Cancer Detection using Frequency Pattern Ant Colony Optimization

[1]Ritu Shukla, [2]Prof.Dilip Motwani
[1]PG scholar, [2]Assistant Professor
[1]Department of Computer,
[1]SLRTCE, Mira Road

***Abstract* - Over the past few decades, to computerized diagnostic tools, intended to aid expert in making sense out of the welter of data. Due to improvements in biometric instrumentation and automation, it is easy to collect a lot of experimental data in molecular biology. It is extremely important for Analysis of such data as it leads to knowledge discovery that can be validated by experiments. Several machine learning and data mining techniques are presently applied for identifying cancer using gene expression data. This paper proposes that consecutive application of two algorithms will result in finding the gene causing cancer. Using association rule mining and swarm intelligence will result in finding an accurate gene which is responsible for causing cancer. In this paper we present a system for diagnosis of cancer using FP (frequent pattern mining) growth algorithm. Ant colony optimization to predict the possibility of cancer .Ant colony algorithm is employed as evolutionary algorithm to optimize the obtained set of association rules. We are using FP algorithm to conclude whether the tumour is malignant or benign tumour. Using these two algorithms the usage of memory will also be less**.

***Index Terms* - Microarray data, frequency pattern, Apriori, ant colony optimization**

I. INTRODUCTION

CANCER is a major cause of all the natural mortalities all over the world. Nearly 13 percent of deaths caused are due to cancer [1]. It is a disease getting constantly challenged by many researchers. Some advancement has been reported for its clinical prevention and cure and there has been a noticeable decline in the lives' lost [2], but they are not quite adequate [3]. The lack of affordable treatment and early detection is the crux of this hostile situation. It is becoming hard for the perennial biomedical scientists and researchers to deal with cancer. The growth is due to the division and multiplication of cells which takes place in body. When the appropriate division levels have been achieved, the process is deactivated. In an abnormal case cells continue to replicate and form lumps in the body. Cancer is an abnormal and uncontrollable growth of cells in the body that turn malignant.

Cancer classification is an important problem for both clinical treatment and biomedical research. Accurate diagnosis of cancer types can enhance efficacy and reduce toxicity of medical treatment for cancer patients. In the past, cancer classification has been relying on subjective judgment from professional pathologists. Currently, microarray experiments can be employed to screen gene expression levels from normal and cancer tissue samples. The comparisons of microarray results between normal and cancer cells can provide the important information of cancer diagnosis and treatment. However, Microarray experiments provide enormous amount of data that require application of advanced computational methods such as data mining techniques, to discover the useful information and knowledge. The remarkable feature of gene expression microarray data for cancer classification is the number of variables (genes) far exceeding the number of samples. in such a high dimensional space work is extremely difficult and traditional statistical methodologies in classification and prediction do not work well.

This paper proposes and compares use of different data mining algorithm for detection of cancer by different algorithm. We emphasize that in data mining the goal is to discover knowledge that is not only accurate, but also comprehensible for the user. Comprehensibility is important whenever discovered knowledge will be used for supporting a decision made by a human user. After all, if discovered knowledge is not comprehensible for the user, he/she will not be able to interpret and validate it. In this case, probably the user will not have sufficient trust in the discovered knowledge to use it for decision making. This can lead to incorrect decisions. There are several data mining tasks, including classification, regression, clustering, dependence modeling, etc. Each of these tasks can be regarded as a kind of problem to be solved by a data mining algorithm. Therefore, the first step in designing a data mining algorithm is to define which task the algorithm will address.

