ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Detection of Cancer Using Biclustering

Sayana Sunny[1], M.Pratheba[2]

PG Scholar, Applied Electronics, SNS College of Engineering, Coimbatore, India[1]

Assistant Professor, Department of EEE, SNS College of Engineering, Coimbatore, India[2]

ABSTRACT— Cancer is one of the common diseases occurring among the people all over the world. It can be due to various reasons such as different habitats, environmental disorders etc. Cancer being detected at early stages can save millions, if effective treatment is provided. It can cause damage to any part of body. One of the means to detect cancer is Biclustering algorithm, which is not an accurate method as only one row and column can be segmented simultaneously. Clustering and biclustering methods are the primary techniques involved in analyzing gene expression data which include grouping of genes, classification of genes and classification of a sample. Apart from classical clustering methods, biclustering is being preferred to analyze biological datasets, due to its ability to group both genes across conditions simultaneously. Furthermore, the probability of the cancer is identified and its type i.e. benign, suspicious or malignant.

KEYWORDS- Biclustering, Image Processing, Benign, Malignant

I. INTRODUCTION

In computer vision, image segmentation is the process of partitioning a digital image into multiple segments such as sets of pixels, also known as superpixels. The goal of segmentation is to simplify and change the representation of an image into something that is more meaningful and easier to analyze. The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image. Each of the pixels in a region are similar with respect to some characteristic or computed property, such as color, intensity, or texture.

Image segmentation is the division of an image into regions or categories, which correspond to different objects or parts of objects. Over a period of up to 5 years, a cancer may duplicate itself up to 20 times [11]. Although there are various traditional methods for detecting cancer, an automated system was developed known as CAD (Computer Aided Detection) [12].Cancer is abnormal growth of cells which never die. Regular cells in the body follow a systematic way of growth, separation, and destruction. Programmed cell death is termed as apoptosis and when this process resolves, cancer cells are formed. Dissimilar regular cells, cancer cells do not experience programmatic death and instead continue growth and division of cells which results in a mass of abnormal cells that grows out of control. There are over 100 different types of cancer and each is classified by the type of cell that is initially affected. Cancer harms the body when damaged cells divide uncontrollably to form masses of tissue which are also called as tumors. When a tumour successfully spreads to other parts of the body and grows, invading and destroying other healthy tissues, it is said to have metastasized. This process is called metastasis and its result is a serious condition that is very crucial for treatment [11].

Breast cancer is a malignant tumor that starts in the cells of the breast. A malignant tumor is a group of cancer cells that can grow into surrounding tissues or spread to distant areas of the body. The disease occurs almost entirely in women, but men can get it, too. Despite an increased global effort to end breast cancer, it continues to be the most common cancer and the second leading cause of cancer deaths in women in the United States. In 2011, an estimated 230,480 new cases of breast cancer are expected among women in the United States. The number of victims of this disease can reach 40,000 or more each year. Thus it is very important to have more research on breast cancer and various methods to detect it [13].

Biclustering was first used by Cheng and Church [5], [6] in gene expression data analysis. It belongs to a distinct class of clustering algorithms that perform synchronous row-column clustering. Biclustering algorithms have also been proposed and used in some application fields such as co-clustering, bi- dimensional clustering, two-mode clustering and subspace clustering [3]. Biclustering is an important technique in two way data analysis. Biclustering is an extremely useful data mining tool used for identifying patterns, where different genes are correlated based on the subset of conditions in the gene expression dataset. This methodology is effectively applied to extract finer details about the behavior of genes under certain experimental samples [9]. Thus biclustering can be very well used for detecting cancer.

Mammography still remains the first step for breast cancer screening and investigation though it is less accurate in patients with dense breast tissue, implants or other factors that result in complex breast tissue. There are various testing options apart from mammography for breast cancer diagnosis and detection like Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Extreme Drug Resistance (EDR). There are also cancer detection methods available using tumors and risk markers.

Many biclustering algorithms have been used so far [2], [5]. Some of them are explained as follows:

A. Spectral Biclustering

Spectral biclustering approaches use techniques from linear algebra to identify bicluster structures in the input data. In this model, it is assumed that the expression matrix has a hidden checkerboard like structure that we try to identify using eigenvector computations. The spectral algorithm was applied to human cancer data and its results were used for classification of tumour type and identification of marker genes.

B. The SAMBA Algorithm

The SAMBA algorithm [5],[6] (Statistical-Algorithmic Method for Bicluster Analysis) uses probabilistic modeling of the data and graph theoretic techniques to identify subsets of genes that jointly respond across a subset of conditions, where a gene is termed responding in some condition if its expression level changes significantly at that condition with respect to its normal level

C. Cheng and Church's Algorithm

The algorithm constructs one bicluster at a time using a statistical criterion i.e., a low mean squared residue (the variance of the set of all elements in the bicluster, plus the mean row variance and the mean column variance). Once a bicluster is generated, its entries are filled with random numbers, and the process is repeated continuously for specific effects. After removing row, column and sub matrix averages, the residual level should be as small as possible. To discover more than one bicluster, Cheng and Church [5], [6] suggested repeated application of the biclustering algorithm on modified matrices.

