Mammographic mass detection based on robust learning algorithms

Cao, Aize

2005

Cao, A. Z. (2005). Mammographic mass detection based on robust learning algorithms.
Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/4879>

<https://doi.org/10.32657/10356/4879>

Nanyang Technological University

Downloaded on 26 Apr 2023 19:59:16 SGT

Mammographic Mass Detection Based on Robust Learning Algorithms

Cao Aize

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University

in fulfillment of the requirements for the degree of

Doctor of Philosophy

2005

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

July 10, 2005

.....

Date

Aize Cao

.....

Cao Aize

*To my family,
for their encouragement and love.*

Acknowledgements

First and foremost, I would like to express my sincere gratitude and appreciation to my supervisor, Associate Professor Song Qing, for his intelligent guidance and patient support over the past years. His theoretical and analytic insight into the problems helped greatly in the accomplishment of this thesis. Also, he has given me precious methodologies that are valuable to my real life. Without his instructions and help, this thesis will not be possible.

I would like to thank all my colleagues at the Intelligent Machines Research Laboratory, for their support and help towards the completion of my study. Particularly, I want to say thanks to Dr Hu Wenjie, Dr Wang Lei, Mr Yang Xulei, and Mr Wang Zhimin, for their valuable discussion on some work and sharing with me their ideas during the progress of this research. Thanks are due to all my friends, no matter where they are, for their help and encouragement.

I would like to acknowledge the School of EEE, Nanyang Technological University, for providing the financial support, research facilities and opportunity for this study.

Finally, I would like dedicate the thesis to my husband, my parents, mother-in-law, my sister, and my lovely son for their understanding and constant encouragement. Words could not express my appreciation to my dearest husband, Xin Fuxue, for his love, encouragement, and continuous support during all these three years. Without all these, it would be not possible for me to complete this work.

Summary

Automatic processing of medical images is a promising research topic that has attracted much attention. The automatic processing of mammograms, which plays an important role in the screening and diagnosis of breast cancer, is investigated in this thesis. The ability of automatically processing a large number of digital mammograms can relieve the radiologists from the burden of time-consuming work. This thesis focuses on developing machine learning algorithms of image segmentation and pattern recognition as a means for breast mass detection in digitized mammograms.

Image segmentation is a challenging technique for breast cancer detection due to the variability of normal breast tissues. Clustering is one of the important methods used in the literature for the unsupervised medical image segmentation, in particular, the soft clustering algorithms have certain advantages compared to the conventional hard clustering algorithms like k-means for the soft segmentation. Among the soft clustering methods, deterministic annealing (DA) has a unique character of global optimization by maximizing Shannon entropy to avoid local optima, a problem encountered by many fuzzy c-means (FCM) related algorithms. The dissimilarity measure in the standard DA is a function of the distance between the intensity-based feature vector and the centroids without the knowledge of the shape information of the target object, which can not lead to satisfactory segmentation for mammographic images with weak contrast between the mass feature and the ones of normal breast tissue. The dissimilarity measure solely using the intensity

information may create patches wherever the intensities of normal breast tissue is close to the intensities of masses. The scattered patches lead to undesirable data partitions. To improve the robustness of the DA methods in mammographic mass segmentation, a new dissimilarity measure with the inclusion of a circular shape function for DA clustering (DACF) is first proposed for circumscribed mass segmentation. With the inclusion of the shape information, the differentiation between the mass region and normal breast tissue can be enhanced.

In addition to the well-defined ones, masses with irregular, spiculated margins, architectural distortion and asymmetry in shapes have high tendency to be malignant. The boundaries of these masses are irregular and fuzzy, which are often embedded in the parenchymal tissues. Moreover, the digitized mammographic images are polluted in terms of noises and artifacts which can lead the segmentation results deviated from the optimal grouping. Once the outliers (noisy points or regions) are identified, mass segmentation can be improved by rejecting the noisy data. Historically, the performance of classical implementation of fuzzy clustering methods is limited because they are subject to initialization and noise. In cope with the outlier estimation, a robust information clustering (RIC) algorithm based on minimax optimization of mutual information is proposed for mass segmentation in an automatic detection scheme. For the minimax criteria in RIC algorithm, the minimization of mutual information of inter-clusters leads to the deterministic clustering. The maximization of the mutual information provides a robust density estimation of probability soft margin to phase out outliers. The concern of RIC in the automatic mass detection system is to find the locations of suspicious regions that may contain mass in mammograms with relatively high detection rate and small number of false positive (FP).

Besides mammography screening, mass pattern analysis is another important aspect for breast cancer detection. The thesis also considers the case when the mass

regions are extracted from mammograms manually instead of automatic methods. The ROIs will be classified into normal or abnormal, benign or malignant cases by a semi-automatic detection system where a new classifier of vicinal support vector machine (VSVM) is investigated for the malignant cases study. It is known that the SVM is a promising classification technique developed by Vapnik and his colleagues. One of the main assumption of the standard SVM is that all samples in the training set are independent and identically distributed (i.i.d). However, in many practical engineering applications, the distribution of the obtained training samples are not i.i.d and often subject to different vicinities. VSVM has the property of dealing with data with unknown density distribution that are smooth and symmetric in the vicinities of any point. It was first proposed by Vapnik, but has the limitation of the derived vicinities in the input space, and the partition ability of the vicinities is rather limited. The thesis extends the concept of the deterministic annealing clustering method proposed by K.Rose in input space to feature space, as a general clustering algorithm to partition the whole training samples into different soft vicinal areas for the application of VSVM. VSVM proposed here acts as an enhancement learning algorithm for the purpose of medical diagnosis to classify the mammographic masses into benign and malignant.

In summary, a few novel and robust learning algorithms are proposed and investigated in this thesis for the approaches of fully and semi-automatic detection of breast masses in digitized mammograms. The simulation results show their better performances compared to other existing algorithms for the automatic processing of mammograms and they are hopeful to facilitate the clinical decisions.

Table of Contents

Acknowledgements	i
Summary	ii
List of Tables	x
List of Figures	xii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Objectives	4
1.3 Major Contributions of the Thesis	6
1.4 Organization of the Thesis	8
2 Medical Background and Literature Review	10
2.1 Introduction	10
2.2 Breast Cancer and Mammographic Images	11

TABLE OF CONTENTS

vi

2.2.1	Breast Structure	11
2.2.2	Know About Breast Cancer	12
2.2.3	Breast Imaging Modalities	14
2.2.4	Mammographic Images	16
2.3	Review of Related Projects	17
2.3.1	Mammographic Image Processing	20
2.3.2	False Positive Reduction	23
2.3.3	Mammographic Mass Diagnosis	24
2.4	Review of Clustering Algorithms for Medical Image Segmentation . .	27
2.4.1	Fuzzy C-Means Clustering	30
2.4.2	Robust Fuzzy C-Means Clustering	32
2.4.3	Deterministic Annealing Clustering	34
2.5	Review of Statistical Learning Theory for Pattern Recognition . . .	35
2.5.1	Empirical Risk Minimization	37
2.5.2	Structural Risk Minimization	37
2.5.3	Support Vector Machine	38
2.6	Concluding Remarks	42
3	Deterministic Annealing Clustering Incorporating Circular Func-	
	tion for Mass Segmentation	43
3.1	Introduction	43

TABLE OF CONTENTS

vii

3.2	A Brief Review of the Mass Constrained Deterministic Annealing . . .	45
3.3	Deterministic Annealing Clustering Incorporating Circular Function (DACF)	48
3.4	Experimental Results	53
3.4.1	Segmentation Results	54
3.4.2	Weighting Parameter Analysis	57
3.4.3	Convergence Time Analysis	58
3.5	Conclusion	59
4	Robust Information Clustering for Automatic Mass Detection	68
4.1	Introduction	68
4.2	Gain Adaptive Contrast Enhancement	72
4.3	Raw ROI Extraction	75
4.4	The Proposed Robust Information Clustering (RIC) Algorithm	77
4.4.1	Basics of Information Theory	77
4.4.2	Robust Information Clustering Algorithm	80
4.5	Experimental Results	84
4.5.1	The Performance of RIC for Outliers Estimation and Mam- mographic Image Segmentation	84
4.5.2	FP Reduction	86
4.6	Detection Results	87
4.7	Discussion	89

TABLE OF CONTENTS

viii

4.8	Conclusion	91
5	Vicinal Support Vector Machine for Mass Pattern Analysis	97
5.1	Introduction	97
5.2	One-Class Support Vector Machine	100
5.3	The Proposed Vicinal Support Vector Machine (VSVM)	103
5.3.1	VSVM with Input Space Partition	103
5.3.2	VSVM with Feature Space Partition	105
5.3.3	Kernel Based Deterministic Annealing for VSVM	108
5.4	Data Acquisition	111
5.4.1	Region of Interest	111
5.4.2	Feature Extraction	112
5.5	Experimental Results	115
5.5.1	Abnormal Cases Detection by One-Class SVM	115
5.5.2	Malignant Cases Detection by VSVM	116
5.6	Conclusion	119
6	Conclusion and Recommendations	120
6.1	Conclusion	120
6.1.1	Segmentation of Circumscribed Mass	121
6.1.2	Detection of Masses Embedded in Glandular or Dense Glandular Breast Tissue	122

