Classification of Microcalcification in Digital Mammogram using Stochastic Neighbor Embedding and KNN Classifier

S. Mohan Kumar
Research Scholar,
Karpagam University,
Coimbatore, Tamil Nadu, India.

G. Balakrishnan, PhD.
Director, IGCE,
Trichy, Tamil Nadu, India.

ABSTRACT

Breast cancer has become a common health problem in developed and developing countries during the last decades and also the leading cause of mortality in women each year. Mammogram is a special x-ray examination of the breast made with specific x-ray equipment that can often find tumors too small to be felt. In this paper, the classification of microcalcification in digital mammogram is achieved by using Stochastic Neighbor Embedding (SNE) for reducing high dimensionality data into relatively low dimensional data and K-Nearest Neighbor (KNN) Classifier. This system classifies the mammogram images into normal or abnormal, and the abnormal severity into benign or malignant. Mammography Image Analysis society (MIAS) database is used to evaluate the proposed system. The experiments demonstrate that the proposed method can provide better classification rate.

Keywords

Stochastic Neighbor Embedding, K-Nearest Neighbor, Digital mammograms, microcalcifications

1. INTRODUCTION

Today breast cancer is the most frequent form of cancer in women above 40. The World Health Organization's International Agency for Research on Cancer estimates that more than 150,000 women worldwide die of breast cancer each year. A computer-aided diagnosis (CAD) system for the automatic detection of clustered microcalcifications in digitized mammograms is presented in [1]. It consists of two main steps. First, potential microcalcification pixels in the mammograms are segmented out by using mixed features consisting of wavelet features and gray level statistical features, and labeled into potential individual microcalcification objects by their spatial connectivity. Second, individual microcalcifications are detected by using a set of 31 features extracted from the potential individual microcalcification objects.

A computerized scheme for detecting early-stage microcalcification clusters in mammograms is proposed [2]. It developed a novel filter bank based on the concept of the Hessian matrix for classifying nodular structures and linear structures. The mammogram images were decomposed into several sub images for second difference at scales from 1 to 4 by this filter bank. The sub images for the nodular component (NC) and the sub images for the nodular and linear component (NLC) were then obtained from analysis of the Hessian matrix.

A computer aided decision support system for an automated diagnosis and classification of breast tumor using mammogram is presented in [3]. It differentiates two breast diseases namely benign masses and malignant tumors. From the preprocessed mammogram image, texture and shape features are extracted. The optimal features can be extracted by using a feature selection scheme based on the Multi Objectives Genetic Algorithm (MOGA).

A new method of feature extraction from Wavelet coefficients for classification of digital mammograms is proposed in [4]. A matrix is constructed by putting Wavelet coefficients of each image of a building set as a row vector. It consists then on selecting by threshold, the columns which will maximize the Euclidian distances between the different class representatives. The selected columns are then used as features for classification. A novel methodology for the classification of suspicious areas in digital mammograms is presented in [5]. It is based on the fusion of clustered sub classes with various intelligent classifiers. A number of classifiers have been incorporated into the methodology and evaluated on the well known benchmark digital database of screening mammography (DDSM).

Detecting the abnormalities in mammogram by using local contrast thresholoding and rule based classification is presented in [6]. Classification of Microcalcification Using Dual-Tree Complex Wavelet Transform and Support Vector Machine is proposed in [7]. It consists of two phases. At the offline phase, training for the SVM is conducted using some training data to find the support vectors. At the online phase, a mammogram to be classified inputted into the system and then classified by the SVM.

A novel opposition-based classifier has been developed [8] which classifies breast masses into benign and malignant categories. An MLP network with a novel learning rule, called Opposite Weighted Back Propagation (OWBP), has been utilized as the classifier. The features include circularity, Zernike moments, contrast, average gray level, NRL derivatives and SP. It evaluated the classifier has been trained with both traditional BP and OWBP learning rules.

The fractal modeling of the mammographic images and their background morphology is presented in [9]. For fractal modeling, the original image is first segmented into appropriate fractal boxes followed by identifying the fractal dimension of each windowed section. Then used two dimensional box counting algorithm after which based on the order of the computations; they are placed in an appropriate matrix to facilitate the required computations. Finally using eight features identified as characteristic features of tumors extracted from mammogram images.