For the discovery of the pattern we will be using association analysis. Association analysis, it came into prominence by the help of barcode technology which resulted construction of transactional databases in markets. Later it was thought that it would be beneficial to find frequently purchased items in the markets in order to increase sales. In [3], association rule mining was introduced as a new data mining technique which could be used for market basket analysis. Association rule mining, searches for items frequently purchased together when the market domain is considered. For instance, through frequent pattern mining, it can be found that sugar and tea are purchased together frequently. Placing these items closer or going on a discount may increase sales. Mining frequent patterns is used in many applications, however, it is not that much applied to bioinformatics. Microarray is an effective technique used in molecular biology to analyze gene expression in different conditions such as control versus drug treatment or healthy versus disease conditions in many organisms. Huge amount of data has been generated already and it is difficult to handle such large amount of data for analysis. Mining frequent patterns in gene expression data seems applicable and may help bioinformatics researchers. In the gene expression analysis context, we are looking for which genes are frequently

expressed together in different experiment conditions. This will lead us to knowledge of genes that are expressed association in defined conditions hereby this will facilitate the molecular biologists to define components of biological pathways with further studies. In this study we applied a frequent pattern mining algorithm to a DNA microarray gene expression data set. Applying frequent pattern algorithm, firstly, gene expression data will be transformed into transactional dataset like market basket datasets [5]. Over-expressed genes will be considered as purchased items and experiment conditions as transactions. After transformation, frequent pattern mining algorithm will be applied on this generated transactional dataset. At last, we will gain the knowledge to help us determine the possible interactions between different genes. After finding the pattern for the gene expression usage of ant colony optimization (ACO) for classification in data mining which will also predict the gene causing cancer. Data mining will have set of data to express patterns in a way which can be used for intelligent decision making. In this paper (as with many other data mining applications) the knowledge is presented in the form of a rule. A rule consists of an antecedent and a consequent. The antecedent is a set of attribute value pairs, for instance, the attribute age with a value of 17. More precisely a rule is of the form:

IF <attribute> = <value> AND <attribute> = <value> AND … THEN <class>

The class part of the rule (consequent) is the class predicted by the rule for the records where the predictor attributes hold. The class should be an attribute value whose prediction is "useful" for the user as there is no point in discovering knowledge that does not serve any purpose. An example rule might be IF gene expression = p53 missing THEN cancer is breast cancer, which would be useful for a user looking for a relationship between gene expression and type of cancer. The goal of classification in terms of data mining is to use the patterns, and so rules generated from a set of training data and apply them to a set of data with an unseen class, and hopefully predict the correct class.

Ant colony optimization (ACO) is a recently proposed meta-heuristic approach which mimics the behavior of ant colonies path finding from nest to food source and can be applied in solving hard combinatorial optimization problems or the control of multi-agent systems [9]. To improve the cancer classification performance, it is necessary to remove features that are irrelevant to cancers. In the paper, a two-stage classification method is proposed. The ant colony optimization algorithm is introduced to select critical genes relevant to cancers.

After the data collection the data will be preprocessed that is removal of noisy data then the different algorithms will be used on that dataset.

II. METHODOLOGY

2.1. Data collection

Microarray technique is simply depicted mRNA is isolated from organism of interest in control and test conditions than mRNA is converted to cDNA with reverse transcription reaction than, cDNA samples from control and test conditions labeled with different florescent tags. The two pools of labeled cDNA are mixed and hybridized to the DNA microarray containing a full set of tens of thousands of DNA sequences based on genomic or complimentary DNA (cDNA) sequences after hybridization, the chip is washed. Finally, the microarray array is scanned using a specialized fluorimager, and the color and intensity of each spot is determined. According to the colors and intensities of each spot expression levels of genes can be calculated by contrasting control versus treatment arrays [12]. As a result of a microarray experiment, datasets containing long lists of measurements of spot intensities and intensity ratios which are generated either by pair wise comparison of two samples or by comparing several samples to a common control are generated. Differentially expressed genes or co-regulated genes with related function can be identified by analyzing the generated dataset with some complicated softwares [13].