TABLE OF CONTENTS

ix

6.1.3	Mass Pattern Analysis	123
6.2	Recommendations	124
6.2.1	Incorporation of the Spatial Information	124
6.2.2	Multi-modality Images Registration	125
6.2.3	Multi-layer Agent Detection System	125
6.2.4	Database Setup	126
Author's Publications		127
Bibliography		129

List of Tables

3.1	The probability of segmentation error (PSE) by DACF, DA and FCM.	56
3.2	Average segmentation error (PSE) of DACF with different weighting parameter β_1 and β_2	58
3.3	The convergence time of DACF and FCM for the segmentation results in Fig.3.3 and Fig.3.5. The number of cluster is two for DACF, and two to six for FCM.	59
3.4	The convergence time (second) of DACF for mdb002 mass segmentation with different weighting parameter β_1 and β_2	65
3.5	The convergence time (second) of DACF for mdb023 mass segmentation with different weighting parameter β_1 and β_2	65
3.6	The convergence time (second) of DACF for mdb025 mass segmentation with different weighting parameter β_1 and β_2	65
3.7	The convergence time (second) of DACF for mdb315 mass segmentation with different weighting parameter β_1 and β_2	65
4.1	The detection results of the number of suspicious regions per image.	88
5.1	The classification results by SVM and VSVM.	117

LIST OF TABLES

xi

6.1 Performance of the proposed algorithms. 121

List of Figures

2.1	The anatomy of the human breast [105].	12
2.2	Noninvasive breast cancer. A. ducts, B. lobules, C. dilated section of duct to hold milk, D. nipple, E. fat, F. pectoralis major muscle, G. chest wall/rib cage. (a) Normal breast anatomy, enlargement: A. normal duct cells, B. basement membrane, C. lumen (center of duct). (b) Ductal Carcinoma in situ (DCIS), enlargement: B. ductal cancer cells. (c) Lobular Carcinoma in situ (LCIS), enlargement: B. lobular cancer cells.	13
2.3	Invasive breast cancer. (a) Invasive ductal carcinoma (IDC), enlargement: B. ductal cancer cells breaking through the basement membrane. (b) Invasive lobular carcinoma (ILC), enlargement: B. lobular cancer cells breaking through the basement membrane. (c) Vascular and lymphatic invasion, enlargement: A. normal duct cells, B. cancer cells, C. basement membrane, D. lymphatic channel, E. blood vessel, F breast tissue.	14
2.4	Two basic views of mammography. (a) Sketch map of the two views. (b) Mammogram with MLO view. (c) Mammogram with CC view. .	17

2.5	The basic abnormalities on mammograms. (a) Microcalcification. (b) Benign mass with well-defined border. (c) Spiculated mass with star-shaped border.	18
2.6	Diagram of bottom-up control strategy for breast mass detection in digital mammograms.	20
2.7	Illustration of clustering methods in image segmentation. Pixel intensities in an image are randomly distributed. Suppose the intensity-based features belong to three groups or clusters, represented by 1,2,3 in the figure. It shows that the pixels belong to the same groups or clusters may not locate at the same region, i.e. they can be partitioned into unconnected in different subregions.	29
2.8	Relationship between ERM and SRM	39
2.9	a. A separating hyperplane with small margin. b. A separating hyperplane with larger margin.	40
3.1	(a) One-dimensional input density of the data samples, the sample data are generated as 6 Gaussian distribution with the identical weights and identical variances, the density is a linear superposition of these 6 Gaussian distributions. (b) Computed cluster splitting and phase transition behavior as temperature T decreasing. Note that the six cluster centers are recovered gradually as temperature goes down.	48
3.2	ROIs that contain masses with circumscribed shape in MiniMIAS database. The two ROIs in mdb132 are processed independently. . . .	61
3.3	Segmentation results by DACF for all the ROIs in Fig.3.2.	62
3.4	Segmentation results by DA for all the ROIs in Fig.3.2.	63

3.5	Segmentation results by FCM for all the ROIs in Fig.3.2.	64
3.6	Segmentation results by DACF with different parameter values of β_1 and β_2 for the case of mdb002. (a) $\beta_1 = 1, \beta_2 = 1$. (b) $\beta_1 = \frac{1}{2}, \beta_2 = 5$. (c) $\beta_1 = \frac{1}{4}, \beta_2 = 10$. (d) $\beta_1 = \frac{1}{8}, \beta_2 = 20$	66
3.7	Segmentation results by DACF with different parameter values of β_1 and β_2 for the case of mdb023. (a) $\beta_1 = 1, \beta_2 = 1$. (b) $\beta_1 = \frac{1}{2}, \beta_2 = 5$. (c) $\beta_1 = \frac{1}{4}, \beta_2 = 10$. (d) $\beta_1 = \frac{1}{8}, \beta_2 = 20$	66
3.8	Segmentation results by DACF with different parameter values of β_1 and β_2 for the case of mdb025. (a) $\beta_1 = 1, \beta_2 = 1$. (b) $\beta_1 = \frac{1}{2}, \beta_2 = 5$. (c) $\beta_1 = \frac{1}{4}, \beta_2 = 10$. (d) $\beta_1 = \frac{1}{8}, \beta_2 = 20$	67
3.9	Segmentation results by DACF with different parameter values of β_1 and β_2 for the case of mdb315. (a) $\beta_1 = 1, \beta_2 = 1$. (b) $\beta_1 = \frac{1}{2}, \beta_2 = 5$. (c) $\beta_1 = \frac{1}{4}, \beta_2 = 10$. (d) $\beta_1 = \frac{1}{8}, \beta_2 = 20$	67
4.1	Flowchart of the automatic mass detection system for digitized mammograms.	71
4.2	The block diagram of the GACE preprocessing procedure, where $I_{CO}(x, y)$ is the output of GACE.	72
4.3	Plot of the gain value function of GACE filter. It is a positive quadratic function, where the slop is increasing.	73
4.4	Original and the enhanced images by GACE. a) Original image. b) Enhanced image by GACE.	74
4.5	Histograms of the original and enhanced images. a) Histogram of original image. b) Histogram of the image enhanced by GACE. . . .	74

4.6	Illustration of components in a mammogram. There are two major areas in a mammogram, breast region and non-breast region. The breast region contains pectoral muscle, breast tissues or mass. The non-breast region contains dark background. For some mammograms there are some background objects with high intensities, as indicated in the figure.	76
4.7	Raw ROI extraction. It is the case of mdb111 in MiniMIAS that contains an asymmetry malignant mass surrounded by dense glandular breast tissues. (a) Original image. The circled region contains mass. In addition to the breast region, there are two more other objects in the background. (b) GACE enhancement. (c) Breast region, obtained by first step of thresholding. (d) Raw ROI, extracted by second step of adaptive thresholding.	92
4.8	Clustering performance of RIC and DA. (a). The distribution of original data. It includes two clusters with overlapped data. Different symbols refer data to different clusters. (b). Clustering result of RIC. The symbols 'o' represent data that belong to outlier cluster. (c). Clustering result of DA.	93
4.9	The performance of RIC on outliers identification and image segmentation for mammographic image data. The cluster number is two. (a) Original ROI that contains a case of mass. (b) Outliers (dark points) that are estimated by RIC. (c) Segmentation result by RIC, where dark points represent outliers. (d). Histogram of the original ROI. (e) Histogram of the original image with outliers removed. (f) Segmentation result by DA, cluster number is two.	93
4.10	FROC analysis for the detection systems based on RIC, DA and FCM.	94

4.11	Detection results for mdb181. (a) Original image. The circled region contains mass. (b) Detection result of RIC.	94
4.12	Detection results for mdb198. (a) Original image. The circled region contains mass. (b) Detection result of RIC.	95
4.13	Masses that are missed. (a) The mass region is removed together with pectoral muscle. (b)-(d) The masses are missed during suspicious cluster selection.	96
5.1	Mammograms that contain normal and abnormal cases. The abnormal cases in (b) and (c) are illustrated in small windows. (a) Mammogram with normal case. (b) Mammogram with benign mass. (c) Mammogram with malignant mass.	98
5.2	Data distribution in the feature space. The hypersphere containing most of the data in one class described by center a and radius R . Data on the boundary are support vectors, data that are outside boundary have $\xi_i > 0$	101
5.3	Selected ROI. (a) Original mammographic image from MIAS. (b) Region of interest, the inner square area is the bounding box of mass region.	112
5.4	The performance of one-class SVM for abnormal cases analysis versus Gaussian RBF kernel width σ	116
5.5	ROC curve for VSVM and standard SVM for malignant masses classification. The performance of VSVM with cluster number $K=2, 5, 10$ are present.	118

Chapter 1

Introduction

1.1 Background and Motivation

With the development of image processing techniques, medical image processing has become an invaluable tool for diagnostic imaging today. Different modalities of medical image, such as magnetic resonance imaging (MRI), computed tomography (CT), digital mammography, provide effective means of noninvasive diagnosis of diseases. The increasing size and large number of medical images make the techniques of automatic image processing and diagnosis necessary. In particular, computer based algorithms for the delineation of special regions of interest(ROI) are a key point in assisting radiologists. Mammography is a very common and useful image modality in diagnosis of breast cancer [75]. There are several types of breast abnormalities that are visible in mammograms, namely, clustered microcalcifications, mass lesions, and architectural distortions. Mass and clustered microcalcification are two major signs of early breast cancer that can be detected in mammograms before they are palpated. The detection of masses in digitized mammograms is specially investigated in this thesis.

