IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. V (May-Jun. 2016), PP 62-67
www.iosrjournals.org

Detection of Breast Cancer by the Identification of Circulating Tumor Cells Using Association Rule Mining

R. Nedunchelian[1], S. Jananee[2]
[1]*Professor & Head,* [2]*Student, Department of Computer Science and Engineering,*
Sri Venkateswara College of Engineering.
nedun@svce.ac.in, jananeesridhar@gmail.com

Abstract: *Circulating Tumor Cells (CTCs) are cells that have shed into the vasculate from the primary tumor and circulate into the blood stream. In this proposed work, the major genes causing the breast cancer is identified by the principle of Association Rule. The trained set and training set is made to upload on the data store. By associating each row of a training set to all the rows of the trained data is done and the report is generated. The Baum Welch process is called for the estimation of actual probabilities and emission probabilities by calculating its log likelihood factor which gives the high Priority gene values that are responsible for the cause of cancer. Based on this cell category is split into three clusters such as carcinoma level, metastasis level and Kaposi sarcoma. On each cluster it finds the highest priority value in it and classifies into high, low and medium values. On extraction of these higher gene values yields the major responsible genes causing breast cancer. Finally obtained results are validated through hierarchical clustering.*

Index Terms: *Circulating tumor cells, Association Rule, Cluster, Baum Welch, actual probabilities, emission probabilities, High Priority Genes.*

I. Introduction

Breast cancer is the group of cancer cells that starts developing in the cells of breast. The term Breast cancer refers to a malignant tumor that has developed from cells in the breast. BC starts with in the cells of the breast as a group of cancer cells that can invade its surrounding tissues or spread to other areas of the body. In general the cancer-related death (BC) is the consequences of tumor cells that start spreading from the primary tumor and forms metastases in resident organs. Cancer metastasis is the main cause of cancer-related death and the dissemination of tumor cells through the blood circulation is an important intermediate step that also exemplifies the switch from localized to systemic disease. Circulating tumor cells in the peripheral blood (PB) arise from the primary tumor and they are indicative for the tumor aggressiveness and metastasis. Several discriminate factors have to be identified in detecting the BC.

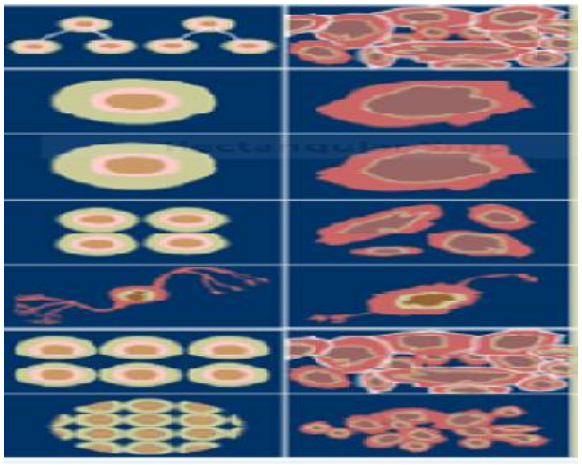

Fig 1.1 NORMAL CELLS VS CANCER CELLS

These are also processed by applying the data mining techniques to the datasets. The process of obtaining the golden information from the raw data is termed as data mining. These data are collected from the Wisconsin databases and GEO Databases. The raw data will not be sufficient to manipulate, for this data pre-processing is made. The data pre-processed will be rich in information which omits the missing values and attributes. Data modeling involves a logic solution with the help of decision trees and decision rules. Data modeling gives an interpretation and conclusion to the whole process.

A. Association Rule Mining:

Association rule mining is the discovery of association relationships among a set of items in a dataset. Association rule mining has become an important data mining technique that correlates the presence of set of items with another range of values for the set of variables. Association rule mining is used to extract association from the market based data which was suggested by **Agarwal et al (1993)**. It has also proved to be useful in many other domains such as microarray data analysis, recommender systems, and network intrusion detection.

