Using K-Means Clustering Technique To Study Of Breast Cancer

R.Radha[1]
[1]Department of Computer Science,
S.D.N.B. Vaishnave College of Women,
Chrompet, Chennai, Tamilnadu-India
e-mail: radhasundar1993@gmail.com

P.Rajendiran[2]
[2]Department of Computer Science, Vidyaa Vikas Educational Institutions, Tiruchengode, Namakkal,Tamilnadu- India
e.mail:pr.res2011@gmail.com

***Abstract*- Breast cancer is one of the most common cancers worldwide. In developed countries, among one in eight women develop breast cancer at some stage of their life. Early diagnosis of breast cancer plays a very important role in treatment of the disease. With the goal of identifying genes that are more correlated with the prognosis of breast cancer, we use data mining techniques to study the gene expression values of breast cancer patients with known clinical outcome. K-means clustering is used to compare the result based on test data. As a result, a set of genes are identified that are potential bio marks for breast cancer prognosis which can categorize the patients based on the certain attributes. A comparison is made on gene expression levels that are discovered with gene subsets identified from similar studies using clustering techniques.**

Keywords – Clustering, Breast Cancer, Gene, K-means, Tumor

I. INTRODUCTION

Cancer is a disease characterized by uncontrolled cell growth and proliferation. For cancer to develop, genes regulating cell growth and differentiation must be altered. These mutations are then maintained through subsequent cell divisions and are thus present in all cancerous cells. Approximately one – third of patients with early stage Breast Cancer(BC) experience disease recurrence after initial diagnosis[1]. Genome –wide association studies(GWAS) have made it possible to identify single nucleotide polymorphisms(SNPs), here in called genetic variants, that are associated with an increased risk of developing breast cancer[2][3]. Results from GWAS are providing valuable class about allelic architectures and are improving our understanding of the emerging genetic susceptibility landscape of breast cancer. However, despite these remarkable achievements, significant challenges remain. Although many compelling genetic variants have been found and replicated in multiple independent GWAS,[2][3] they explain only a small fraction of the variation. Importantly GWAS do not typically inform the broader context in which the genetic variants operate, leading to the development of breast cancer subtypes. As a consequence they provide limited insights about the different sub types of breast cancer.

The advent of micro array technology has made possible the identification of molecular signatures and molecular classification of subtypes of breast cancer based on mRNA expression profile[4][5]. However, although these primary analyses have identified clinically actionable biomarkers, they have been unsuccessful in determining which genes have casual roles as opposed to merely being consequences of disease states[2].

Incidence and mortality differ along racial lines[6]. Small BC tumors are often asymptomatic, but longer tumors often present as a painless, palpable mass[6]. Other, less common symptoms include breast pain and physical changes to the breast or nipple. The majority of patients with invasive BC are diagnosed at early stage of disease, and only 6% of patients presents with meta static disease[7]. In addition, BC recurrence varies considerably over time and is influenced by adjuvant therapeutic modalities. Despite recognized variablitity, most recurrences occurs within the first 5 years of diagnosis[8].

II. METHODOLOGY

Clustering involves partitioning a set of data points into non overlapping groups, or clusters of points where points in a cluster are more similar to one another than to points in other clusters[9]. In general clustering algorithms are classified into two categories[9][10] (hard clustering algorithms and fuzzy clustering algorithms). In hard clustering, each data point belongs to one and only one cluster, where in fuzzy clustering, each data point is allowed to have membership in more than one clusters.

A. K-Means Clustering

K-means clustering(K-Means)[11] is a simple and fast method used commonly due to its straightforward implementation and small number of iterations. This algorithm divides the data set into k disjoint subsets. An estimation of the number of clusters(k) is made by the user and calculated as an input where the computer randomly assigns each gene to one of the k-clusters. The distance between each gene and the center of each cluster is promptly calculated resulting in an optimal grouping of data to clusters where the genes inside every cluster are as close to the centre of the cluster as possible while at the

978-1-4799-2877-4/14 $31.00 © 2014 IEEE
DOI 10.1109/WCCCT.2014.64

same time there is maximal distance between genes of different clusters.

B. Experimental Setups

M.Zwitter and M.Soklic [13] presented a dataset that is based on the full field of breast cancer domain (Table 1) collected at university Medical centre , Institute of oncology, Ljubljana, Yugoslavia. This data set has 9 attributes with 286 tuples. In this database the attribute meno_pause is classified into three categories as “premeno”, “ge40” and “it40” depending on size of the tumor. The clusters are based on these turmor size.

TABLE- 1 – INFORMATION ON ATTRIBUTES OF DATASET.

Attribute	Type	Details on attributes
age	Integer	Patients age in integer
meno_pause	Categorial	premeno,ge40,it40
tumor_size	Integer	tumor_size in integer
inv_nodes	Integer	inv_nodes in integer
node_caps	Categorial	yes,no,?
deg_malig	Intger	deg_malig in integer
breast	Categorial	breast direction left,right
breast_quad	Categorial	breast_quad_ directions
irradiat	Categorial	yes,no

This study used Python with Orange open source software (a collection of machine learning algorithms for data mining tasks) for experiments[12], and also k-means clustering modules from python with orange software.