Breast cancer is a major public health problem in many countries, with incidence and mortality high in western nations. It has been a leading cause of fatality among all cancers for women, especially women in USA and Europe, nearly one in ten women may develop breast cancer during their lifetime [63]. The American Cancer Society estimated about 43,900 deaths and 178,700 new cases of invasive breast cancer among women in the United States in 1998. Mortality will be reduced if early detection is executed that allows treatment before cancer is spread to other parts of the body, with survival rate higher than 97% in USA [37]. The principal modality of detecting early signs of breast cancer is X-ray mammography, which is by far the most common imaging method used due to its effectiveness, low cost and high sensitivity. The American Cancer Society has recommended that women between the ages of 40 and 49 to have a mammography exam every other year, and women over 50 to have an exam every year. The incidence of breast cancer will increase with ages up to 75-80 years old [60]. This kind of screening program generates a huge number of mammograms to be examined by a limited number of radiologists and mammographers, which may result in misdiagnosis due to human reading errors caused by visual factors [35]. In order to help the radiologists to improve the accuracy and efficiency of diagnosis in mammography screening, computer aided detection system is introduced in the mammography screening process.

There are two types of mammography: screening and diagnosis. The aim of screening mammography is to find the abnormal regions in a mammogram early. While the diagnosis is to determine whether an abnormal case detected is benign or malignant. Although mammograms are the best tool to find cancers in the early stages, it should be noted that a mammogram may miss certain mass which is called false negative (FN). Also, a mammogram may indicate a ROI to be malignant but it turns out finally to be a benign case, which is called false positive (FP). Although the image quality is a factor to affect the diagnosis decision, one of the main causes of many missed diagnoses is attributed to human factors, such as varying criteria,

distraction by other image features, or simple oversight [37] [53]. Both errors of FN and FP in diagnosis of breast cancer are costly. A FN detection will make patients lose their best chance to treat cancer earlier, and a FP detection will cause an unnecessary biopsy. The detection accuracy of mammogram interpretation can be improved when a mammogram is double read by two radiologists, or reread by the same radiologist [14]. Consequently, the cost of these readings is significant. In addition to that, manual readings are labor intensive and long time work may make radiologists lose some main signs of malignancy. Modern computer based technology can offer a significant reduction of costs and obtain high quality results [16]. Therefore, the development of effective algorithms for automatic mass detection and analysis are of great interest for double reading a mammogram [62].

It is a challenging task to detect masses automatically in digital mammograms. There are some kinds of masses which may be malignant. From the morphogenesis point of view, the masses can be distributed into three types: the circumscribed masses with distinct border and circular shape (well-defined), ill-defined masses with irregular shape and less distinctive border, and the spiculated lesions which appear star-shaped with blurred border. Among these three types of masses, the circumscribed masses are relatively easier to be detected since they are high density radiopaque. The more difficult ones to be detected are ill-defined masses due to the less distinctive borders. The most difficult one is the spiculated lesion which is almost always malignant [50]. The balk in breast mass detection is the variation of normal breast tissue, low contrast and fuzzy margin of the masses, which are partially covered by fat, glandular and density breast tissue in mammographic image. In addition, varied size and shape of masses may fail to generate templates that present the common morphological properties of masses.

In order to increase the detection accuracy and reduce FP and FN effectively, suspicious regions of interest that may contain masses are necessary to be extracted

from the normal breast tissue by image processing techniques and be partitioned into benign and malignant groups by pattern recognition algorithms. One generic automatic mammogram analysis system consists of a two-step process. The first step is intended to detect mass candidates from global or local mammograms. The malignant masses should not be missed during this step, i.e. FN should be close to zero. The second step is an evaluation of the malignancy in order to reduce the number of FP. The selected features in the second step should represent some characters of each candidate region that are useful to identify benign and malignant masses.

Various kinds of image processing algorithms solving the mass localization problems can be found in literature, namely, thresholding [69], bilateral-subtraction [24], Markov random field [31], fuzzy set [66], pixel based feature segmentation [68], region growing [81], iris filter [33], model based method [36], and wavelet [58], etc. Most computerized detection systems do not use one method solely, instead they have combined several image processing and pattern recognition techniques together in order to achieve better detection performance.

1.2 Objectives

The primary objective of this thesis is to study machine learning algorithms for breast mass detection in digital mammograms. It is motivated by the requirement of developing more effective and robust unsupervised clustering algorithms and supervised classifiers based on information and statistical learning theories, for image segmentation to localize region of interest (ROI) and classification of the ROIs, as a help of mammography screening and breast cancer diagnosis that are explained in the previous section.

The task of localization of suspicious regions and boundary finding seems simple

to human eyes, on the contrary, the computerized automatic mass detection and diagnosis in mammograms are not as easy as it appears. It is still difficult for machine to simulate how the human beings perceive, process and store the visual information. The noise, uneven illumination, low contrast and variation between individuals in medical images make the detection and analysis even more complicated. Breast mass detection in digital mammograms is a more challenging task due to the complex structures in appearance and signs of abnormalities. The fuzzy features between benign and malignant masses enhanced the difficulty of diagnosis.

There are other difficulties in the related area. One is that the data sets used in different research studies are not identical. Although researchers may study the same database, while they may chose different mammograms, i.e. not all the mammograms in a particular database are taken into account due to the algorithms may be sensitive to different kinds of images. Another problem is the evaluation of the proposed detection results. In medical image analysis, the most commonly used approach is to validate the results with the radiologist's manual marks. Annotating the boundary and region of a mass manually is a tedious and time-consuming work, and there is no guarantee for a perfect scheme.

Although many early approaches for computerized mass detection in mammograms have led to impressive results, several fundamental issues remain unresolved. For instance, a breast mass is missed in mammograms in part due to the density and contrast of the mass are not optimal for both human and machine observers. In different computer aided detection (CAD) systems, the criteria are not unique, which affect more for the machine observer than human perception. Furthermore, various CAD systems prefer different pattern recognition techniques, there may also exist the mismatch between the data structure and classifier architecture or mismatch between the class boundaries and decision boundaries.

In this thesis, several machine learning algorithms, including information theory

based clustering and statistical learning theory based support vector machine are investigated for mass segmentation and pattern recognition. More specifically, the work in this thesis can be targeted at: 1) image segmentation to locate mass region by incorporating spatial information to increase the accuracy of mass areas and reduce the patched regions. 2) automatic mass detection for masses embedded in glandular and dense glandular breast tissues based on robust information clustering algorithm to ensure that the suspicious masses to be isolated from their background breast tissue with relatively high detection rate and small number of FP. 3) mass pattern analysis to classify the ROIs that may contain masses into normal and abnormal, benign and malignant cases by a semi-automatic detection system, where a new classifier of vicinal support vector machine is proposed for the malignant case classification.

1.3 Major Contributions of the Thesis

The thesis studies a number of new machine learning algorithms in the fields of image segmentation and pattern recognition with the applications in mammographic image segmentation and mass pattern classification. The localization of suspicious masses, automatic detection of masses from global mammograms with weak contrast, and semi-supervised mass diagnosis system for local regions have been investigated. The success of the proposed approaches can be attributed to the study of the information-based and statistical learning theory based new algorithms, with help from the prior knowledge of mammographic images in MiniMIAS database. The contributions are summarized as follows:

- A new algorithm to segment masses in the mammographic images is proposed. The algorithm, with the name of deterministic annealing clustering incorporating circular function (DACF), makes use of the incorporation of

intensity-based feature with circular shape function in the dissimilarity measure of objective function. The proposed algorithm is robust compared to the original DA and fuzzy c-means (FCM) methods in mass area segmentation. It is believed that no previous research have used the DA method incorporating spherical shape function for breast mass segmentation. Comparative studies based on the segmentation results by DACF, DA and FCS were carried out for all the mammograms that contain circumscribed masses in MiniMIAS database, which demonstrated good performance of the proposed DACF.