2.2. Data pre-processing

Genes, cDNA clones, or expressed sequence tags [ESTs] most often constitute DNA sequences that are scanned by microarray experiments, conditions contingent. They could include time series data of any biological process, viz., life cycle of a yeast cell, or a collection of varied tissue samples, e.g., normal versus cancerous tissues. For such, a gene expression matrix is acquired, which for clear reasons, contains gene data, not resisting a compendium of noise, missing values and inappropriate data. Data pre-processing is vital before any cluster analysis. Study on promoter sequences [4] and enhancer sequences [4] helps deriving transcription factors of an associated gene. Regulating transcription is the most common method of gene control. The action of transcription factors permits genes to be controlled during development and also in different cell types [5]. It is guaranteed to prune irrelevant data during preparation for further analysis. The process employs gene expression measures to discover co-regulated genes. To limit "noise" in the data, replication is used, thus ensuring precipitation of genes whose expression levels mark the outliers [6]. For this, methods like fold-change and significance analysis of microarrays (SAM) are put to work. Fold-change is a simple method, while SAM operates on certain statistical assumptions. Fold-change functions on choosing/discarding genes with a predetermined threshold level (usually a factor of 2) [7]. It then compares this level with the mean level of the gene expression, thus choosing/rejecting genes on basis of the calculation. SAM emphasizes the use of t-tests to calculate false discovery rate (FDR). The analysis based on t-tests can determine the probability of difference in gene expression getting monitored by chance. The highlight of the problem is that clustering genes on basis of their expression levels is not the sole criteria, with other factors such as experimental conditions, external factors (epigenetics), criteria for choosing base platform for genes (common ground), and their [genes] association with other genes. Certainly, a microarray consists of vivid groups of co-expressed genes. A common strategy is to commence with the ones that are related to some ordinal biological function and/or out

of a preliminary clustering result. Supervised and semi-supervised learning techniques use existing domain knowledge to filter particular genes [7], [8]. Clustering is thus a highly conditioned, specific filtering of gene expression profiles.

2.3. Data post-processing

2.3.1 Association Rule

Association rule mining aims at discovering associations between items in a transactional database. Given a set of transactions D = {T1,…..,Tn} and a set of items I = {i1, …., im} such that any transaction T in D is a set of items in I, an association rule is an implication of the form A B where the antecedent A and the consequent B are subsets of a transaction T in D, and A and B have no common items. For the association rule to be acceptable, the conditional probability of B given A has to be higher than a threshold called minimum confidence. Association rules mining is a two-step process, in the first step frequent item-sets are generated (i.e. item-sets whose support is no less than a minimum support) and in the second step association rules are derived from the frequent item-sets obtained in the first step. Association rule is also used in bioinformatics, web usage mining[14].

Apriori algorithm

Apriori is Latin word which mean "from what comes before". It uses bottom up strategy. It uses Breadth first search Method (B.F.S).It is used in various treatment of disease like Cancer. It is the most commonly used Method in Frequent pattern mining. This method find frequency itemset using candidate generation method. In this itemset are sorted in lexicographic manner. Apriori property is any subset of Frequency itemset is always frequent. Pseudo Code for Apriori algorithm:

1. Join Step: Ck is generated by joining Lk-1 with itself.
2. Min support: It is the minimum support used for searching frequency pattern that satisfy this constraint.
3. Min Confidence: It is used for finding the strong association rule that satisfy this threshold.
4. Prune Step: If (k-1) is not frequent itemset then the subset of (k) is also not frequent.

Ck: - Candidate itemset of size k.
Lk :-Frequency itemset of size k.
L1 ={Frequency items};
For (k=1;Lk!=0;k++) do begin
Ck.
Ck+1 =Candidate generated from Lk.
For each transaction'd' in database do increment the count of candidate in Ck+1 that contained in d.
Lk-1 = candidate in Ck-1 with min support End Return L1 U L2……Lk; [14]

A. Fp-growth algorithm

Fp-growth algorithm is used for generation of frequency itemset without candidate generation thus improves performance of algorithm. This method make used of Divide and Conquer strategy. This take place in 2 steps:-

Step 1: It compresses the input database showing frequency itemset into fp-tree.fp-tree is build using 2 passes on dataset.
Step 2: It then divide fop-tree into set of conditional dataset and mines them separately. Thus extract the frequency item set from fp-tree.