An association rule is of the form, **X Y**

Where X = and Y = are sets of genes items, with x_i and y_j being distinct items for all *i* and all *j*. This association states that if a gene is chosen as a victim X, it is also likely to choose Y. In general, any association rule has the form LHS (left-hand side) RHS (right-hand side), where LHS and RHS are sets of items. Association rules should supply both support and confidence.

Association Rule Generation:

The goal of mining association rule is to generate all possible rules that exceed some minimum user-specified support and confidence thresholds. The problem is thus decomposed into two sub problems:

1. Generate all item sets that have a support that exceeds the threshold. These sets of items are called **large item sets.** Note that large here means large support.

2. For each large item set, all the rules that have a minimum confidence are generated as follows: for a large item set X and Y X, let Z = X - Y;

Then if support (X)/support (Z) minimum confidence, the rule Z Y (i.e., X - Y Y) is a valid rule. [Note: In the previous sentence, Y X reads "Y is a subset of X."]

Association Rules among Hierarchies:

The association occurs among hierarchies of items. Typically, it is possible to divide items among disjoint hierarchies based on the nature of the domain. Items can be categorized into classes and subclasses that give rise to hierarchies. It should also have enough confidence and support to be valid association rules of interest. Therefore, the application area has a natural classification of the item sets into hierarchies, discovering an association within the hierarchies is of no particular interest. The ones of specific interest are associations across hierarchies.

II. Related Studies

In [1] authors suggested an indirect method is used for the identification of major factors in the peripheral blood that reveals the presence of such CTC. By using the selected publicly available breast cancer and peripheral blood microarray data sets they used a two-step elimination procedure for the identification of several discriminate factors.

The First stage aims to extract the gene signatures associated with pair wise differentiation between cell types or disease states. The Second stage considers the intersection of the previous signatures of the blood and tissue samples. A comparison is made between the cancer tissue, normal tissue and the peripheral blood samples of the cancer patients and normal individuals. So these comparisons lead to the closer association with the existence of CTCs. The second stage considers the intersection of these signatures. This intersection of all three comparisons is expected to derive a gene signature, which is indicative of the presence of CTCs and it can be directly compared with factors used for the isolation of such cells.

Progression model and metastatic predestination model provide evidence about tumor progression toward metastasis. In order to gain further insight on our results, we test the Performance of the remaining 24 genes in the intersection of all the three comparisons on two independent datasets. As a result among 24 common gene signatures only seven were identified and meant to be responsible for breast cancer.

Most cancer events are diagnosed in the late phases of the illness and so the early detection in order to improve breast cancer outcome and survival is very crucial. Predictions of the other patients are realized through seven different algorithms and the accuracies of those have been given. During the prediction process, Rapid Miner 5.0 data mining tool is used to apply data mining with the desired algorithms [2].

The spread of cancer relates to the detachment of malignant cells into blood. The explosion of genomic sequence and molecular profiling data has illustrated the complexity of human malignancies. In a tumor cells dozens of different genes may be aberrant in structure or copy number, hundreds and thousands of genes may be differently over expressed [3].

The information regarding secondary metastasis tumor is discussed. Metastatic colorectal cancer relies on detachment of aggressive malignant cells from the primary tumor into the bloodstream and, concordantly, the presence of this CTC is associated with a poor prognosis [4]. Specific alterations in cancer might be indicative of its ability to diffuse such genes can indirectly predict existence of Circulating Tumor Cells without the need to detect or extract them [5].

III. Methods & Procedures

Breast Cancer Data Sets:

Data sets used in this work are collected from Wisconsin repository and GEO database. Data pre-processing is employed to remove the missing values, redundant data etc... The following set of attributes given in table 3.1 is to be considered while choosing the datasets which is listed in the below table. Along with this some of the major factors are considered along with the detection of genes.

TABLE 3.1 SET OF ATTRIBUTES

S.NO	ATTRIBUTES
1.	Sample code number
2.	Category
3.	Clump thickness
4.	Uniformity of cell shape
5.	Uniformity of cell size
6.	Marginal adhesion
7.	Single epithelial cell size
8.	Bare nuclei
9.	Bland chromatin
10.	Normal nucleoli
11.	Mitosis
12.	Images
13.	Class

Proposed Work

The work focuses on identifying the rest of genes by building the association rule for the identification of breast cancer. The Training data set is pre-processed and allowed to store in a Data store. Where the trained data is provided as a reference set. Based on the mapping of the each row of a training set to all the rows of the trained data set is made and the report is generated.

