INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

BREAST CANCER ANALYSIS IN MEDICAL MINING BASED ON MARKOV CHAIN MONTE CARLO EXPECTATION MAXIMIZATION

[1]Ms. P.Geetha M.Sc., M.Phil**. [2]Ms. M.Subha** M.Sc., M.Phil. M.C.A., (PhD)

[1]Asst. Professor, PG & Research Dept. of Computer Science, Kaamadhenu Arts and Science College, Sathyamangalam, India. E-Mail: geetha.gsg@gmail.com
[2]Asst. Professor, PG & Research Dept. of Computer Science, Kaamadhenu Arts and Science College, Sathyamangalam, India. E-Mail: subha.gmanoharan@gmail.com

ABSTRACT: - Hyper plastic lesions of the breast include Usual Ductal Hyperplasia (UDH), a focal expansion of the number of cells in a terminal breast duct, and a typical Ductal Hyperplasia (ADH), in which a more abnormal pattern of growth is seen, and Ductal Carcinoma In Situ (DCIS) which is associated with an increased risk of developing breast cancer. The main work is easily found out the breast cancer quickly. With a Content Based Image Retrieval (CBIR) system, users will be able to retrieve relevant images based on their contents. Mont Carlo Markov Chain model EM algorithm is helpful for cell segmentation. CBIR researchers have typically followed two distinct directions based on modeling the contents of the image as a set of attributes which is produced manually by using an integrated feature -extraction / object-recognition system.

Keywords- Data Mining, Image Mining, Content Based Image Retrieval (CBIR), Feature Extraction, MCMC – EM Algorithm**.**

INTRODUCTION

Ductal Carcinoma in Situ (DCIS) is the most common type of non-invasive breast cancer. An example of a normal hyper plastic response would be the growth and multiplication of milk-secreting glandular cells in the breast as a response to pregnancy, thus preparing for future breast feeding. The objective of the system designed with the clinical management of patients in mind. Once identified as actionable, determining the true subtype of a lesion and does not change initial patient management much. Pixel data are modeled by a four-component Markov Chain Monte Carlo model. Finally we retrieve the Content-based image retrieval (CBIR) aims at finding images of interest from a large image database using the visual content of the images in the training result. The purpose is to present an image conceptually, with a set of visual features such as color, texture, and shape.

DATA MINING

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods. Data mining is the search for relationships and global patterns that exist in large database but are ‘hidden’ among the vast amount of data, such as a relationship between patient

data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. In the public sector, data mining applications initially were used as a means to detect fraud and waste, but have grown to also be used for purposes such as measuring and improving program performance [5].

IMAGE MINING

Image mining is the discovery of patterns from a collection of images. The fundamental challenge is to determine how low-level, pixel representation contained in an image or an image sequence can be effectively and efficiently processed to identify high-level spatial objects and relationships. It involves preprocessing, transformations and feature extraction, evaluation and interpretation and obtaining the final knowledge. Image mining is more than just an extension of data mining to image domain. It is an interdisciplinary endeavor that draws upon expertise in computer vision, image processing, image retrieval, data mining, machine learning, database, and artificial intelligence [3].

CONTENT BASED IMAGE RETRIEVAL SYSTEM (CBIR)

With CBIR systems, querying is facilitated through generic query classes. Examples of some query classes include color, texture, shape, attributes, and text and domain concepts. Color and texture queries allow users to formulate the description of the images to be retrieved in terms of like color and texture. Queries can also be posed with regard to the text associated with the images. In a medical setting, image retrieval is not only based on image content but also on the physician's diagnosis, treatment, etc. (i.e., additional textual data). We should also point out that CBIR differs from traditional database systems in that images are retrieved based on a degree of similarity and that records are usually retrieved from databases because of exactly matching specified attribute values [4].

FEATURE EXTRACTION

- **Cell Region Segmentation using GMM-EM**

Pixel data are modeled by a four-component Gaussian mixture model (GMM). The expectation maximization (EM) algorithm is implemented using a*, b* channels to estimate the parameters of the GMM model. The resulting mixture distribution is used to classify pixels into four categories. The remaining pixels are considered cell regions and images containing these regions are used in the next stage [9].