A novel semi-supervised k-means clustering is proposed for outlier detection in mammogram classification is proposed in [10]. The shape features are extracted from the digital mammograms, and k-means clustering is applied to cluster the features, the number of clusters is equal with the number of classes. A novel Genetic Association Rule Miner (GARM) is applied with this reduced feature set to construct the association rules for classification. The performance is analyzed with rough set using Receiver Operating Characteristic (ROC) curve analysis.

Texture analysis based on curvelet transform for the classification of mammogram tissues is presented in [11]. The most discriminative texture features of regions of interest are extracted. Then, a nearest neighbor classifier based on Euclidian distance is constructed. The obtained results calculated using 5-fold cross validation. The approach consists of two steps, detecting the abnormalities and then classifies the abnormalities into benign and malignant tumors.

A new classification approach using Support Vector Machines (SVM) for detection of microcalcification clusters in digital mammograms is proposed in [12]. Classifying data is a common task in machine learning. The MC (Microcalcification) detection is formulated as a supervised learning problem and apply SVM as a classifier to determine at each pixel location in the mammogram if the MC is present or not.

In this paper, an automatic classification of microcalcification in digital mammograms based on SNE and KNN classifier is presented. The remainder of this paper is organized as follows: The methodologies and proposed method used for the proposed system is described in sections 3 and 4. The experimental results are given in section 5.

2. METHODOLOGY

The proposed system for the classification of microcalcification in digital mammograms is built based on SNE and by applying KNN for building the classifiers. In this following section the theoretical background of all the approaches are introduced.

2.1 Stochastic Neighbor Embedding

SNE is a probabilistic approach to the task of placing objects, described by high-dimensional vectors or by pair-wise dissimilarities, in a low-dimensional space in a way that preserves neighbor identities. A Gaussian is centered on each object in the high-dimensional space and the densities under this Gaussian (or the given dissimilarities) are used to define a probability distribution over all the potential neighbors of the object. The aim of the embedding is to approximate this distribution as well as possible when the same operation is performed on the low-dimensional "images" of the objects. A natural cost function is a sum of Kullback-Leibler divergences, one per object, which leads to a simple gradient for adjusting the positions of the low-dimensional images.

For each object, i and each potential neighbor,j the asymmetric probability is calculated by the formula (1) that i would pick j as its neighbor is given by

$$pij = \frac{\exp\left(-d_{ij}^{2}\right)}{\sum_{k\neq i} exp\left(-d_{ik}^{2}\right)} \qquad (1)$$

The dissimilarities, d_{ij}^{2} , may be given as part of the problem definition (and need not be symmetric), or they may be computed using the scaled squared Euclidean distance ("affinity") between two high-dimensional points,X_i; X_j

$$d_{ij}^{2} = \frac{\left\|X_i - X_j\right\|^2}{2\sigma_i^2} \qquad (2)$$

where σ_i is either set by hand or found by a binary search for the value of σ_i that makes the entropy of the distribution over neighbors equal to $\log k$. Here, k is the effective number of local neighbors or "perplexity" and is chosen by hand. In the low-dimensional space, the Gaussian neighborhoods are used with a fixed variance so the *induced* probability qij that point i picks point j as its neighbor is a function of the low-dimensional *images* y_i of all the objects and is given by the expression

$$qij = \frac{exp\left(-\left\|y_i - y_j\right\|^2\right)}{\sum_{k\neq i} exp\left(-\left\|y_i - y_j\right\|^2\right)} \qquad (3)$$

The aim of the embedding is to match these two distributions as well as possible. This is achieved by minimizing a cost function which is a sum of Kullback-Leibler divergences between the original (pij) and induced (qij) distributions over neighbors for each object is given by

$$C = \sum_i \sum_j pij \log \frac{pij}{qij} = \qquad \sum_i KL(P_i||Q_i) \qquad (4)$$

The minimization of the cost function in Equation 4 is performed using gradient method. The gradient has the simple form as

$$\frac{\partial C}{\partial Y_i} = 2\sum_j (y_i - y_j)\left(p_{ij} - q_{ij} + p_{ij} - q_{ij}\right) \qquad (5)$$