- A robust clustering algorithm, called robust information clustering (RIC), is proposed for the masses segmentation with low contrast. RIC is developed based on minimax optimization of the mutual information. The algorithm is robust in the sense that both peak and valley of image intensity histogram are estimated automatically. In mass segmentation, the presence of outliers (noises or pixels around the boundary of a mass) make it difficult to determine the code vector of each input pattern. RIC has the ability to identify the outliers and isolate the mass body from its background breast tissues with relative small cluster number and less patched regions.
- The proposed RIC algorithm is applied in a fully automatic breast mass detection system, which employs RIC in the second phase of mass segmentation. The RIC segmentation is carried out based on the raw ROI that is first extracted from global mammograms by two steps of adaptive thresholding. The purpose of the automatic detection scheme by employing RIC is to find the rough location of the suspicious masses in mammograms with relatively high detection rate and small number of FP. Experiments based on the mammograms in MiniMIAS database demonstrate that the proposed RIC method is more robust in mass location detection.
- A new algorithm for classifying the benign and malignant diseases is proposed.

The algorithm, called Vicinal Support Vector Machine (VSVM), is used to minimize the vicinal risk of input patterns in the kernel space and set up a decision function by employing vicinal kernel. The traditional VSVM proposed by Vapnik [112] can only deal with the data in the input space in the form of vicinal kernels corresponding to different vicinities of the training points. For the proposed VSVM algorithm, the whole training data are clustered into different soft vicinal areas in feature space by kernel based deterministic annealing (KBDA) method. The choice of different number of clusters makes VSVM be adaptive to different data structures in the input space.

- The proposed VSVM is applied to a semi-automatic breast mass diagnosis system. The analysis system includes two steps. First, one-class Support Vector Machine (SVM) [113] is applied for the abnormal cases analysis, where only normal cases are served as training samples. VSVM is further investigated for the malignant cases identification. The aim of this step is to decide whether a detected abnormal case is benign or malignant. The whole diagnosis procedure is based on the ROIs that are selected manually.

1.4 Organization of the Thesis

The thesis is organized according to the respective tasks of mammographic mass detection for mammography screening and cancer diagnosis. The rest of this thesis can be divided into the following chapters.

Chapter 2 provides a survey on the mammographic mass detection methods. The medical background of the breast cancer and the mammography imaging modalities are introduced. Literature review about relative projects about mass detection, clustering method for image segmentation and statistical theory for pattern recognition are presented.

Chapter 3 investigates a new clustering algorithm of deterministic annealing incorporating circular function (DACF) for mammographic image segmentation. The objective function in DACF is composed of both intensity related feature and spatial shape information for mass segmentation. The DACF algorithm is illustrated in detail based on a brief review of standard DA method. Experimental results for the circumscribed masses segmentation, the convergence time and the weighted parameter analysis are provided.

Chapter 4 demonstrates the proposed robust information clustering (RIC) algorithm for automatic mass detection. In this algorithm, the mini-max of mutual information is used, which is suitable for outliers identification. The experimental results for RIC in simulated data and mammograms demonstrate good performance of the RIC in mass segmentation as compared with DA and FCM. Finally the full image detection scheme is performed and results are given.

Chapter 5 addresses a new classifier of vicinal support vector machine (VSVM) for breast mass pattern analysis. The system aims to classify a region of interest into normal, benign or malignant case. The basic ideas of one-class support vector machine and VSVM are first introduced. The proposed kernel based VSVM, within which a kernel based DA (KBDA), is presented for malignant mass recognition. The detection scheme is further implemented and experimental results are provided.

Chapter 6 provides the conclusion of the thesis and recommendations for the future work in breast cancer detection.

Chapter 2

Medical Background and Literature Review

2.1 Introduction

The task of computer aided system in breast cancer detection in mammograms is to identify suspicious objects (tumors) of interest by simulating human perception and understanding of the image content. The inputs of a computerized system consist of the digital images that scanned from X-ray film, and its output could be a concise description of the object of interest in a mammogram. However, because the mammographic images are often noisy and complex, and the shapes of masses are various greatly, it is difficult to produce perfect extraction of the objects of interest. To locate and recognize tumors presented in the digital mammograms with different types of background breast tissues, or mammography under various environments, lots of research work must be carried out to develop the theoretical and practical studies. Among different approaches, clustering algorithms have shown their potential for image segmentation, and support vector machines work more effectively for

pattern recognition, both of which have evoked much research interests.

The objective of this chapter is to review the medical background of breast cancer, and the major and significant works in the field of breast cancer detection in mammograms. A survey of clustering algorithms for image segmentation and statistical learning theory for pattern recognition are provided in this chapter. The fundamental knowledge about the breast cancer and mammographic images is introduced in section 2.2. Followed by a brief review of the related projects on breast mass detection in digital mammograms in section 2.3. The standard clustering learning algorithms for image segmentation and statistical learning theory for pattern recognition are highlighted in section 2.4 and section 2.5 respectively.

2.2 Breast Cancer and Mammographic Images

2.2.1 Breast Structure

The breast is composed of two main types of tissue: glandular and stromal tissue. Glandular tissue house the milk-producing glands (lobules) and the ducts, while stromal tissues include fatty and fibrous connective tissues of the breast. In detailed, each breast contains several milk glands with ducts that carry milk to the nipples. About 15 to 20 ducts come together near the areola (circular area around the nipple) to form reservoirs of milk to be drawn from the nipple. In addition, the breast also contains lymph vessels and lymph nodes that trap bacteria, cancer cells or other substances that may be in the lymphatic system [105]. The anatomy of the breast structure is shown in Fig.2.1.

Figure 2.1: The anatomy of the human breast [105].

2.2.2 Know About Breast Cancer

Breast cancer refers to a malignant tumor that has developed from cells, the basic components of tissues in the breast. For the cycling of body tissues, old cells die and new cells grow to take place of them. Once the normal process is broken, i.e. new cells grow when body does not need it, as well as the old cells do not die as body requires, a mass of tissue called tumor will be formulated from these extra cells. It is called ductal when a cancer is seen only in the ducts. So as to the lobular if it occurs only in the lobules. It is called *in situ* when the tumor is confined to the ducts or lobules, otherwise, it is considered as *invasive* or *infiltrating*, as shown in Fig.2.2 and Fig.2.3 [74]. The common belief is that breast cancer originates from duct as the infiltrating ductal carcinoma.

All tumors, which are either benign or malignant, are not necessary to lead to cancer. The benign tumors are not treated as cancer since they are rarely life-threatening, such as fibroadenomas or papillomas, which are abnormal growths but cannot spread

Figure 2.2: Noninvasive breast cancer. A. ducts, B. lobules, C. dilated section of duct to hold milk, D. nipple, E. fat, F. pectoralis major muscle, G. chest wall/rib cage. (a) Normal breast anatomy, enlargement: A. normal duct cells, B. basement membrane, C. lumen (center of duct). (b) Ductal Carcinoma in situ (DCIS), enlargement: B. ductal cancer cells. (c) Lobular Carcinoma in situ (LCIS), enlargement: B. lobular cancer cells.

outside of the breast to other organs and will not grow back once they are removed. On the contrary, the malignant tumors are more serious than benign cases due to they are life-threatening. They can grow back once being removed. They are not only invasive to damage nearby tissues and organs, but can also spread to other organs.

No one knows the exact reasons of breast cancer. But women within the age of forty to fifty-nine do have high risk to this kind of disease. It is by far the most common cancer among women, particular for women in western countries, nearly one in ten will develop breast cancer during their lifetime. Once the cancer case is detected in the early stage, or it can be diagnosed without invasive methods, women can be relieved from life-threaten and anatomy.

Figure 2.3: Invasive breast cancer. (a) Invasive ductal carcinoma (IDC), enlargement: B. ductal cancer cells breaking through the basement membrane. (b) Invasive lobular carcinoma (ILC), enlargement: B. lobular cancer cells breaking through the basement membrane. (c) Vascular and lymphatic invasion, enlargement: A. normal duct cells, B. cancer cells, C. basement membrane, D. lymphatic channel, E. blood vessel, F breast tissue.

2.2.3 Breast Imaging Modalities

There are several imaging modalities which can be used for breast cancer detection and diagnosis. For example, X-ray computed tomography (CT) has the ability of three-dimensional X-ray imaging by acquiring many X-ray views. Although much more information is obtained by this kind of technique than techniques with only two-dimension information, it is not practically used for breast screening because of being over exposed to X-ray dose.