C. Phase – I

In K-means clustering, the distance measure used is Manhattan, and random setting is used for cluster initialization. Based on the above settings the resulted clusters optimization report is given in Table -2. In this all the data points (286) were clustered with the help of k-Means clustering method. In this we obtain maximum number of clusters as two, and found the score from one cluster to another cluster as different. In this the average probability of correct classification is found to be 66.99%.

TABLE – 2 CLUSTER OPTIMIZATION REPORT.

k	Best	Score
1		1606.5
2		1478.5

k-Means Clustering

Settings

Distance measure: Manhattan
Initialization: Random
Restarts: 4

Optimization

Minimum num. of clusters: 2
Maximum num. of clusters: 3
Scoring method: Between cluster distance

Data

Examples: 286
Attributes: 9 (age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad, irradiat)
Class: recurrence

D. Phase – II

In phase-II of K-means clustering, to improve the classification accuracy distance to centroids is used as scoring method, Euclidean is used as distance measure, random setting is used for cluster initialization. Based on the above settings the clusters optimization report obtained is given in Table -3. Here all the data points (286) were clustered with the help of k-Means clustering method. In this we obtain maximum number of clusters as two, and found that the score from one cluster to another cluster are different. In this the average probability of correct classification is found to be 74.98%.

k-Means Clustering

Settings

Distance measure: Euclidean
Initialization: Random
Restarts: 4

Optimization

Minimum num. of clusters: 2
Maximum num. of clusters: 6
Scoring method: Distance to centroids

Data

Examples: 286
Attributes: 9 (age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad, irradiat)
Class: recurrence

TABLE – 3 CLUSTER OPTIMIZATION REPORT.

k	Best	Score
1		0.0781
2		0.0770

E. Phase – III

In phase –III of K-means clustering, to further improve the classification accuracy distance to centroid is used as scoring method, Pearson Correlation is used for distance measure, agglomerative is used for cluster initialization. Based on the above settings the clusters optimization report is obtained is given in Table -4. In this all the data points (286) were clustered with the help of k-Means clustering method. In this we obtain maximum number of clusters as two, and found the score from one cluster to another cluster as different. In this Phase – III we obtain the optimization report as same for all the clusters. In this the average probability of correct classification is found to be 100%.

k-Means Clustering

Settings

Distance measure: Pearson Correlation
Initialization: Agglomerative clustering
Restarts: 4

Optimization

Minimum num. of clusters: 2
Maximum num. of clusters: 6
Scoring method: Distance to centroids

Data

Examples: 286
Attributes: 9 (age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad, irradiat)
Class: recurrence

TABLE – 4 CLUSTER OPTIMIZATION REPORT.

k	Best	Score
1		286.0
2		286.0

The comparative study of K-Means clustering algorithm based on different distance, scoring and clustering initialization method is given in Table – 5.

TABLE -5 COMPARATIVE STUDY OF K-MEANS CLUSTERING ALGORITHM FOR DATA POINTS WITH PREDICTION OF TUMOR SIZE

K-means clustering Phase	Distance	Scoring	Initialization	Average probability (%)
Phase -I	Manhattan	Between cluster distance	Random	66.99%
Phase – II	Euclidean	Distance to Centroids	Random	37.37%
Phase – III	Perason Correlation	Distance to Centroids	Agglomerative	100.00%

III. RESULTS AND DISCUSSION

All the data points(286) were clustered by K-Means clustering algorithms. Clustering results were compared with the breast cancer domain fields with the help of scatter plot diagram to predicate the tumor size dependency on menopause (genes).

In Phase – I clustering format it is found that in cluster (C1) tumor with maximum size are grouped in ge40 category and minimum size are grouped in it40 and very small size are grouped in premeno category. In cluster (C2) tumor with maximum size are grouped in premeno, minimum size are grouped in ge40 and very small size are grouped in it40 category. Fig.1 differentiate two clusters which are based on the field of tumor_size.

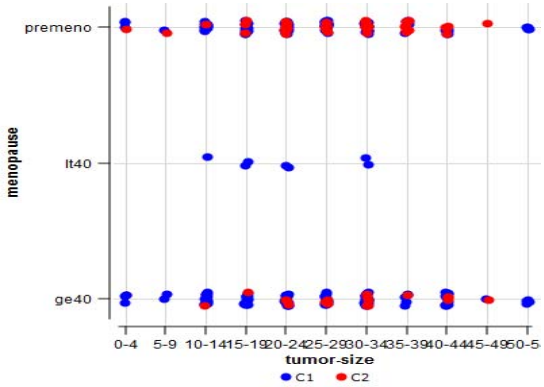

Fig.1. Two clusters based on tumor_size and menopause – Phase I

In Phase – II clustering format , it is found that in cluster (C1) tumor with maximum size are grouped in premeno category and minimum size are grouped in ge40 and very small size are grouped in it40 category. In cluster (C2) tumor with maximum size are grouped in ge40, minimum size are grouped in premeno and very small size are grouped in it40 category. Fig.2 differentiate two clusters based on the field of tumor_size.