Baum Welch algorithm is called for the estimation of actual probability and emission probabilities by calculating its log likelihood which gives the high Priority gene values. From this cluster based cell category is split into three clusters. On selecting these cluster finds the highest value in it and classifies into high, low and medium values. On extracting these high gene values yields the major possible genes causing breast cancer.

Finally obtained results are validated through hierarchical clustering. The hidden Markov model seeks to improve the given model to one that it could have more likely generated given output sequence. It does so by using both forward and backward algorithm and using the output of these two algorithms to create temporary variables which can then be used to improve the initial probabilities, transition probabilities and emission probabilities.

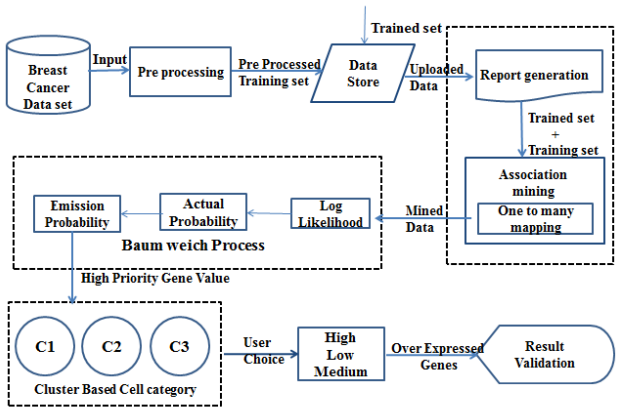

Fig 3.1 PROPOSED ARCHITECTURE

The figure 3.1 represents the proposed system architecture. Pre-processing is made with the raw data to extract the useful gene information. Factors such as Family history, age, alcohol consumption, smoking, obesity, Birth and breast feeding, physical activity, Endogenous Hormone Exposure, Exogenous Hormone Exposure, Age at first Menstruation, Breast tissue , breast feeding are considered as the major factors causing breast cancer.

Table 3.2 Factors Causing Breast Cancer

S. NO	FACTORS	INDICATORS
1.	Family history	Inheriting gene mutation
2.	Age	Above 50 years
3.	Smoking	After pregnancy
4.	Alcohol Consumption	Regular consumption
5.	Obesity	Increased weight of body
6.	Birth and breast feeding	Lack of Breast feeding at time of birth
7.	Physical activity	Due to lack of physical activity
8.	Endogenous Hormone Exposure	High rate of Presence
9.	Exogenous Hormone Exposure	High rate of Presence
10.	Age at first Menstruation	Later Period of menstruation
11	Breast tissue	Dense Breast Tissue (thickness)
12.	Breast feeding	Lack of Breast feeding

Report Generation:

Data is collected from various sources and pre processed. Pre-processing removes the inconsistent data, missing values, redundant values and gives the consistent data for further implementation. The pre-processing involves data cleaning and gives only the valuable information in terms of counts. The datasets are collected from Wisconsin database, UCI machine learning repository, Affymetrix database and GEO database. These data may contain noisy, missing and inconsistent data. Likewise the training sets are processed and shown in the fig 3.2 below.