Magnetic resonance imaging (MRI) is a 3D imaging technique and a highly useful adjunct to mammography for detecting breast cancer early. It can find some breast cancers that mammography cannot detect although there is little or no contrast between abnormalities and normal glandular tissue with standard MRI, particularly for the breast cancer detection for pre-menopausal women. But it is not effective for identifying microcalcifications, one of earliest signs of breast cancer. In addition,

the identification of abnormalities currently requires a contrast agent to be injected into the patient's bloodstream. MRI is still a much more expensive modality than X-ray mammography. Patients are recommended to take MRI detection only if they are young to be deemed as high risk of breast cancer or if a mammography screening produces ambiguous results.

Ultrasound is already used in breast cancer assessment to diagnose cyst which are easily discriminated due to their low density, but it can not provide good spatial resolution as mammography. Breast ultrasound is developing quickly, the early reports of the applications of color flow ultrasound to breast cancer diagnosis were extremely encouraging. In breast cancer detection, ultrasound is frequently used to evaluate breast abnormalities that are found with screening mammography, diagnosis mammography, or during a physical performed clinical breast examination.

PET (positron emission tomography) and SPECT (single photon emission computed tomography) are nuclear imaging techniques. The images are created by detecting gamma rays produced by man-made radioactive atoms injected into the patient. Gamma detectors move around the patient to collect data. A computer uses the data gathered by the sensors to construct multicolored two- or three-dimensional images that show the location of the compound acts in the body. While the role of PET in the detection of breast cancer is still under consideration due to its poor spatial resolution, long decay rates and high expense.

Mammography, the topic of this thesis, is a type of X-ray imaging technique to create high contrast and resolution film (or digital) image of breast by using low dose of X-rays. It can show changes in the breast before it is palpable. A lump can be found by mammography, which is a major modality of determination of breast cancer. Once a lump is identified to be a case of abnormality, further breast imaging test such as ultrasound or breast biopsy may be performed. In breast, the fat is relatively transparent to X-ray and abnormalities which have denser tissue are

visible as brighter areas in a mammogram. A problem with X-ray mammography is that glandular tissue is relatively opaque to X-rays. Therefore, it is difficult to detect a mass that is embedded in dense glandular tissues.

2.2.4 Mammographic Images

The digital mammographic images come from two approaches. One is called digital mammography, which is a new technology for producing X-ray mammograms of the breast. In digital mammography, the film normally used for recording the image is replaced by an electronic X-ray detector, such as solid state detector. Another approach is the digitized mammogram, which is obtained by using high-resolution scanner for the original X-ray film. All experiments of this thesis are based on digitized mammograms.

A standard mammography screening in U.S for one breast has two views, one is from above to bottom (cranio-caudal view, CC), the other one is from an oblique or angled view (mediolateral-oblique, MLO). With the MLO view, the pectoral muscle is depicted obliquely from above and visible down to the level of the nipple or further down. For the CC view, the entire breast glandular tissues are depicted. The fatty tissues closest to the breast muscle appear as a dark strip on the X-ray and behind that it should be possible to make out the pectoral muscle, see Fig.2.4.

There are two types of mammography exams: screening and diagnosis. Screening mammography is to show up abnormalities at an early stage when they are too small to be palpable. Early detection of small breast cancers by screening is recommended to women who are above 40 years old, or women before age 40 who have symptom. Diagnostic mammography is an X-ray examination of the breast in a woman who has a breast complaint by self-exam, or has an abnormality found during screening. It is to determine whether an abnormal one is benign or malignant, including the

Figure 2.4: Two basic views of mammography. (a) Sketch map of the two views. (b) Mammogram with MLO view. (c) Mammogram with CC view.

size and location of breast abnormalities. Typically, multiple images of breast are interpreted during diagnostic mammography, which is beyond of the study.

When evaluating a mammogram, a radiologist will first locate a suspicious region of interest in the mammogram and looks for certain signs or characteristics indicative of abnormality. Among these signs are the presence of clustered microcalcifications and masses. There are a number of different kinds of masses which may be malignant. From the shape point of view, there are three types of masses in breast cancer: the circumscribed mass with distinct border and circular shape (well-defined), ill-defined mass with an irregular shape and less distinctive border, and the spiculated lesions which appear star-shaped with blurred border, as illustrated in Fig.2.5.

2.3 Review of Related Projects

Mammography is sensitive to screening and diagnosis of breast cancer. Mass, one of the main signs of breast cancer, is difficult to be detected because it can be simulated or obscured by normal breast parenchyma [30]. It is really a time consuming task for the radiologists to read mammograms not only in terms of quantity, but also quality.

Figure 2.5: The basic abnormalities on mammograms. (a) Microcalcification. (b) Benign mass with well-defined border. (c) Spiculated mass with star-shaped border.

Long time work may make them to lose some main signs of malignancy. Therefore, it is desirable to develop computer-aided detection and diagnosis (CAD) system to provide a second opinion to radiologists in the interpretation of mammograms.

In the standard mammographic mass detection, radiologists will exam the whole image and localized regions that are suspicious. The suspicious regions will be checked further by the same or other radiologists. It seems to be an easy task of human beings to perceive and understand an image, which in fact is a complicated procedure that is still not fully understood. How do human beings perceive, what features do they grasp to describe a mass, and what standard do they set up to identify the masses are still open problems for computer analysis systems. With the development in image processing and computational abilities, computer-based mammographic image processing technique plays a vital role in the mammography screening and diagnosis. Developing automatic mammographic image analyzing and diagnostic system has attracted super interest of many researchers. The goal of a CAD system is to imitate the effect of human vision by electronically perceiving and understanding a mammographic image. But the wide gap between nature vision and computer vision makes it difficult for a CAD system to perform as well as human beings do. Many investigators have developed soft computing techniques for the

analysis of mammograms in order to shorten the gap and driven by the promise that a CAD system can improve breast cancer detection.

It is a challenging task for masses detection in digital mammograms. The features of masses vary greatly in size, shape, and density. Therefore, many computer techniques have employed multi-concepts for the detection of masses in mammographic images. The basic control strategy of these methods employs bottom-up scheme [6], which starts from the original raw image data, works upward to enhancement of the preselected features and removal of irrelevant details using application-dependent image processing techniques, location of suspicious areas by image segmentation method, object description and recognition of these areas into normal, benign or malignant cases. A simple three-step of bottom-up approach is: pre-processing, image segmentation, and pattern recognition, as shown in Fig.2.6. Each step may consist of several sub-steps. In the pre-processing, the information that is concerned in the original image is highlighted by sub-steps such as contrast enhancement, filtering, etc. The modified image is more convenient for further processing. In segmentation, the modified image is segmented according to different purposes and regions of interest is extracted. The final pattern recognition procedure consists of two sub-steps of feature extraction and classification.

Ackerman [57] applied computerized image processing techniques and Bayesian classification scheme to tackle the classification problem of analyzing the breast lesions on xeroradiograms as early as 1971, which may be the earliest work for breast cancer diagnosis. From then on, a large number of image analysis techniques has been developed for mammographic image enhancement, object localization and pattern recognition. However, due to the character of mammographic images, many image interpretation techniques have been found to be application dependent. Thus, researchers try to combine and modify some existing algorithms to create the best techniques for the mammographic image analysis. The research work done in the

Figure 2.6: Diagram of bottom-up control strategy for breast mass detection in digital mammograms.

automatic mammographic image processing and mass pattern analysis in digital mammograms will be reviewed in this section.

2.3.1 Mammographic Image Processing

Localization of the mass is considered a critical and the first step in automatic detection of mammographic masses. The mass should not be missed before any further analysis. Multiscale concepts have been employed by many computer aided techniques for the detection of masses in mammographic images. Lai *et al* [102] first proposed to detect circumscribed masses based on a combination criteria including the shape, brightness contrast and uniform density of tumor areas. A suspicious is defined in the study as an area which 1) is brighter than its surrounding tissue, 2) has uniform density inside the area, 3) has an approximately circular shape of varying size, and 4) has fuzzy edge. They used modified median filtering to enhance mammogram images and template matching to detect breast mass. Brzakovic *et al* [16] used thresholding and fuzzy pyramid linking for mass localization. Qian *et*

al [116] have proposed using modified and weighted median filtering to enhance the digitized image prior to object identification.

Li *et al.* [31] proposed a Markov random field approach that lies in the category of region based algorithm to do breast mass detection. The system is performed in two groups: segmentation and classification, which are the two common steps in the automatic CAD algorithms. The region of interest is first extracted from the images by adaptive thresholding. The second step segmentation is achieved by a modified Markov random field model-base method. N.Petrick *et al.* [69] proposed a two-stage adaptive density-weighted contrast enhancement (DWCE) filter in conjunction with a Laplacian-of-Gaussian edge detector for mass detection. The DWCE is first applied to the entire mammogram to locate regions containing potential masses. The second DWCE is applied to these local regions and provides enhancement for the final definition of the breast structure. A combined adaptive DWCE and region-growing segmentation algorithm is further developed by this group [70]. Instead of the the edge detector in [69], the region-growing technique is applied in the paper to each of the identified structure.