Fp-Tree Structure is as shown below:-

1. One root labelled as "null" with a set of item prefix sub trees as children, and a frequent-itemheader table.
2. Each node in the item-prefix sub tree consists of three fields:-
   i. Item-name: registers which item is represented by the node.
   ii. Count: the number of transactions represented by the portion of the path reaching the node.
   iii. Node-link: links to the next node in the FP-tree carrying the same item-name, or null if there is none.
3. Each entry in the frequent-item-header table consists of two fields:-
   i. Item-name: as the same to the node;
   ii. Head of node-link: a pointer to the first node in the FP-tree carrying the itemname.

PSEUDO CODE FOR FP-GROWTH ALGORITHM

Input: constructed FP-tree.
Output: complete set of frequent patterns.
Method: Call FP-growth (FP-tree, null). Procedure FP-growth (Tree, α)
{
1) If Tree contains a single path P then
2) For each combination do generate pattern β U α with support = minimum support Of nodes in β.
3) Else for each header ai in the header of Tree do {
4) Generate pattern β = aiUα with support = ai.support;
5) Construct β.s conditional pattern base and then β.s conditional FP-tree Tree β
6) If Tree β = null 7) Then call FP-growth (Tree β, β)}

} [14]

2.3.2. Ant Colony Optimization (ACO)

An artificial Ant Colony System (ACS) is an agent-based system which simulates the natural behavior of ants and develops mechanisms of cooperation and learning. ACS was proposed by Dorigo et al. (Dorigo, 1996, P.1-13) as a new heuristic to solve combinatorial-optimization problems. This new heuristic, called Ant Colony Optimization (ACO), has been shown to be both robust and versatile – in the sense that it can be applied to a range of different combinatorial optimization issues. ACO is also a population-based heuristic. This is advantageous because it allows the system to use a mechanism of positive feedback between agents as a search mechanism. There has lately been a growing interest in developing rule discovery algorithms based on other kinds of population-based heuristics – mainly evolutionary algorithms (Rafael, 2002, P.190-208 &Dorigo, 1999, P.137-172). In the next subsection we provide a brief overview of Artificial Ant Colony, and then describe the proposed algorithm which named FUZZY-ACO.Artificial ant colony An Artificial Ant Colony for pattern classification problem as followed: Number of ants in Ant Colony: number of fuzzy rules for classification. Feature of Ants: features and attributes in breast cancer domain. The amount of pheromone associated with each Ant: The amount of evaluation function associated with each rule. Pheromone updating: optimize the current classifier fuzzy rule-base.

Current Ant Colony: current rule-base. Modify Current Ant Colony: modify current rule-base. New Ant Colony: new current rule-base. Calculate the fitness of new Ant Colony: Calculate the evaluate function for new rule base. Admission of new Ant Colony with specific probability and improvement of fitness: Admission of new rule-base with specific probability and improvement of evaluation function

ACO Based Gene Selection Algorithm

The paper proposes an ACO based gene selection method to find out the critical genes for cancer classification when using microarray data. Real ants are capable of finding the shortest path from a food source to the nest even there is an obstacle between them. The ants communicate with each other according to its chemical trail called pheromone. The main characteristics of ant colony are distributed computation, positive feedback and the use of a constructive greedy heuristic [15]. The original application of ant system is the traveling salesman problem (TSP). Given a set of n towns, the TSP can be stated as the problem of finding a minimal length closed tour that visits each town once. In this paper, each gene is viewed as a town (node) on the TSP problem. The nodes on the tour generated by the ant colony are the selected genes for cancer classification. Given a set of n genes, let () ij t τ be the intensity of pheromone trail between gene pair (i, j) at time t. Each ant at time t chooses the next gene, where it will be at time t + 1. Let d (d < n) be the number of genes we want to choose from the original gene set. Therefore, if we call an iteration of the ACO algorithm the m moves carried out by the m ants in the interval (t, t + l), then every d iterations of the algorithm (which we call a cycle) each ant has completed a tour. The trail intensity is subsequently updated according to the following equation:

$$\tau_{ij}(t+1) = \rho\tau_{ij}(t) + \sum_{k=1}^{m} \Delta\tau_{ij}{}^{k}(t) \qquad (1)$$

where ρ is a coefficient such that (1-ρ) represents the evaporation of trail between time t and t + 1, and $\Delta\tau_{ij}^{k}(t)$ is the pheromone trail laid on edge (i,j) by the kth ant between time t and t + 1; it is given by

$$\Delta\tau_{ij}{}^{k}(t) = \begin{cases} \dfrac{Q.G_k}{n_k} & if\ \ kth\ ant\ select\ \ edge(i,j)\ in\ \ its\ tour \\ 0 & otherwise \end{cases} \qquad (2)$$

Where Q is a constant and Gk is the tour length of the kth ant. Usually the tour length is defined by the sum of the Euclidean distance between two nodes. However, the Euclidean distance between two genes is meaningless. Let$\sigma_i$be the standard deviation of the ith (i = 1,…, n) gene expression level in all of the samples. In this paper, the tour length of the kth ant is defined as

$$Gk = \sum_{i=1}^{d} \sigma i \qquad (3)$$

The coefficient ρ must be set to a value less than 1 to avoid unlimited accumulation of trail [15].

In our experiments, we set the intensity of trail at time 0 to

$$\tau_{ij}(0) = \left(c + \frac{\sigma_i + \sigma_j}{2} \right) \qquad (4)$$

Where C is a constant.

The transition probability from gene i to gene j for the kth ant is defined as

$$p_{ij}^{k}(t)=\begin{cases}\dfrac{[\tau_{ij}(t)]^{\alpha}\cdot[\eta_{ij}(t)]^{\beta}}{\sum_{k\in allowed_k}[\tau_{ik}(t)]^{\alpha}\cdot[\eta_{ik}(t)]^{\beta}} \\ 0\end{cases} \quad (5)$$

Where η_{ij} is called the visibility and is defined as
$\eta_{ij} = \sigma_i + \sigma_j$ allowed is a set containing the genes not been selected by the kth ant, and where α and β are parameters that control the relative importance of trail versus visibility. Let $G^k(t)$mean the set of genes selected by the kth ant at time t, G(t) be the set of genes selected by the m ants at time t.

$$G(t)=\bigcup_{k=1}^{m} G^{k}(t) \quad (6)$$

Given the above definitions, the proposed ACO based gene selection algorithm is simply stated as follows.

(1) Initialize:

Set t: = 0 {t is the time counter}
Set NC: = 0 {NC is the cycles counter}
For every edge (i, j) set an initial value τij(0) according to Eq. (4). Place the m ants on the different nodes

(2) For k := 1 to m do Place the starting node of the kth ant in $allowed_k$.

(3) Repeat (d - 1) times
For k := 1 to m do Choose the node j to move to, with the maximum probability $p^k_{ij}(t)$given by Eq. (5)
Move the kth ant to the node j

(4) For k := 1 to m do
Compute the length G_k, of the tour described by the kth ant
Update the tour found
For every edge (i, j)
For k:= 1 to m
Compute τ_{ij} (t+1) according to Eq. (1)

$$G(t)=\bigcup_{k=1}^{m} G^{k}(t)$$

Set t: = t + 1

(5) If (NC <NCmax)
Then Set NC := NC + 1
Goto step 2
Else
Output G(t)
Stop

III. ANALYSIS

As compared to apriori algorithm fp-growth is more efficient as time required to execute is less than apriori and also memory utilization [16] is less in fp-growth. So we find that fp-growth is more efficient to use as comparison shown in table Table1 .As these both algorithms will be using ant colony optimization as prediction for the cancer

Table1. Comparison of fp-growth algorithm & apriori using ant colony optimization for cancer gene detection