Cell Identity	Clump_Thickness	Uniformity_of_Cell_Shape	Class	Bland_Chromatin	Normal_Nucleoli	Carcinoma Level	Metastasis Level	Kaposi sarcoma	Epstein
100001	8	20045	2	876	4	2.9583437266151162	1.1585160202360878	34.27288600648541	6.408094435
100001	8	20045	2	876	4	2.9583437266151162	1.1585160202360878	34.27288600648541	6.408094435
100001	8	20045	2	876	4	2.9583437266151162	1.1585160202360878	34.27288600648541	6.408094435
100002	6	4665	4	356	4	11.082529474812434	0.18568665377176016	20.578778135048232	7.543520309
100002	6	4665	4	356	4	11.082529474812434	0.18568665377176016	20.578778135048232	7.543520309
100002	6	4665	4	356	4	11.082529474812434	0.18568665377176016	20.578778135048232	7.543520309
100003	17	8712	2	862	4	0.4132231404958678	2.1111111111111111	8.723599632690542	269.4444444
100003	17	8712	2	862	4	0.4132231404958678	2.1111111111111111	8.723599632690542	269.4444444
100003	17	8712	2	862	4	0.4132231404958678	2.1111111111111111	8.723599632690542	269.4444444
100004	6	77813	4	227	2	0.339274928353874	0.9356060606060606	3.1742767917957155	22.727272727

Fig 3.2 REPORT GENERATION

Classification Based On Association Rule:

The trained set and the training sets are uploaded are uploaded on the data store. The map function is called by associating the training set and trained set by selecting one tuple associating it with all the tuple of the trained data set. Likewise all the full association is made between the trained and the training set. One-to-many mapping is made both the sides of the trained and training sets. After associating these data the report is generated by combining both the trained and the training set. The associated data is transformed to the Baum Welch process and it is shown in the below Fig 3.3

Cell Identity	Clump_Thickness	Uniformity_of_Cell_Shape	Class	Bland_Chromatin	Normal_Nucleoli	Carcinoma Level	Metastasis Level	Kaposi sarcoma	Epstein
100001	8	20045	2	876	4	2.9583437266151162	1.1585160202360878	34.27288600648541	6.408094435
100001	8	20045	2	876	4	2.9583437266151162	1.1585160202360878	34.27288600648541	6.408094435
100001	8	20045	2	876	4	2.9583437266151162	1.1585160202360878	34.27288600648541	6.408094435
100002	6	4665	4	356	4	11.082529474812434	0.18568665377176016	20.578778135048232	7.543520309
100002	6	4665	4	356	4	11.082529474812434	0.18568665377176016	20.578778135048232	7.543520309
100002	6	4665	4	356	4	11.082529474812434	0.18568665377176016	20.578778135048232	7.543520309
100003	17	8712	2	862	4	0.4132231404958678	2.1111111111111111	8.723599632690542	269.4444444
100003	17	8712	2	862	4	0.4132231404958678	2.1111111111111111	8.723599632690542	269.4444444
100003	17	8712	2	862	4	0.4132231404958678	2.1111111111111111	8.723599632690542	269.4444444
100004	6	77813	4	227	2	0.339274928353874	0.9356060606060606	3.1742767917957155	22.727272727

Showing 1 to 10 of 2,990 entries

Fig 3.3 ASSOCIATION PROCESS

E. Processing of Baum Welch Algorithm:

Baum-Welch is an iterative procedure that helps to find the unknown parameters of Hidden Markov Model (HMM) by calculating the actual and emission probabilities. The HMM describes joint probability of a collection of 'hidden' and observed discrete random variables. It works by maximizing a proxy to the log likelihood, and updating the current model to be closer to the optimal model and the likelihood calculation is shown in fig 3.4. Each iteration on Baum Welch guarantees to increase the log-likelihood of the data.

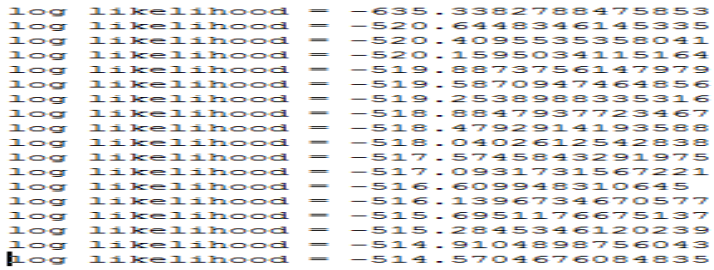

Fig 3.4 LIKELIHOOD CALCULATION

The Baum-Welch algorithm is used for learning the model parameters of the Hidden Markov Model. Baum-Welch is simply an instantiation of the more general Expectation-Maximization (EM) algorithm and the probabilities of Baum Welch are shown in the below figure 3.5. The EM algorithm is used to find locally maximum likelihood parameters of a statistical model in cases where the equations cannot be directly solved. These models involve latent in addition to unknown parameters and known data observations. To solve these two sets of equations numerically, pick the arbitrary values for one values for one of two sets of unknowns, use these new values to find a better estimate of first set, and keep alternating between two until the resulting values both converge to fixed points.