Sameti *et al* [66] used fuzzy sets to partition the mammographic image data. A model-based vision algorithm is developed in [115] using difference-of-Gaussian filters to detect masses and computed nine features based on size, circularity, contrast and law texture features. Kupinski *et al.* [59] proposed to use seeded segmentation algorithm as radial gradient index and simple probabilistic models on lesion segmentation.

Naga and Rangaraj [71] proposed to employ Gaussian smoothing and sub-sampling operations as preprocessing steps in mass detection. The mass portions are segmented by establishing intensity links from the central portions of masses into the surrounding areas. Li *et al.* [36] used a statistical model supported approach for enhanced segmentation and extraction of suspicious mass areas from mammographic

images. Zheng *et al.* [58] suggested several artificial intelligent techniques as fractal dimension analysis, multiresolution markov random field, etc, combined with discrete wavelet transform for detection of masses in mammograms. A template matching scheme based on mutual information [27] is proposed recently. In this method, the computer assisted detection scheme utilizes a knowledge databank of mammographic regions of interest with known ground truth. Each ROI in the databank serves as a template. The CAD system follows a template matching approach with mutual information as the similarity metric to determine the abnormalities of a query ROI.

Some investigators concentrated on using the asymmetry between the right and left breast images to determine possible mass locations. Yin *et al* [25] uses both linear and nonlinear bilateral subtraction that involves linking multiple subtracted images to locate initial candidate masses. While the structure asymmetry in the method used by Lau *et al* [110] is between the two breast images. Yin *et al* [26] further developed a bilateral registration technique, which is based on determining correspondence between control points on the skin line. The corresponding points of the right breast were registered with those of the left breast using a least-squares method, allowing translation and rotation. Hadjiiski *et al* [52] developed an automated regional registration to identify corresponding lesions in temporal pairs of mammograms. Saskia *et al* [107] implemented and validated four methods for temporal mammogram registration. The first two methods are based on nipple location and center of mass of the breast area. The third method uses mutual information that is computed from the joint probability distribution of the images' intensities. The fourth one is a warping approach.

However, it is difficult to compare the effectiveness of these methods since each used a unique set of digitized mamommograms.

2.3.2 False Positive Reduction

In order to achieve a high detection rate, the image processing algorithms are normally introduced together with other technologies for detecting masses in mammograms. In image processing step, all the suspicious regions are singled out for further analysis. According to different processing methods, the number of located regions can be different, most of which are non-mass regions. In addition, a large number of potential regions may be found simultaneously together with these mass-like regions. The goal of FP reduction is to reduce the number of FP regions without losing a significant number of true masses.

Normally FP reduction occurs during both stages of image segmentation and classification. It is an very important stage in computerized mass detection system for the help of mammography screening. The initial purpose of mammographic screening is to find any abnormal regions in a mammogram. It is a decision between normal breast tissue and potential mass region. Petrick *et al* [69] have illustrated detailed of an automatic detection system. FP reduction was first applied in the first step of global object edge detection. In this stage, altogether 11 morphological features such as the number of edge pixels, the total object area, the object's contrast, circularity and rectangularity, etc, were used for FP reduction. The extracted morphological features were used in a sequential classification scheme: a simple threshold classifier, followed by an LDA classifier, and finally is the back propagation neural network. The local FP reduction was the same as global FP reduction. Texture features were used for the final decision between masses and normal structures. The detection system indicated 4.4 FPs per image at a true positive (TP) of 90% and 2.3 FPs per image at a TP rate of 80%. In their later study [70], each object after DWCE filter and region growing is classified as a breast mass or normal tissue based on extracted morphological and texture features. After FP reduction, the detection resulted in 4.2 FP per image at a 90% true positive (TP) and 2.0 FPs per image at an 80% TP

fraction.

Guido *et al* [29] have studied 132 mammograms that contain 71 malignant tumors. A sensitivity of approximately 75% is reached at 1 FP per image. The system in [33] analyzed 1212 images. The sensitivity was obtained at 90.5% with 1.3 FP per image. A tree-type classification strategy was applied in [58] to determine whether a given region is suspicious for cancer. The detection method resulting the sensitivity of 97.3% at 3.92 FPs per image for 322 mammograms. Rangayyan *et al* [71] have not only studied FP reduction for the normal and abnormal cases identification at a sensitivity of 81% with 2.2 FPs per image, but also classified the detected abnormal cases into benign and malignant masses, a procedure of mammographic diagnosis.

2.3.3 Mammographic Mass Diagnosis

In order to decide the malignancy of mass candidates identified by mass detection system or extracted manually, machine learning algorithms should be carried out by various CAD systems and a knowledge database can be constructed in a multidimensional feature space. As the first step, the selected features should have the property of discrimination, reliability, independence and optimality. In detail, discrimination and reliability indicate that features for patterns in various classes should have significantly different values, for patterns in the same class should have similar values. Independence means that different features should not be strongly correlated to each other. The optimality refers to the optimal number of features that are not redundant and can represent each class well for reducing the complexity of the classifier.

The localized suspicious areas are diagnosed as normal, benign and malignant cases by the mammographic appearance [86]. The primary features that indicate malignancy are related to the mass' density, geometric and texture features, where

the geometric features mostly refer to size, shape and border. Most benign masses are homogeneous and possess well-defined boundary. The boundary of the malignant mass are typically fuzzy or ill-defined, comprising fine linear strands extending irregularly outward from the center of the mass since they often infiltrate the surrounding tissues. The shape of the mass can be used to discriminate between benign and malignant masses. Many studies have therefore focused on analyzing the shapes of the masses.

Kilday [41] developed a set of seven shape features based on circularity and radial distance measure (RDM) from the centroid to the points on the boundary. The features included compactness, mean of RDM, standard deviation of RDM, entropy of the RDM histogram, area ratio, zero crossing and boundary roughness. The classification accuracy of 51% is obtained for fibroadenomas, cysts and cancers classification. Huo [120] proposed two spiculated measures obtained from the radial edge gradient information surrounding the periphery of the located regions. Skeleton analysis methods using the iris filter to detect spiculated lesions is addressed in [34] and moments of distances of contour points from the centroid, compactness of the boundary are used in [89]. Lori *et al* [54] utilized the multiresolution shape features by wavelet transform method. Detailed information of local boundary has been analyzed in [90].

The formulated shape-based features are effective to distinguish benign from malignant tumors with distinguished shape information. They failed to identify the benign and malignant cases that possess the similar shapes. Although the majority of benign masses in mammograms are well circumscribed, some do possess irregular or spiculated distortions. On the other hand, some malignant masses do have well-defined margins [64]. Moreover, mammographic mass boundaries are usually blurry, making it difficult for automated detection scheme to detect precisely the mass boundaries. Considering the benign masses typically to be more homogeneous

compared with the malignant masses, some research work focus on features based on the gradient and texture information contained by mass to perform classification task.

In [12], the breast tissues are partitioned into 144 sections per image. Ten normalized statistics for each section were used as texture features to classify 2270 mammographic sections of eight patients. Six out of the ten features are reported to yield best-performance with false positive rate of 26% and a false negative of 0.6%. Spatial gray-level dependence (SGLD) based texture features are investigated in [5] for the classification of masses and normal tissues. For a dataset of 135 ROIs, the method results in 89% sensitivity and 76% specificity in the training step and 76% sensitivity and 64% specificity in the test step. Chan *et al* [21] have also studied the effectiveness of texture features based on SGLD matrices to identify masses from normal breast tissue. In their method, altogether 168 ROIs with masses and 504 normal ROIs were examined by a linear discriminant classifier. The area under the receiver operating characteristics (ROC) curve of 0.89 and 0.86 for training and testing groups.

Gray-level co-occurrence matrices (GLCM) have been analyzed by some researchers. Kinoshita *et al* [100] employed feed-forward neural network by using the combination of shape and GLCM based texture features for the classification of benign and malignant breast lesions. An accuracy of 81% was reported for a dataset of 38 malignant and 54 benign lesions. Gradient-based and GLCM based texture features are further studied in [72]. The classification method used is based on posterior probabilities computed from Mahalanobis distances. The corresponding accuracy using jack-knife classification were observed to be 0.52 and 0.73 for 39 images from MIAS database and the combined database respectively.

