906	10.2366
	1554187_at	LOC554206	16.9199	15.6306	18.8628	17.3767	14.8512	15.015
	230021_at	TICRR	25.4832	32.9972	27.267	26.6506	25.6564	21.8105
	238226_at	TMEM255B	29.5791	26.6391	27.1048	25.821	27.9684	28.0039

Figure 4: Sample Data

Fig.3 represents the expert's documental data. The documental data obtained from expert is stored in the data base, the documental data is nothing but the history of gene expressions obtained from the cancer affected patient previously. Fig.4 represents the sample data. The sample gene expression from the patient for whom the cancer diagnosis has to be done is given as input. The frequency of the dataset is found, which is shown in table 1. In the dataset 1, 2, 3 show the various types of lung cancer symptoms. 0 indicates the cancer symptom is not found. ? means the data is unknown, which signifies that the symptom can be presented or not and it is not clear.

Table 1: Frequency of Data Set

Type	Frequency	No of samples taken
1	590	697
0	107	-
3	1734	2107
?	5	-
2	1500	1631

The analyses is that for 107 cases there is no detection and for 5 cases the result is ambiguous means the possibility and non-possibility of symptoms are equal. For applying ant colony optimization, set the types as the ant and the support value as the pheromone.

Support of 1: $590/697=84.64\%$

Support of 2: $1500/1631=91.05\%$

Support of 3: $1734/2107=82.30\%$

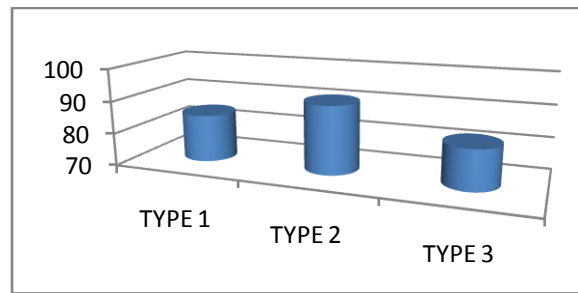


Figure 5: Individual Type Accuracy

Detection Accuracy

1. Proposed ACO 87.7
2. DPSO 68.33
3. PSO 65.44
4. PART 48.14
5. SMO 48.14
6. Naive Bayes 56.67
7. KNN 45
8. Classification Tree 46.67