Item	Location	Session	Carcinoma Level	Metastasis Level	Kaposi sarcoma	Epstein	Pappiloma
0.0	0.0	0.0	0.491980	0.507677	0.0	0.000344	0.0
0.0	0.0	0.0	0.466106	0.000000	0.0	0.533894	0.0
0.0	0.0	0.0	0.345276	0.654724	0.0	0.000000	0.0

Fig 3.5 BAUM WELCH PROBABILITIES

IV. Results and Discussions

Extraction of Higher Priority Genes:

The High Priority Gene values are extracted from the Baum Welch processing by finding the actual probabilities and emission probabilities by the Hidden Markov Model which is shown in the figure below fig 3.6. Based on the Cluster Based Category is organized as carcinoma level, metastasis level and Kaposi sarcoma as CPC, CPM and CTR. It is calculated from Baum Welch emission probabilities. It generates value for each clusters based on this values it categories into High, Low and Medium gene values. The High valued genes are considered as the major responsible breast cancer causing Genes. Fig 3.7 shows the final result analysis.

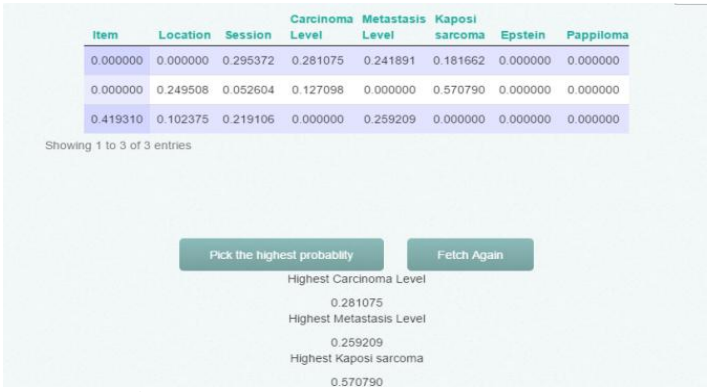

Fig 3.6 HIGH PRIORITY GENE VALUES

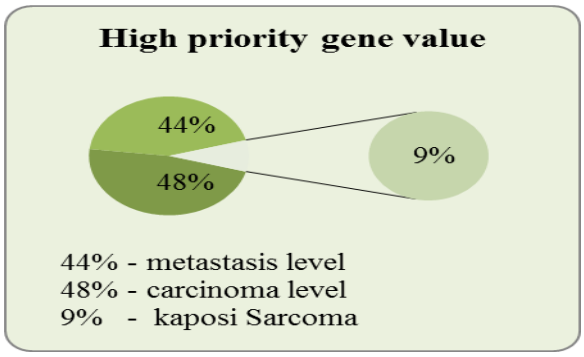

Fig 3.7 RESULT ANALYSIS

V. Conclusion

The detection of the circulating tumor cells helps in the identification of the major genes in the breast cancer. This is done by using the association rule technique of the data mining which is used for detecting the possible genes that plays a major role in detecting the breast cancer. So the proposed work focuses on developing an association rule for the efficient identification of the possible genes by using data mining technique – Association Principle. The obtained results are approximately equal to the biomedical results.

Future Work

Future work focuses on the prediction of the breast cancer which is very useful for the human survival since it the second deadly disease in the cancer world. On building these kind of new technology will help them to even predict the diseases at the earlier stage which can help the survival of the human being. Other direction will focus on the finding the stage of the BC by analyzing the datasets.

REFERENCES

[1] Stelios Sfakianakis, Ekaterini S. Bei, Michalis Zervakis, *Member, IEEE*, Despoina Vassou, and Dimitrios Kafetzopoulos, "On the Identification of Circulating Tumor Cells in Breast Cancer" IEEE Journal Of Biomedical And Health Informatics, Vol. 18, No. 3, May 2014.