Parameter	FP-Growth Algorithm & ant colony	Apriori Algorithm & ant colony
Memory utilization	As no Candidate are generated so memory utilization is less.	Due to candidate generation memory utilization is more.
No of Scans	Scan the database twice only.	Multiple scans for generating candidate sets.
Time Required	Time require is less.	Time require is more as for generating candidate set require more time
Efficiency	More Efficient	Less efficient

IV. CONCLUSION AND FUTURE SCOPE

This research work provides a valuable knowledge on cancer causing gene. It is observed that a reliable and precise classification of tumors is essential for successful diagnosis and treatment of cancer. By allowing the monitoring of expression levels in cells for thousands of genes as well as, microarray experiments may lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more informative classification. The ability to successfully distinguish between tumor classes (already known or yet to be discovered) using gene expression data is an important aspect of this novel

approach to cancer classification. Also annotated is that comparing the activity of genes in a healthy and cancerous tissue may give some hints about the genes that are involved in cancer. By using the frequency pattern technique over apriori algorithm the usage of memory is less and thereafter for the classification and the prediction ant colony optimization is used. It is conclude that by using the frequency pattern and ant colony optimization the memory usage is less. Also, moving towards an era of personalized or precision medication, NGS and Gene Therapy are making their mark there.

REFERENCES

[1] Data and Statistics. World Health Organization, Geneva, Switzerland, 2006.
[2] R. Siegel, J. Ma, Z. Zou, and A. Jemal, "Cancer statistics," CA: A Cancer J. Clinicians, vol. 64, pp. 9–29, 2014.
[3] "International Agency for Research on Cancer (IARC)," WHO,W. Stewart and C. P. Wild eds., World Cancer Report, 2014.
[4] PubMedHealth- U.S. Nat. Library Med., (2012). [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002267/
[5] 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.
[6] N. Pasquier, C. Pasquier, L. Brisson, and M. Collard, (2008). "Mining gene expression data using domain knowledge," Int. J Softw. Informat, vol. 2, no. 2, pp. 215–231, [Online] Available: http://www.ijsi.org/1673-7288/2/215.pdf.
[7] K. M Williams, "Statistical Methods for analysing microarray data: Detection of differentially expressed genes" Inst. Signal Process., Tampere Univ. Technol. Tampere, Finland, Dep. Biology, Univ. York, York, U.K., 2004.
[8] B. Vilo, "MiniReview-Gene expression data analysis," FEBS Lett. vol. 480, pp. 17–24, 2000.
[9] Dorigo, M., Bonabeau, E., & Theraulaz, G.., "Ant algorithm and stigmergy", Future Generation Computer System, Vol. 16, pp. 851-871, 2000.
[10] http://web.iyte.edu.tr/~barisyildiz/publications/hibit2011_pp229-233.pdf
[11] Barış Yıldız and Hatice Şelale. "Mining Frequent Patterns from Microarray Data". In The 6th International Symposium on Health Informatics and Bioinformatics, (HIBIT 2011), İzmir, Turkey, pp. 229-233, May 2-5, 2011.
[12] S. Albelda, D. Sheppard, "Functional Genomics and Ex-pression Profiling," *Am. J. Respir. Cell Mol. Biol.*, vol. 23, pp. 265-269, 2000.
[13] A. Schulze and J. Downward, ―Navigating gene expres-sion using microarrays — a technology review*," Nature Cell Biology*, vol. 3, pp. 190-195, 2001.
[14] Data Mining Techniques For Diagnosis And Prognosis Of Breast Cancer Jaimini Majali et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6487-6490
[15]. Dorigo, M., Maniezzo, V., & Colorni, A., "The ant system: optimization by a colony of cooperating agents", IEEE Trans. Systems, Man, Cybernetic. B, Vol. 26, No. 1, pp. 29-42, 1996.
[16] Kazi Mahbub Mutakabbir, Shah S Mahin & Md. Abid HasanMining Frequent pattern within a Genetic Sequence Using unique Pattern Indexing and Mapping Techniques 3rd INTERNATIONAL CONFERENCE ON INFORMATICS, ELECTRONICS & VISION 2014