In this study, a mammographic image is given as an input to the system. By using image processing the algorithm filters out the unwanted part from the mammographic image. It considers only the white part from the entire mammographic image and discards the black portion from the mammographic image, thus emphasizing the most significant portion of the image which would be useful further processing. A specific threshold value is then decided above which there are the chances for occurrence of cancer.

In order to study the various ways for detection of cancer the most important thing is getting the mammographic images of the cancerous patients. The most renowned hospital for cancer is Aswini Hospital. Mammography images of cancerous patients from Aswini hospital are collected. With the help of some of the doctors of that hospital it was very easy to collect all the required information about the breast cancer such as the various methods available to detect it, what new can be introduced in the field.

II. MATERIALS AND METHODS

Biclustering algorithms synchronously cluster both rows and columns. These types of algorithms are applied to gene expression data analysis to find a subset of genes that shows similar behavioral-pattern under a subset of conditions [3].The concept of Gene Expression Data Matrix to detect cancer. It will be working with a rxc data matrix, where each element a_{ij} will be a replaced with a real value. In the case of gene expression matrix, a_{ij} represents the expression level of gene i under condition j. Table 1 illustrate the arrangement of a gene expression matrix [3], [7].

TABLE 1. GENE EXPRESSION DATA MATRIX

	Condition 1	…	Condition j	…	Condition c
G1	a11	…	a1j	…	a1c
Gene …	…	…	…	…	…
Gi	ai1	…	Aij	…	Aic
Gene …	…	…	…	…	…
Gr	ar1	…	Arj	…	Arc

The general case of a data matrix P, with set of rows A and set of columns B that is, the data matrix is P =(A,B) where the element a_{ij} corresponds to a value representing the relation between row i and column j. Such a matrix P, with r rows and c columns, is defined by its set of rows, A= $\{a_1,\ldots,a_r\}$ and its set of columns, B= $\{b_1,\ldots,b_c\}$. Use (A, B) to denote the matrix P. Consider the data matrix P as mentioned above. Define a group of genes as a subset of genes that shows similar behavioral-pattern across the set of all conditions. Similarly, a group of conditions is a subset of conditions that shows similar behavioral-pattern across the set of all genes. A bicluster is a subset of genes that shows similar behavioral-pattern across a subset of conditions, and vice versa [3].

For detecting cancer the most important step is data collection, which includes getting the mammographic images of the tissues of cancerous patients. Certain mammographic images of cancerous patients were acquired from Aswini Hospital [4]. Biologically cancer is detected using a technique known as microarray. A microarray is typically a glass slide on to which DNA molecules are fixed in an orderly manner at specific locations termed spots. A microarray may be composed of thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely resemble to a gene. The DNA in a spot may either be genomic DNA or short stretch of oligo-nucleotide strands to be identical to a gene. It is used for image processing, transformation and normalization. [1], [6]

An attempt has been made to develop a computerized system for detection of cancer which makes use of image processing technique. Image processing involves the following steps:

a) An input mammogram image is processed and formation of bicluster in the image.
b) Determination of the tumour by the process of segmentation.

After image processing it would transform this image to an rXr matrix. This matrix helps in easy grouping of genes, classification of a new gene and classification of a new sample. Given the data matrix P, as mentioned above, it define a group of rows as a subset of rows that shows similar behavioral- pattern across the set of entire columns. Similarly, a cluster of columns is a subset of columns that exhibit similar behavioral- pattern across the set of entire rows. Now after converting the image to matrix format it considers a threshold value for determining whether it is a cancerous tissue image or not. A schematic diagram of our implementation is as follows:

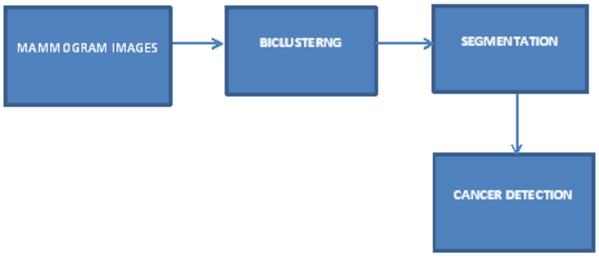

Figure1. Schematic representation of cancer detection application

A mammographic image is given as input to the system, which then is converted to matrix. Then next procedure is to preprocessing the input image. Using spectral coclustering algorithm A cluster is formed thus the segmentation is performed in the biclustering based cancer detection.