The gradient descent is initialized by sampling map points randomly from an isotropic Gaussian with small variance that is center around the origin. For speed up the optimization and avoid been stuck in local optima, a momentum term is added to the gradient [4]. The current gradient is added to an exponentially decay sum of previous gradients in order to determine the changes in the coordinates of the map points at each iteration of gradient search. Mathematically, the gradient with a momentum term is given by [4]

$$y^{(t)} = y^{(t-1)}\eta\frac{\partial J}{\partial y_i} + \alpha(t)\left(y^{(t-1)} - y^{(t-2)}\right) \qquad (6)$$

Where $y^{(t)}$ indicate the solution at iteration$t$, η indicates the learning rate, and $\alpha(t)$ represents the momentum at iteration t. In the early stages of the optimization, after the each iteration, a random jitter is added to the map points. Then gradually reducing the variance of this noise performs a type of simulated annealing that helps the optimization to escape local minima in the cost function.

2.2 K-NN Classifier

The k-nearest neighbor algorithm (K-NN) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance-based learning where the function is only approximated locally and all computation is deferred until classification. In K-NN, an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor. The neighbors are taken from a set of objects for which the correct classification is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

3. PROPOSED METHOD

The proposed system for the classification of microcalcification in digital mammograms mainly consists of two different stages which include the feature extraction stage and classification stage. All the stages are explained in detail in the following sub sections.

3.1 Feature Extraction Stage

Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which over fits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. Figure 1 shows the block diagram of feature extraction stage of the proposed system based on SNE.

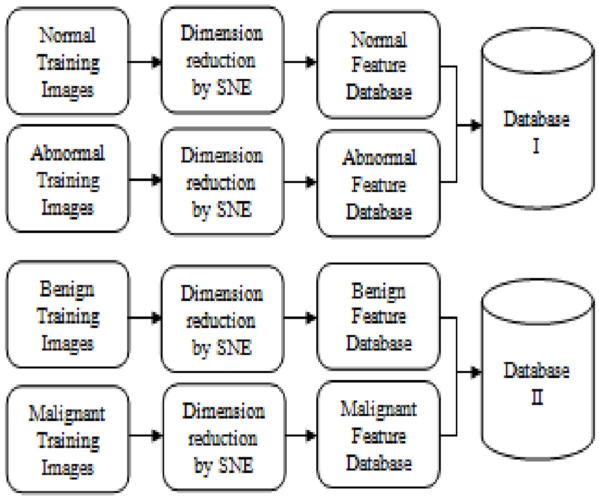
Normal Training Images
Dimension reduction by SNE
Normal Feature Database
Abnormal Training Images
Dimension reduction by SNE
Abnormal Feature Database
Database I
Benign Training Images
Dimension reduction by SNE
Benign Feature Database
Malignant Training Images
Dimension reduction by SNE
Malignant Feature Database
Database II

Fig 1: Block diagram of the feature extraction stage of the proposed system

The well known microcalcification area in the MIAS mammogram images are given to the feature extraction stage. The known microcalcification area which was given by the MIAS database is separated from the whole image. The size of the extracted ROI is 256 x 256. This high dimensional data is reduced into a relatively low dimensional data by using SNE and this reduced data set is stored in the database as feature. Database-I is constructed by using the training images of normal and abnormal images and used in the initial stage classifier. Database-II is constructed by using the training images of benign and malignant images and used in the final stage classifier.

3.2 Classification Stage

Classification phase executes two phases. In the first one, the classifier is applied to classify mammograms into normal and abnormal cases. Then the mammogram is considered abnormal if it contains tumor (microcalcification).

Finally, the abnormal mammogram is classified into malignant or benign in the final stage. In this classification stage, KNN classifier in every phase is trained at specific number of training set in each category. The block diagram of the classification stage of the proposed system based on KNN classifier is shown in Figure 2.