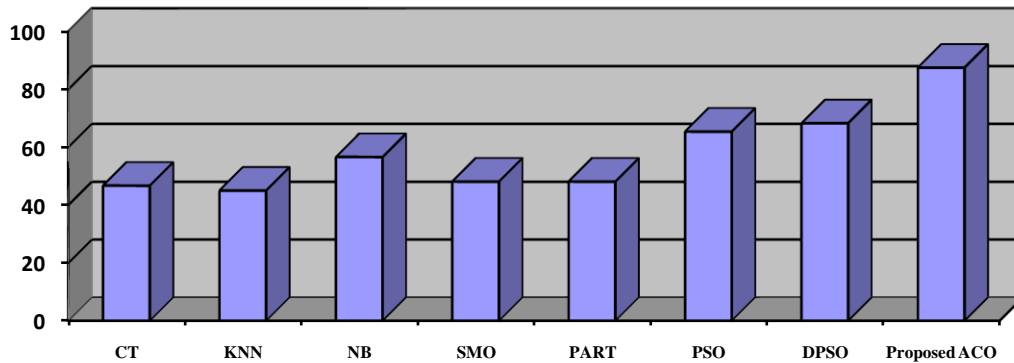


Figure 6: Detection Accuracy Comparison

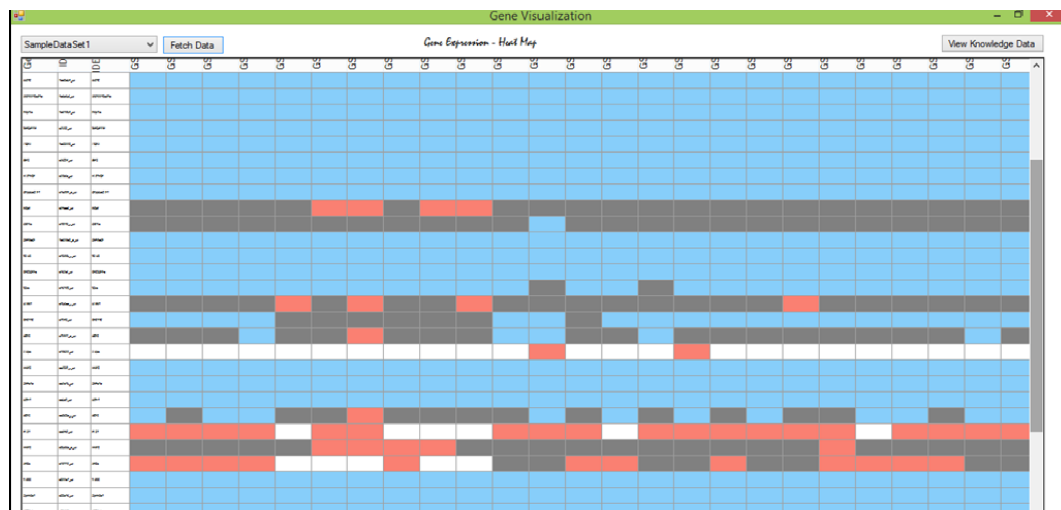


Figure 7: Heat Map

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. Fractal maps and tree maps both often use a similar system of color-coding to represent the values taken by a variable in a hierarchy. Heat map is generated from the extracted gene expression. Each individual gene is examined and the characteristic of individual gene is obtained.

The Fig.7 represents the heat map generated from the sample data taken. The heat map has four different colors each color indicating different characteristics. The blue color indicates normal genes, red color indicates affected genes, and grey color indicates the genes that are likely to be affected and the white color indicate that genes are in safe condition.

Conclusion and Future Work

Predicting Cancer by analyzing gene and converting the gene expression is the proposed concept of the project, which leads to identifying and analyzing the cancer result set. Controlling gene activity from gene to functional protein & phenotype has also been analyzed 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 present state. Mining gene expression analysis is done by Semantic ontology, which 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 variation in efficiency exist, none of the established approaches is uniformly surpassing to others. The standard of algorithm is necessary, but it is not itself a guarantee of the quality of a specific data analysis.

References

- [1] Elizabeth, G. M., and Giovani, P., 2004, "Clustering and classification for gene expression data analysis," Johns Hopkins Univ., Dept. Of Biostatist Working Paper 70.
- [2] Shay, E., 2003, "Microarray cluster analysis and applications," Available: <http://www.science.co.il/enuka/Essays/Microarray-Review.pdf>.
- [3] Jiang, D., Tang, C. and Zhang, A., 2004, "Cluster analysis for gene expression data: A Survey," IEEE Trans. Knowl. Data Eng., vol.16, no.11, pp. 1370-1386.
- [4] Roff, D. A. and Preziosi, R., 1994, "The estimation of the genetic correlation: The use of the jack knife," Heredity, vol. 73, pp.544-548.
- [5] Scharl, T. and Leisch, F., 2006, "Jack knife distances for clustering time course gene expression data," in proc. ASA biometrics, p.8.
- [6] Pasquier, N., Pasquier, C., Brisson, L. and Collard, M., 2008, "Mining gene expression data using domain knowledge," Int. J. Softw. Informat, vol. 2, pp. 215-231.
- [7] Collard, B., 2008, "An ontology driven data mining process" Inst. TELECOM, TELECOM Betagne.
- [8] Su, Y., Murali, T. M., Pavlovic, V., Schaffer, M. and Kasif, S., 2003, "RankGene: Identification of diagnostic genes based on expression data," Bioinformatics, vol. 19, no. 12, pp. 1578-1579.
- [9] Dudoit, S., Fridlyand, J. and Speed, T. P., 2002, "Comparison of discrimination methods for the classification of tumours using gene expression data," J. Amer. Statist. Assoc., vol. 97, no. 457, pp. 77-87.
- [10] Raza, K. and Mishra, A., 2012, "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.
- [11] Sharuya Jauhari and Rizvi, S.A.M., 2014, "Mining Gene Expression Data Focusing Cancer Therapeutics: A Digest," IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.11, no.3, pp. 533-547.
- [12] Christian Blum, 2005, "Ant Colony Optimization: Introduction and recent trends," Elsevier, Physics of Life Reviews 2, pp. 353-373.