[2] Zehra Karapinar Senturk and Resul Kara , "Breast Cancer Diagnosis Via Data Mining: Performance Analysis of Seven Different Algorithms" vol. 4, no. 2, 2014.

[3] J. Barbaz´an, L. Alonso-Alconada, L. Muinelo-Romay, M. Vieito, A. Abalo, M. Alonso-Nocelo, S. Candamio, E. Gallardo, B. Fern´andez, I. Abdulkader, M. de Los A´ ngeles Casares, A. Go´mez-Tato, R. Lo´pez and M. Abal, "Molecular characterization of circulating tumor Cells in human metastatic colorectal cancer," PloS One, vol. 7, 2012.

[4] Timothy J. Molloy, Paul Roepman, Bjorn Naume, Laura J. vant Veer A Prognostic Gene Expression Profile That Predicts Circulating Tumor Cell Presence in Breast Cancer Patients, 2012.

[5] Y. Park, T. Kitahara, T. Urita, Y. Yoshida, and R. Kato, "Expected clinical applications of circulating tumor cells in breast cancer," World J. Clin. Oncol., vol. 2, no. 8, pp. 303–310, 2011.

[6] D. C. Danila, K. Pantel, M. Fleisher, and H. I. Scher, "Circulating tumors Cells as biomarkers," Cancer J., vol. 17, no. 6, pp. 438–450, 2011.

[7] G. Shin, T.-W. Kang, S. Yang, S.-J. Baek, Y.-S. Jeong, and S.-Y. Kim, "GENT: Gene expression database of normal and tumor tissues," Cancer Inf., vol. 10, pp. 149–157, 2011.

[8] E. Obermayr, F. S. Cabo, M. K. Tea, C. Singer, M. Krainer, M. Fischer, J. Sehouli, A. Reinthaller, R. Horvat, G. Heinze, D. Tong, andR. Zeillinger, "Assessment of a six gene panel for the molecular detection of circulating tumor cells in the blood of female cancer patients," BMC Cancer, vol. 10, no. 1, p. 666, 2010.

[9] S. Riethdorf and K. Pantel, "Advancing personalized cancer therapy by Detection and characterization of circulating carcinoma cells," Ann. New York Acad. Sci., vol. 1210, no. 1, pp. 66–77, Oct. 2010.

[10] G. Shin, T.-W. Kang, S. Yang, S.-J. Baek, Y.-S. Jeong, and S.-Y. Kim,*"GENT: Gene expression database of normal and tumor tissues," CancerInf.*, vol. 10, pp. 149–157, 2011.

[11] H. G. LaBreche, J. R. Nevins, and E. Huang, *"Integrating factor analysis and a transgenic mouse model to reveal a peripheral blood predictor of breast tumors," BMC Med. Genomics*, vol. 4, p. 61, 2011.

[12] 11.J. Aaroe, T. Lindahl,V.Dumeaux, S. Saebo, D. Tobin, N. Hagen, P. Skaane, A. Lonneborg, P. Sharma, and A. L. B. Dale, "*Gene expression profiling of peripheral blood cells for early detection of breast cancer," Breast CancerRes.*, vol. 12, no. 1, p. R7, 2010.

[13] I. B. Pau Ni, Z. Zakaria, R. Muhammad, N. Abdullah, N. Ibrahim, N. Aina Emran, N. Hisham Abdullah, and S. N. A. Syed Hussain, "*Gene expression patterns distinguish breast carcinomas from normal breast tissues: The Malaysian context," Pathology—Res. Practice*, vol. 206, no. 4, pp. 223–228, Apr. 2010.

[14] G. Marot, J.-L. Foulley, C.-D. Mayer, and F. Jaffr´ezic, *"Moderated effect size and P-value combinations for microarray meta-analyses," Bioinformatics*,vol. 25, no. 20, pp. 2692–2699, Oct. 2009

[15] https://www.scribd.com/doc/306698398/Data-Warehousing