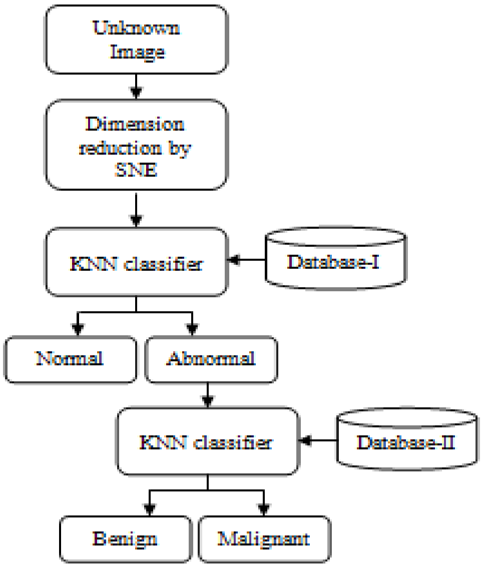

Fig 2: Block diagram of the classification stage of the proposed system

3.2.1 Initial Stage Classifier

In the initial stage classifier, the given unknown ROI from the digital mammogram image is tested for normal or abnormal category. The given high dimensional unknown ROI image is reduced into a relatively low dimensional dataset by using the SNE. This reduced dataset is initially tested with the trained KNN classifier which uses DATABASE-I. Table 1 shows the number of training and testing images used for the initial stage classifier.

Table 1. Number of training set and testing set for initial stage classifier

Type of image	No of training Images	No of Testing Images
Normal	66	99
Abnormal	17	25

3.2.2 Final Stage Classifier

In the final stage classifier, the abnormal ROI image from the initial stage classifier is further classified into Benign or Malignant. The reduced dataset of unknown ROI image is again tested with the trained KNN classifier which uses DATABASE-II. Table 2 shows the number of training and testing images used for the final stage classifier.

Table 2. Number of training set and testing set for final stage classifier

Type of image	No of training set	No of Testing set
Benign	8	12
Malignant	9	13

4. EXPERIMENTAL RESULTS

To assess the performance of the proposed system, many computer simulations and experiments with MIAS database images were performed. The performance of the proposed system is carried on 99 normal images and 25 microcalcification images. Among the 25 abnormal images, there are 12 benign and 13 malignant images available. All the images are considered for the classification test. The classification rate obtained using the SNE data sets are show in Table 3. From the table 3, it is clearly found that all the normal images and malignant images are classified with no error while the abnormal and benign category, over 80 % classification result is achieved. The result obtained from the proposed method is shown in figure3.

Table 3: Classification results of proposed method based on SNE

Mammogram Type	Classification Rate (%)
Normal	100
Abnormal	84
Benign	83.33
Malignant	100

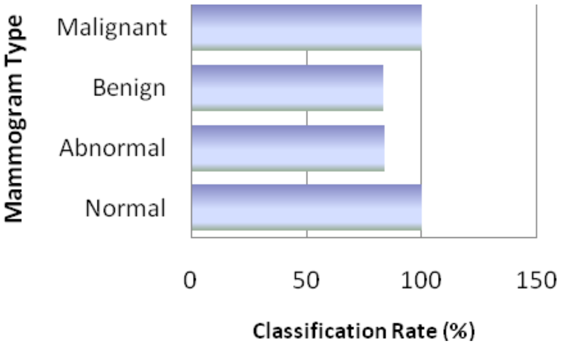

Fig3: Result obtained from the proposed method

5. CONCLUSION

In this paper, the classification of microcalcification in digital mammogram based on SNE and KNN classifier is proposed. The high dimensional data from the ROI image is relatively reduced into low dimensional data set by using the SNE and the reduced data set is used as features to classify the given mammogram images into normal or abnormal as well as benign or malignant. The proposed classification scheme is carried on MIAS database image.

Experimental results show that the proposed system achieves 100% classification rate for normal and malignant cases and over 80% classification rate for benign and abnormal cases. Still, the work is going on to get the better result for abnormal and benign cases.

6. REFERENCES

[1]. Songyang Yu and Ling Guan, "A CAD System for the Automatic Detection of Clustered Microcalcifications in Digitized Mammogram Films", IEEE Transactions on Medical imaging, vol. 19, no. 2, February 2000, pp 115-126.

[2]. Ryohei Nakayama and Yoshikazu Uchiyama, "Computer-Aided Diagnosis Scheme Using a Filter Bank for Detection of Microcalcification Clusters in Mammograms", IEEE Transactions on Biomedical engineering, vol. 53, no. 2, February 2006, pp 273-283.