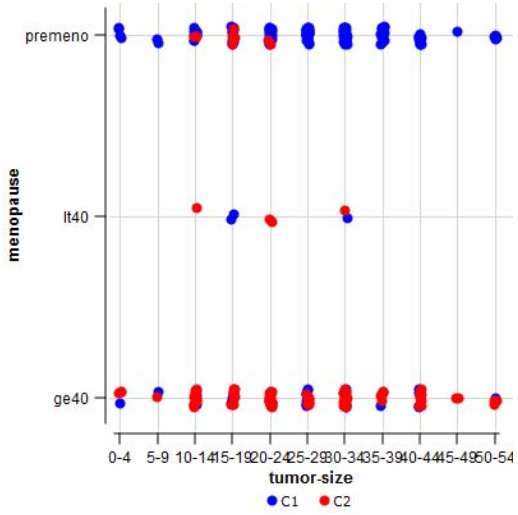

Fig.2. Two clusters based on tumor_size with menopause.-Phase -II

In Phase – III clustering format , it is found that in cluster (C1) tumor with maximum size are grouped in ge40 and premeno category and minimum size are grouped in it40 category. In cluster (C2) we found no values. Fig.3 differentiate two clusters based on the field of tumor_size with menopause.

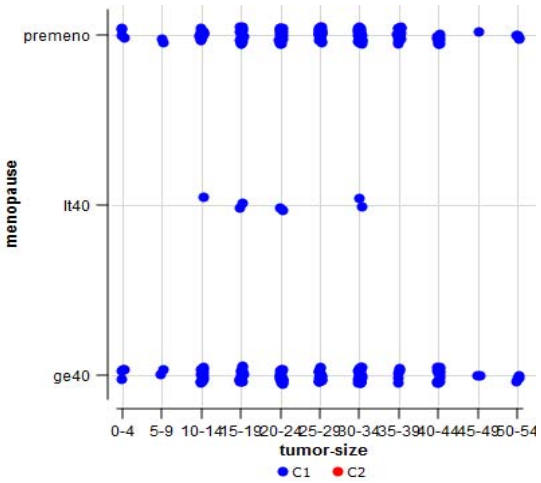

Fig.3.Two clusters based on tumor_size with menopause.-Phase –III

IV. CONCLUSION

In this paper K-means clustering algorithms is applied to study breast cancer attributes and comparison is done by applying different type of distance measure, scoring method and initialization value to identify and differentiate the average probability of correct classification, In this by applying agglomerative clustering 100% correct classification is obtained. Clustering algorithms can be combined with the decision of radiologists to improve accuracy of the cluster. In this study, it is suggested to use various probabilistic clustering algorithms in particular to study breast cancer data.

REFERENCES

[1]. Early Breast Cancer Trialists "Collaborative Group(EBCTCG), Effects of Chemotheraphy and hormonal therapy for early breast cancer on recurrence and 15-year survival: An Overview of the randomized trials*",.Lancet* 365:1687-1717,2005

[2] Hicks C, Asfour R, Pannuti A, Miele L. *"*An integrative genomics approach to biomarker discovery in breast cancer", *Cancer Inform* 2011;10:185-204

[3] Zhang B, Beeghly-Fadiel A, LongJ Wei Z. "Genetic Variants associated with breast-cancer risk: *Comprehensive research synopsis, meta-analysis, and epidemiological evidenc"e. Lancet,*2011;12:477-88

[4] Perou CM,Sorlie T, Eisen MB, et al. *"*Molecular portraits of human breast tumour*"s ,Nature*.2000;406:747-52

[5] Sorlio T. "Molecular portraits of Breast cancer: tumour subtypes as distinct disease entitie*"s. Eur J.Cancer*,2004;40:2667-75

[6] American Cancer Society. Breast Cancer Facts & Figures 2009-2010.Atlanta: American Cancer Society; 2009. Available at :http://www.cancer.org/Research/cancerFactsFigures/BreastCancerFactsFigures/f861009-final-9-08-09-pdf. Accessed April 26,2011.

[7] Ries LAG,YoungJL, KeelGE, et.al.,editors.*SEER Surivial Monograph:Cancer Survival Among Adults*;US SEER Program,1988-2001,Patient and Tumor Characteristics. Bethesda,MD:National Cancer Institute;2007. Available at:http//seer.cancer.gov/ publications/survival/seer_survival_mono_highres.pdf. Accessed September 29,2010.

[8] Sapher.T, TormeyDC, Gray R*: "*Annual hazard rates of recurrence for breast cancer after primary theraphy*".J clin Oncol* 14:2738-2746,1996.

[9] Jain AK & Dubes R.C *,Algorithms for Clustering Data* (Prentice Hall, Englewood Cliff, New Jersey) 1988

[10] Everitt B, Sabine L, Monven L & Leese M*, Cluster Analysis* (Hodder, Arnold) 2001.

[11] S.Tavazoie, D.Hughes, M.J.Campbell, R.J.Cho, G.M.Church, "Systematic determination of genetic Network architecture", *Nature Genet,* pp.281-285,1995.

[12] O'REILLY, Mitchell L Model, *Bioinformatics Programming Using Python.*