Based on the extracted features, some research work focus on the development of decision support classifiers. Li *et al* [35] has set up a knowledge database including

both shape and texture features and developed a probabilistic modular neural network (PMNN), which is considered as nonlinear classifier, to carry out the mapping function of the knowledge database. A multiple circular path convolution neural network (MCPCNN) architecture was specifically designed in [94] for analyzing mass features. The area under ROC curve was obtained ranging from 0.84 to 0.89, where the the feed forward neural network has 0.78 to 0.80 area under ROC. Another kind of neural network technique is illustrated in [77], called hierarchical pyramid neural network.

2.4 Review of Clustering Algorithms for Medical Image Segmentation

The fundamental step in many CAD and image analysis systems is the segmentation of the input image, which is critically dependent on the scene to be sensed, the imaging geometry, configuration and sensor used to transduce the scene into a digital image, and ultimately the desired output of the system. Image segmentation remains a difficult task due to the tremendous variability of object shapes and the variation in image quality. Furthermore, an image often demonstrates details at multiple scales. In particular, medical images are often corrupted by noise and sampling artifacts. The challenging in image segmentation lies in: the complexity of modelling a large amount visual patterns that appear in generic images, and the intrinsic ambiguities in image perception, especially when there is no specific guidance to the understanding. Tu and Zhu [121] believe that a segmentation algorithm generates normally non-unique result, instead it should produce multiple distinct solutions dynamically and endlessly so that these solutions can best preserve the intrinsic ambiguity.

As one type of medical image analysis systems, the breast mass segmentation is an even hard technique in digitized mammographic images. First, the mass shapes vary a lot. Some are circumscribed, some are speculated, and some are architecture distortion, etc. Second, the background breast tissues are different according to the ages, races of women. The breast of young ladies' are mostly dense, and the breast of old people's are mostly fat. Women in Asia possess even dense breast tissue. Third, the mammographic images often contain noisy objects or artifacts. It is really a hard task to model the mass due to its complexity, the variance of image quality, and the intrinsic ambiguities between mass and its background breast tissue.

Methods for performing segmentations vary widely depending on the specific applications that may be divided into eight categories: thresholding, region growing, classifiers, clustering, Markov random field models, artificial neural networks, deformable models, and atlas guided approaches [17]. These approaches belong to three domains: pixel-based methods, region-based techniques, connectivity-preserving relaxation methods. A pixel-based method, such as thresholding, classifier, and clustering techniques, is often a desirable goal in medical image segmentation when it is necessary to identify the disconnected regions belonging to the same tissue class. The shortcoming is that much noisy points (or regions) and many holes inside of an objects will be generated simultaneously. Spatial information is used for region based methods, e.g. region growing, Markov random field [96], etc. These methods have the problem of computational complexity. The main idea of a connectivity-preserving relaxation-based segmentation method is to start with some initial boundary shape and iteratively modify it by applying various shrink or expansion operations according to some energy function. The typical method is deformable model [15], which has the risk to be trapped into a local minimum.

Clustering is originally designed for unsupervised classification of feature patterns by grouping data using a given measure of similarity, a detailed review is given

in [3]. By analyzing the feature points in the input 2D image, clustering algorithms have found wide range of applications on image segmentation. Image segmentation requires that a decision is made about the type of a pixel, i.e. which feature groups it belongs to. It is an observation at each pixel and describes the relation between feature vectors and pattern classes, which can also be stated as partitioning a given set of pixels into subgroups, each of which should be as homogeneous as possible within group, and should be as inhomogeneous as possible between groups. A brief illustration of clustering method in image segmentation is shown in Fig.2.7.

Figure 2.7: Illustration of clustering methods in image segmentation. Pixel intensities in an image are randomly distributed. Suppose the intensity-based features belong to three groups or clusters, represented by 1,2,3 in the figure. It shows that the pixels belong to the same groups or clusters may not locate at the same region, i.e. they can be partitioned into unconnected in different subregions.

There are basically two approaches in pattern clustering, i.e. parametric and non-parametric algorithm [13]. In the parametric clustering approach, a pre-defined probability density function is used to calculate the clustering objective function to achieve the optimal performance, which has limited applications in case of unknown distributions of the input pattern [28]. On the other hand, the nonparametric approach is to partition data without a priori knowledge of the data distribution. The clustering problems is usually made by defining a cost criterion to be minimized. The objective function is to evaluate how well the clustering fits to the data. It is a

basic optimization problem. Most of clustering algorithms use the gradient descent algorithm, such as K-means, fuzzy K-means or fuzzy c-means (FCM), and fuzzy c shell clustering algorithm [85]. In order to prevent convergence to local minima, simulated annealing are incorporated into the clustering procedure.

2.4.1 Fuzzy C-Means Clustering

Fuzzy c-means (FCM) is an extension of classic K-means using the concept of fuzzy logic [39]. Suppose the vectors of input patterns are $\mathbf{x}_i$, $i = 1, \dots, l$, which will be partitioned into c clusters, then the basic objective function of FCM can be described as follows:

$$J_m = \sum_{k=1}^c \sum_{i=1}^l (u_{ki})^m (d_{ki})^2 \quad (2.1)$$

where $m \in (1, \infty)$ is a parameter to increase or decrease the fuzziness. Higher values of fuzziness will make the result fuzzier. For normal data an $m = 2.0$ is used most of the times. The u_{ki} is the membership of pattern $\mathbf{x}_i$ to the k th cluster. Suppose the feature vector of cluster center are $\mathbf{v}_k$, $k = 1, \dots, c$, the dissimilarity measure d_{ki} is calculated as the squared Euclidean distance

$$(d_{ki})^2 = \|\mathbf{x}_i - \mathbf{v}_k\|^2 \quad (2.2)$$

The solutions that minimizing J_m are least-squared error stationary point of J_m . The cluster centers are obtained as

$$\mathbf{v}_k = \frac{\sum_{i=1}^l (u_{ki})^m \mathbf{x}_i}{\sum_{i=1}^l (u_{ki})^m} \quad (2.3)$$

and the membership functions are

$$u_{ki} = \left(\sum_{j=1}^c \left(\frac{d_{ki}}{d_{ji}} \right)^{\frac{2}{m-1}} \right)^{-1} \quad (2.4)$$

The basic procedure to find the solution is Picard iteration.

The FCM algorithms tend to find patterns with feature vector of spherical in shape distribution, and have been applied in image segmentation [42]. According to specific application, image modality and other factors, many improved version of FCM have been proposed, particular spatial constraints are incorporated with fuzzy image clustering. The main concept imposing spatial constraints in [118] is to add a neighborhood voting scheme that decide the possibility of the central pixel belonging to a certain cluster. It has built a fuzzy-system with three inputs, a single output and 11 rules that applies the spatial constraint to the memberships. The influence of the neighbouring pixels on the center in a 3×3 window has been taken into account in the dissimilarity measure in [8]. Good performance for noise reduction was obtained. Elliptic shape function constraint was developed later for the detection of the lip image [98].

In order to overcome the intensity inhomogeneity in magnetic resonance images (MRI), the objective function in FCM is further developed in some research work to obtain optimal image segmentation results. Pham *et al* [18] have proposed an adaptive FCM method for 2-D magnetic resonance images (MRI) segmentation when the intensity in the image are inhomogeneous. They extended the idea further into 3-D MRI image segmentation [19]. A modified version of adaptive FCM method for MRI image segmentation is in [65]. Chen *et al* [95] introduced the FCM with spatial constraints into kernel space, where a distance measure is set up to enhance the clustering performance of anti noise and outliers.

Although there are some differences between these proposed algorithms for image

segmentation, the basic ideas are the same, i.e. the spatial constraints are added in the dissimilarity measure of the objective function in the standard FCM. These algorithms are determined by the character of image data, which are spatial continuous. In other word, the membership of a pixel can not be determined by the features of the pixel solely, but features of its neighborhood pixels. Since the spatial constraint are optimized with the intensity-based features simultaneously during clustering process, no matter the neighborhood pixels or the spatial shape function, some noise points can be annihilated and the intensity homogeneity can be corrected. The optimization of these algorithms often falls into local minima since FCM is a local minimization method.

2.4.2 Robust Fuzzy C-Means Clustering

In the applications of many FCM clustering methods, the real dataset often contains noisy points, which will be also assigned membership to one or more clusters even though they are far away from the cluster prototypes. The existing of noises can decrease the performance of clustering results. Therefore, how to handle the noisy points is an interesting topic that has been discussed in many research work, commonly called robust clustering methods. The robustness means that the performance of an algorithm should not be affected significantly by small deviations from the assumed model and it should not deteriorate drastically due to noise and outliers. A detailed review for robust clustering methods can be found in [92].