- **Individual Cell Segmentation**

Segmentation maps of cell regions obtained in the previous part are converted to gray level images before they are used in this stage. To identify individual cells, we used a watershed algorithm. There are several hundred connected components present in each segmented image. The gray-level intensity are computed for each connected component identified in an ROI (Release Of Information), statistical features involving the mean, standard deviation, median, and mode are computed to obtain features at the ROI level [6].

- **Classification method**

Each slide contains multiple ROI and a positive diagnosis is confirmed when at least one of the ROI in the slide is identified as positive. For a negative diagnose, the pathologist has to rule out the possibility of each and every ROI being actionable.

- **Classification on MCMC-EM**

The MCMC sampler requires data sample from continuous time Markov process, conditional on the beginning and ending states and the paths of the neighboring models. These MCMC-EM clustered features are used for training and testing the binary classifier. Based on the process we extract the large images. And this easy method to analysis the image and increase the size of the database for more extensive testing of the developed system [8].

RELATED WORKS

- Maron & Lozano - Perez, (2007) proposed a framework called Diverse Density algorithm. Since then various variants of standard single instance learning algorithms like Boosting SVM, Logistic Regression

etc. have been modified to adapt to the MIL scenario. A Computed Tomography (CT) scan is a pulmonary embolism/nodule/lesion or not [7].

- D.Wu, J. Bi, and K. Boyer (2009) proposed a min–max framework of cascaded classifier with multiple instance learning for computer aided diagnosis systems have been widely used to assist physicians in interpreting medical images from different modalities such as magnetic resonance imaging (MRI), X-ray, and computed tomography (CT) and to identify potentially diseased regions like lesions or tumors [10].
- Xin & Frank, (2006) proposed relying on the Bayesian automatic relevance determination paradigm, our learning algorithm selects the relevant subset of features that is most useful for accurate multiple instance classification. Experimental results demonstrate that the number of features chosen for optimizing the accuracy of multiple-instance classification is much smaller than that selected in a corresponding single instance learning algorithm [11].
- M. M. Dundar, S. Badve, V. Raykar, R. K. Jain, O. Sertel, and M. N. Gurcan (2010) proposed a Pathology diagnoses are made according to a set of criteria defined by the World Health Organization (WHO). While these criteria are generally easy to identify for most lesions, there are borderline cases where it becomes difficult to deter- mine with lesions [2].

METHODOLOGY

Support vector machines map input vectors to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data. The separating hyperplane is the hyperplane that maximizes the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier. General flow of the project is shown in the below figure (1).

- **Formalization**

$$\{(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \ldots, (\mathbf{x}_n, c_n)\}$$

Where the ci is either 1 or −1, a constant denoting the class to which the point belongs. Each is a p-dimensional real vector, usually of normalized (Normalizing constant) [0, 1] or [-1, 1] values. The scaling is important to guard against variables (attributes) with larger variance that might otherwise dominate the classification. SVM hyperplane form is

$$\mathbf{w} \cdot \mathbf{x} - b = 0$$

By using geometry, we find the distance between the hyperplanes is 2/|w|, so we want to minimize |w|. To exclude data points, we need to ensure that for all i either

$$\mathbf{w} \cdot \mathbf{x_i} - b \leq -1$$

$$c_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1, \quad 1 \leq i \leq n$$