After the formation of cluster the segmentation is achieved. The segmentation process involves removal of unwanted portion of the mammogram image. Then pectoral muscles are removed in the following step. Any process there is a chance of formation of noise from the external process or internal process. To remove the available noise morphological filtering and convolution is performed. Finally centroid is calculated and the peak to signal noise ratio is calculated.

Now by considering some specific threshold value, the final result can be determined. If the result is above the threshold value, then there are chances for that person to have cancer. If it is below the threshold value, then no cancer is present (figure 1).

For executing this, software known as MATLAB has been used; this made this process very simple. We need to first store all the mammographic images in one folder in the computers drive. The input mammogram images commonly used is given in figure 2.

After performing the above procedure the result can be obtained. It could detect the presence of cancer in the selected image. In MATLAB2009 the simulation program is coded and thus simulating the code, the execution of program starts. It then eliminates the unwanted portions of the image and the pectoral regions and obtained the place where cancerous tumor may be present.

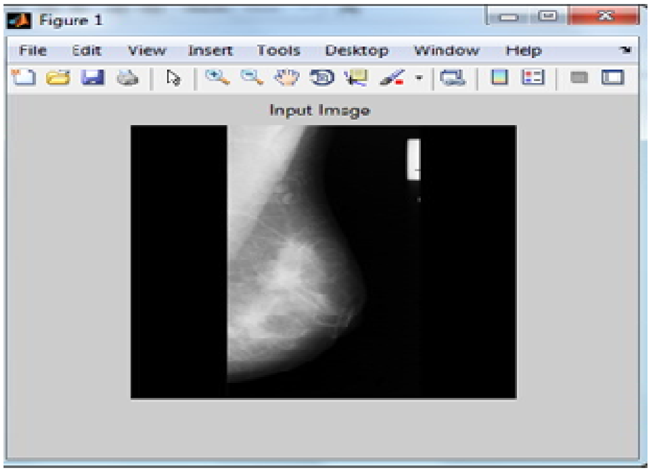

Figure 2: Screenshot after browsing the input image

III. RESULTS AND DISCUSSIONS

After studying various mammographic images, many cancerous and non-cancerous tumors were detected by carrying out the above procedure. The outputs obtained by the segmentation process are given below.

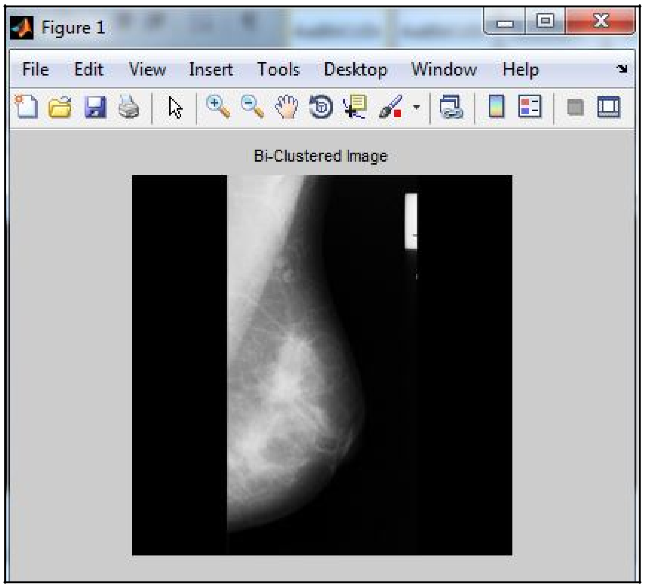

Figure3. Bicluster image formation

Fig 3 shows the biclustered image of a mammogram image. The field of interest considered is the organ image alone. The remaining portions are eliminated in the first stage of the simulation. The background is fixed and the rest of the part eliminated in the label removal procedure.

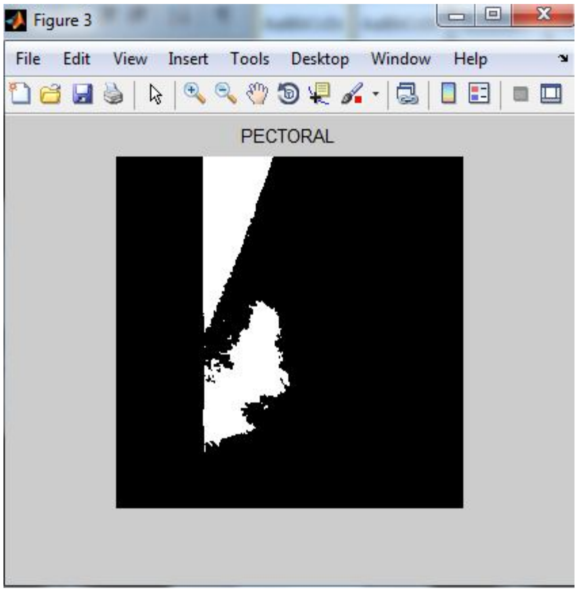

Figure 4. Screenshot of pectoral muscles

Figure 4 shows the pectoral region and the extracted pectoral muscles in the mammogram image. The muscles in the breast is known as pectoral muscles. The pectoral muscles are identified and extracted those regions in this step. Thus the simulated result is obtained.