[3]. M.Suganthi and M.Madheswaran, "Mammogram Tumor Classification using Multimodal Features and Genetic Algorithm", IEEE International Conference on "Control, Automation, Communication and Energy conservation, June 2009, pp 1-6.

[4]. Ibrahima Faye and Brahim Belhaouari Samir, "Digital Mammograms Classification Using a Wavelet Based Feature Extraction Method", IEEE conference on Computer and Electrical Engineering, 2009, pp 318-322.

[5]. Peter Mc Leod and Brijesh Verma, "A Classifier with Clustered Sub Classes for the Classification of Suspicious Areas in Digital Mammograms", IEEE conference on Neural Networks, July 2010, pp 1-8.

[6]. Viet Dzung Nguyen, Thu Van Nguyen and Tien Dzung Nguyen, "Detect Abnormalities in Mammograms by Local Contrast Thresholding and Rule-based Classification", IEEE third International Conference on Communications and Electronics, August 2010, pp 207-210.

[7]. Andy Tirtajaya and Diaz D. Santika, "Classification of Microcalcification Using Dual-Tree Complex Wavelet Transform and Support Vector Machine", IEEE International Conference on Advances in Computing, Control and Telecommunication Technologies, December 2010, pp 164-166.

[8]. Fatemeh Saki and Amir Tahmasbi, "A Novel Opposition-based Classifier for Mass Diagnosis in Mammography Images", IEEE Iranian Conference of Biomedical Engineering, November 2010, pp 1-4.

[9]. Alireza Shirazi Noodeh and Hossein Rabbani, "Detection of Cancerous Zones in Mammograms using Fractal Modeling and Classification by Probabilistic Neural Network" IEEE Iranian Conference of Biomedical Engineering, November 2010, pp 1-4.

[10]. K. Thangavel and A. Kaja Mohideen, "Semi-Supervised K-Means Clustering for Outlier Detection in Mammogram Classification", IEEE Trendz in Information Sciences & Computing, December 2010, pp 68-72.

[11]. Mohamed Meselhy Eltoukhy and Ibrahima Faye, "Curvelet Based Feature Extraction Method for Breast Cancer Diagnosis in Digital Mammogram", IEEE International Conference on Intelligent and Advanced Systems, June 2010, pp 1-5.

[12]. Dheeba.J and Tamil Selvi.S, "Classification of Malignant and Benign Microcalcification Using SVM Classifier",

IEEE International Conference on Emerging Trends in Electrical and Computer Technology, March 2011, pp 686-690.

AUTHOR'S PROFILE

Dr. G. Balakrishnan, M.E., Ph.D., is the Director, Indra Ganesan College of Engineering. He has completed his B.E.(CSE) from Bharathidasan University, Trichy, M.E., (CSE) from PSG College of Technology, Coimbatore and Ph.D., (Image Processing) from University of Malaysia Sabah, Malaysia. He has more than 10 years of Academic and Industrial experience. He has more than 40 publications in various International Journals and Conferences. His Ph.D. research is based on the development of Navigation Aid for the Betterment of Visually Impaired. He is a recognized supervisor for guiding Ph.D. students' under Trichy Anna University, Bharathiyar University, Coimbatore, Mother Theresa University, Kodaikanal. He is the Advisory Council member for several International and National conferences. He has won silver medals for his research contribution in various National and International Research competitions. He was awarded 'The Best Outgoing Researcher Award 'during 2006 by Malaysia University.

Mr. S.Mohan Kumar is a Research scholar of Karpagam University, Coimbatore. He is a Microsoft Certified Professional in MS-SQL Server. He has served as a Software Engineer, Technical Lead (Online Resources) in reputed Software industries in Bangalore and also acted as HOD in well known Engineering Colleges and also in a Deemed University. He is a Life member of Indian Society for Technical Education, Advanced Computing and Communication Society, System Society of India, International Association of Computer Science & Information Technology and Energy Conservation Society. He has presented around 20 Technical Papers in International /National/ State Level Conferences / Seminars and Symposiums. His major area of research work is 'Medical Image Processing'.