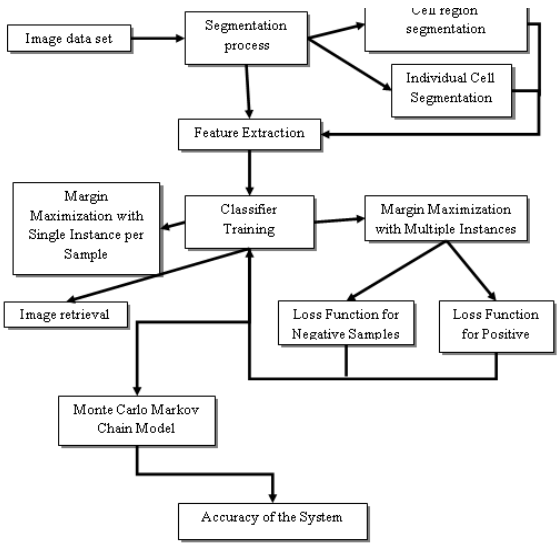

FIGURE 1: GENERAL FLOW OF THE PROJECT

- **Soft Margin**

Corinna Cortes and Vladimir Vapnik suggested a modified maximum margin idea that allows for mislabeled examples. If there exists no hyperplane that can split the "yes" and "no" examples, the Soft Margin method will choose a hyperplane that splits the examples as cleanly as possible, while still maximizing the distance to the nearest cleanly split examples. This work popularized the expression Support Vector Machine or SVM. The method introduces slack variables, ξi, which measure the degree of misclassification of the datum xi

$$c_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1 - \xi_i$$

$$1 \leq i \leq n$$

The objective function is then increased by a function which penalizes non-zero ξi, and the optimization becomes a tradeoff between a large margin, and a small error penalty. If the penalty function is linear, the equation now transforms to

$$\min \|\mathbf{w}\|^2 + C \sum_i \xi_i$$

$$\text{such that } c_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1 - \xi_i \quad 1 \leq i \leq n$$

Subject to (for any i = 1 … n)

$$y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

This constraint in along with the objective of minimizing |w| can be solved using Lagrange multipliers. The key advantage of a linear penalty function is that the slack variables vanish from the dual problem, with the constant C appearing only as an additional constraint on the Lagrange multipliers.

EXPERIMENTAL RESULTS

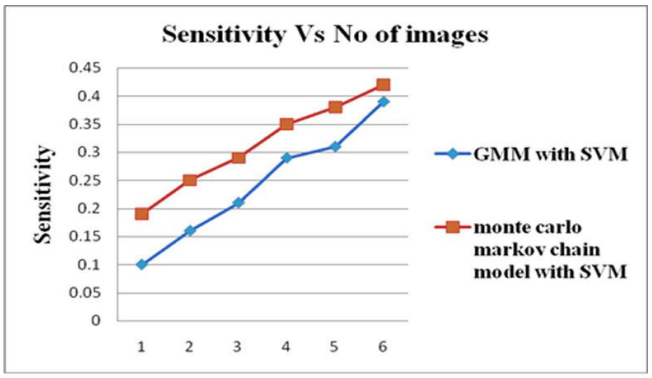

FIGURE 2: SENSITIVITY OF IMAGES

We analyze and compare the performance offered by GMM SVM method with our proposed method of MCMC SVM. Here if the number of images increased the sensitivity then it is increased linearly. Based on the comparison and the results from the experiment show the proposed approach works better than the other

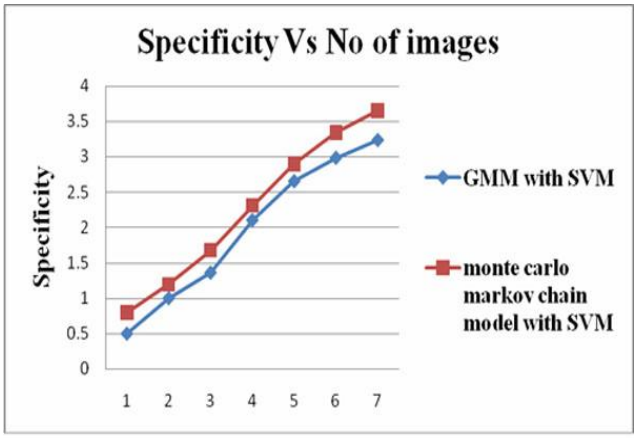

FIGURE 3: SPECIFICITY OF IMAGES

We analyze and compare the performance offered by GMM SVM method with our proposed method of MCMC SVM. Here if the number of images increased the specificity then it is increased linearly. Based on the comparison and the results from the experiment show the proposed approach works better than the other existing systems [1].