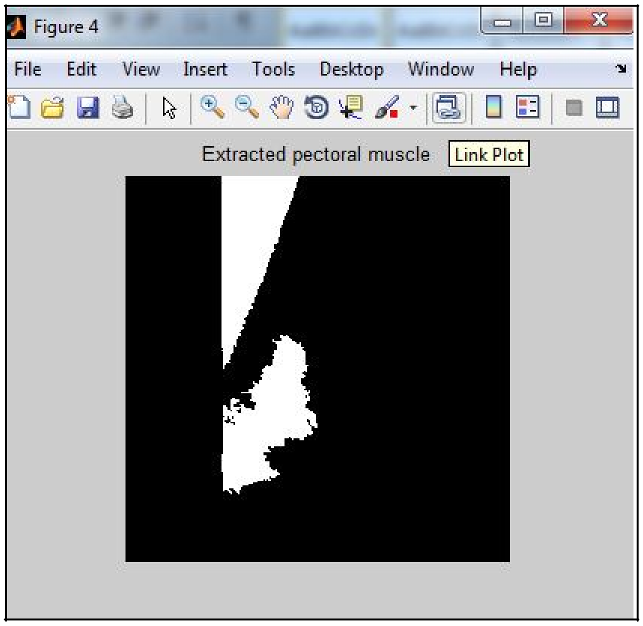

Figure 5. Screenshot of extracted pectoral muscle

The figure 5 describes the pectoral portions are removed from the original input sample mammogram image.

IV. CONCLUSION

The biclustering algorithms are used for the image segmentation. Biclustering is a relatively young area & it has a great potential to make significant contributions to biology and to other fields. Various other applications in biological data analysis, gene network identification, data mining, and collaborative filtering remain to be explored. There are various applications for biclustering. It analyzes data from different individuals suffering from different types of cancer. It contains data collected from several individuals with a particular cancer or healthy people.

V. ACKNOWLEDGMENT

This work is carried out in SNS College of Engineering, Project Lab. The authors would like to thank Aswini Hospital for their support and expertise.

REFERENCES

[1] Kenneth Bryan, Padraig Cunningham and Nadia Bolshakova, "Application of Simulated Annealing to the Biclustering of Gene Expression Data" , IEEE Transactions on Information Technology in Biomedicine vol 10,no 3, July 2006.

[2] Nishchal K Verma, Sheela Meena, Amarjot Singh, YanCui, Shruti Bajpai, "A Comparison of Biclustering Algorithms ", Proceedings of and 2010 International Conference on Systems in Medicine Biology 16-18 December 2010, IIT Kharagpur, India.

[3] Sara C. Maderia and Arlindo L. Oliveira, "Biclustering Algorithms for Biological Data Analysis: A Survey", IEEE Transactions on Computational Biology and Bioinformatics, vol 1, no 1, January-March 2004.

[4] Aswini Hospital ,Patturaickal ,Thrissur

[5] Amos Tanay, Roded Sharan and Ron Shamir, School of Computer Science, Tel-Aviv University, Ramat-Aviv, Tel-Aviv,69978 Israel, Received on January 24, 2002; revised and accepted on March 31, 2002.

[6] Amos Tanay, Roded Sharan and Ron Shamir, "Biclustering Algorithms: A Survey", May 2004.

[7] M. Madan Babu. "An Introduction to Microarray Data Analysis", chapter 11.

[8] Manjunath Aradhya, Francesco Masulli and Stefano Rovetta, "Biclustering of Microarray Data based on Modular Singular Value Decomposition", Proceedings of Computational Intelligence Methods for Bioinformatic s and Biostatistics (CIBB) 2009.

[9] Sebastion Kaiser and Friedrich Leisch, "A Toolbox for Bicluster Analysis", Technical Report No 028, 2008, University of Munich.

[10] Mehul P. Sampat, Mia K. Markey, Alan C. Bovik "Computer Aided Detection and Diagnosis in Mammography"

[11] American Cancer Society (accessed on 15 April2012) http://www.cancer.org/Cancer/BreastCancer/DetailedGuide/breast-cancer-what-is-breast-cancer

[12] Y.Cheng and G.M. Church, "Biclustering of Expression Data", Proc.Eighth Int'l Conf. Intelligent Systems for Molecular Biology (ISMB'00),pp.93-103,2000.