APPLICATIONS

Healthcare management

To aid healthcare management, data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims. For example, to develop better diagnosis and treatment protocols, the Arkansas Data Network looks at readmission and resource utilization and compares its data with current scientific literature to determine the best treatment options, thus using evidence to support medical care. Also, the Group Health Cooperative stratifies its patient populations by demographic characteristics and medical conditions to determine which groups use the most resources, enabling it to develop programs to help educate these populations and prevent or manage their conditions.

Group Health Cooperative has been involved in several data mining efforts to give better healthcare at lower costs. In the Seton Medical Center, data mining is used to decrease patient length-of-stay, avoid clinical complications, develop best practices, improve patient outcomes, and provides information to physicians—all to maintain and improve the quality of healthcare

SCOPE OF THE FUTURE WORK

A number of clinical trials are underway that should shed important light on the diagnosis, evaluation, and treatment of DCIS. Lumpectomy followed by radiation therapy is the most common treatment for DCIS. Research suggests that, while women treated with lumpectomy have slightly higher recurrence rates than women who undergo mastectomy, survival rates between the two groups are very similar.

For treating DCIS, a simple mastectomy — removing the breast tissue, skin, areola and nipple, and possibly the underarm lymph nodes (sentinel node biopsy) — is one option. Breast reconstruction after mastectomy, if desired, can be performed in most cases. Because lumpectomy combined with radiation is equally effective, simple mastectomy is less common than it once was for treating DCIS. Top most Data mining algorithms like Apriori algorithm, k-means algorithm, Page rank, Adaboost to implementing these algorithms in real time situations with include quality metrics for measure the performance of the heart lesions application

CONCLUSION

The proposed approach which request a number of iterative feedbacks to produce refined search results in a large scale image database and extracting the feature into the color, space, text into the pattern mining. High quality of image retrieval on RF can be achieved in a small number of feedbacks. Based on the specificity and sensitivity, the proposed system works better than the other existing systems. The system is developed with 62 cases and tested on 33 cases. An overall accuracy of 87.9% is achieved on the entire test set involving seven well-defined and 26 borderline cases. An accuracy of 84.6% is recorded on borderline cases. This was slightly higher than the average accuracy of nine board-certified pathologists (81.2%) evaluated on the same set.

REFERENCES

[1] Bernhard Schlkopf (Author), Alexander Smola, “Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)”.
[2] Dundar.M.M, Badve.S, Raykar.V, Jain.R.K, Sertel.O, and Gurcan.M.N, (2010) “A multiple instance learning approach toward optimal classification of pathology slides: A case study: Intraductal breast lesions”.
[3] D.Wu, J. Bi, and K. Boyer, “A min–max framework of cascaded classifier with multiple instance learning for computer aided diagnosis”, 2009.
[4] Gudivada.V and Raghavan.V, (1995) “Content-based image retrieval systems”, IEEE Computer, 28(9):18–22, September.
[5] Jeffrey W. Seifert, (2004) “Data Mining: An Overview, Analyst in Information Science and Technology”, Updated December.
[6] Maron & Lozano- Perez, (2007) “Diverse Density algorithm and MIL scenario”.

[7] Murat Dundar.M, Member, IEEE, Sunil Badve, Gokhan Bilgin, Vikas Raykar, Rohit Jain, Olcay Sertel, (2011) "Computerized Classification of Intraductal Breast Lesions Using Histopathological Images", Vol. 58, No. 7, July.
[8] Prieyadharsini.G, Prof.Tamije Selvy.P and Dr.Palanisamy.V, (2012) "Feature Extraction of Intraductal Breast Images Using Gmm", Info Institute of Engineering, Coimbatore, India, Vo2, No.1, February.
[9] Tavassoli.F.A and Devilee.P, (2003) "World Health Organization: Tumours of the Breast and Female Genital Organs (IARC WHO Classification of Tumors)", Cedex, France: IARCPress-WHO.
[10] Wu.D, Bi.J, and Boyer.K, (2009) "A min–max framework of cascaded classifier with multiple instance learning for computer aided diagnosis".
[11] Xin & Frank, (2006) "Relying on the Bayesian automatic relevance determination paradigm".