Different from some improved version of FCM algorithms which have certain robustness to noisy points and outliers, Dave [91] has proposed a robust fuzzy clustering algorithm based on the concept of *noise cluster* or *noisy prototype*. The noisy prototype is a universal term that is always at the same distance from every point in the dataset. Assume that there are c clusters as in the standard FCM, excluding noises in the dataset. Then one cluster of noisy data is added, i.e., the $(c+1)$ th cluster is

the noise cluster. The objective function in FCM is reformulated as

$$\hat{J}_m = \sum_{k=1}^{c+1} \sum_{i=1}^l (\hat{u}_{ki})^m (\hat{d}_{ki})^2 \quad (2.5)$$

$$= \sum_{k=1}^c \sum_{i=1}^l (u_{ki})^m (d_{ki})^2 + \sum_{i=1}^l u_{(c+1)i}^m \delta^2 \quad (2.6)$$

where δ^2 is the distance from pattern $\mathbf{x}_i$ to the noisy prototype. The formulation of the membership and cluster center is the same as in the standard FCM for $k=1$ to c . The main difference is that the constraint on the membership is now defined as

$$0 < \sum_{k=1}^c (u_{ki}) \leq 1 \quad (2.7)$$

instead of the definition in conventional FCM algorithm

$$\sum_{k=1}^c (u_{ki}) = 1 \quad (2.8)$$

Therefore, the membership $u_{(c+1)i}$ is defined to be

$$u_{(c+1)i} = 1 - \sum_{k=1}^c (u_{ki}) \quad (2.9)$$

This approach is shown quite successful in improving the robustness of prototype-based clustering algorithms [91].

Krishnapuram *et al* [84] recast the fuzzy data partition problem into probabilistic frame work and proposed a possibilistic approach for robust c-means clustering. By using this approach, each point belongs to a cluster with certain possibilistic membership. If a noisy point is far away from all the c clusters, its membership to all these clusters will be possibilistic small. It does not affect the clustering results significantly. A pruning approach for robust c-means clustering to identify and

prune the outliers is in [45], which is based on the sizes and shapes of the clusters so that the clustering results are least affected by the noises. It adopts C thresholds to distinguish the noisy points from the c clusters. Some other approaches for robust clustering can be found in [40] and [119].

2.4.3 Deterministic Annealing Clustering

Unsupervised clustering is an important optimization problem since the data analysis does not have the prior knowledge of the probability distribution. Clustering is usually formulated as an optimization problem to minimize an objective (or cost) function. Most of these objective functions are based on a minimum variance criterion and are nonconvex at several local minima. As indicated previously, the fuzzy set clustering methods employ decent algorithms and tend to get trapped in local minima. Simulated annealing or stochastic relaxation [99] is a technique for avoiding local minima with a sequence of random moves being generated to escape local minima. While it is not realistic for many practical applications due to the process requiring a very slow schedule. Since only the structure of the data contributes to the clustering, thus, information theory is an obvious criteria to establish the clustering rule [22].

Rose *et al* [49] proposed the concept of deterministic annealing (DA) for the problem of clustering and vector quantization. The method is based on the principles of information theory of maximizing Shannon entropy and utilizes the idea of annealing process from physical analogy. The cost function is parameterized by a Lagrangian parameter, called temperature. According to Rose [48], the DA method offers three important features: 1) the ability to avoid many poor local optima, 2) applicability to many different structures or architectures, 3) the ability to minimize the right cost function even when its gradients vanish almost everywhere as in the case of the empirical classification error. Due to these advantages, the DA method is used

in many research work for data clustering [108] [104] and image segmentation [109] [101]. This thesis studies several extension algorithms based on the DA method for data clustering and image segmentation.

2.5 Review of Statistical Learning Theory for Pattern Recognition

As indicated in section 2.3, pattern recognition techniques are necessary and important in the steps of false positive reduction and mass pattern diagnosis. Although the selected features between normal and abnormal, benign and malignant cases should have the properties of discrimination, reliability, independence and optimality, while due to the characters of masses in digitized mammograms, the selected features can not always hold such properties, i.e. the feature data may be linearly nonseparable or nonlinear nonseparable. The design of a pattern recognition system involves the problem to estimate the underlying properties of the feature data. Currently, statistical learning theory based Support Vector Machine(SVM) is a popular classifier due to its formulation embodies the Structural Risk minimization (SRM) principle, which has been shown to be superior to traditional Empirical Risk Minimization (ERM) principle that is employed by conventional neural networks.

Statistical learning theory was introduced in the late 1960's, it was a purely theoretical analysis for function estimation from a given collection of data until the middle 1990's, when a new statistical learning algorithm *support vector machine* (SVM) was proposed [112]. This made the statistical learning theory a tool for both theoretical analysis and practical application algorithm for estimating functions.

For a given learning task, with a given finite amount of examples, a learning machine is expected to have best generalization performance if the good balance is achieved

between the accuracy obtained on the particular example and the capacity of the machine, that is, the ability of the machine to learn any training set without error [11]. Actually, a trained machine with too large capacity will overfit the training data and increase generalization error. In other word, a tiny change of a pattern will lead the machine make wrong decision. A machine with very small capacity may recognize different objects belonging to different classes as the same class, it can only recognize the outliers but not detail. In order to get best generalization performance, good balance should be achieved between the capacity and the performance of a learning machine.

There are three basic components for the model of learning from examples [114]:

1. a generator that generates random input vectors $\mathbf{x}$, drawn independently from a fixed but unknown distribution $P(\mathbf{x})$.
2. a supervisor that provides an output value y for every input vector $\mathbf{x}$, according to a fixed but unknown conditional distribution function $P(y|\mathbf{x})$.
3. a learning machine that is capable of implementing a set of function $f(\mathbf{x}, \alpha)$, $\alpha \in \Lambda$.

The problem of learning is to choose the best function, call trained machine, from $f(\mathbf{x}, \alpha)$ that predicts the supervisor's response. In order to achieve this purpose, a risk function is set up based on the loss $L(y, f(\mathbf{x}, \alpha))$ that is measured between the supervisor's response y and the learning machine's response $f(\mathbf{x}, \alpha)$, stated as

$$R(\alpha) = \int L(y, f(\mathbf{x}, \alpha)) dP(\mathbf{x}, y). \quad (2.10)$$

where the joint probability distribution $P(\mathbf{x}, y)$ is unknown and the only available information is contained in the training set.

2.5.1 Empirical Risk Minimization

The empirical risk minimization (ERM) principle is quite general in the classical methods for solving learning problems, such as neural networks. Suppose the training set contains l training samples, $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)$, the expectation of the test error for a trained machine is therefore:

$$R(\alpha) = \int \frac{1}{2} |y - f(\mathbf{x}, \alpha)| dP(\mathbf{x}, y) \quad (2.11)$$

$R(\alpha)$ is called the expected risk, or risk. It provides a true mean error. But $R(\alpha)$ is unknown since the probability distribution function $P(\mathbf{x}, y)$ is also unknown.

As only the training data is given, the training error of the *empirical risk* $R_{emp}(\alpha)$ is a measure that could be used to decide which of the possible functions is preferable.

$R_{emp}(\alpha)$ is defined to be just the measured mean error rate on the training set:

$$R_{emp}(\alpha) = \frac{1}{2l} \sum_{i=1}^l |y_i - f(\mathbf{x}_i, \alpha)| \quad (2.12)$$

The quantity $\frac{1}{2} |y_i - f(\mathbf{x}_i, \alpha)|$ is the loss. $R_{emp}(\alpha)$ is a fixed number for a particular choice of α and for a particular training set $\{\mathbf{x}_i, y_i\}$, no probability distribution appears here, while limited number of training observations will not guarantee a high generalization ability if directly minimizing the empirical risk.

2.5.2 Structural Risk Minimization

As a part of statistical learning theory, VC (Vapnic-Chervonenkis) theory introduces bounds on the expected risk that depend on both the empirical risk and the capacity F , which is a set of functions where f_α is chosen from.

Theorem: With probability at least $1 - \eta$, the inequality

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(\log(2l/h) + 1) - \log(\eta/4)}{l}} \quad (2.13)$$

holds true simultaneously for all functions of the set F , where $0 < \eta < 1$.

This bound shows that in order to achieve small expected risk, both the empirical risk and the ratio between the VC-dimension h and the number of the training points l should be small. The VC dimension is defined as the largest number of points that can be separated in all possible ways using functions of the given set. The second term of this bound is called VC confidence. The VC confidence depends on the chosen class of functions. The right choice of α for controlling $R_{emp}(\alpha)$ is made through experimentation. For controlling h , a structure is introduced, that F is composed of the nested subsets of functions $F = \{f_\alpha, \alpha \in \Lambda\}$, such that

$$F_1 \subset F_2 \subset \dots \subset F_n$$

whose VC-dimensions satisfy

$$h_1 \leq h_2 \leq \dots \leq h_n$$

Training a set of learning machines, one for each subset, then SRM will find a subset of functions that minimizes the bound on the expected risk. the corresponding learning machine is the trained machine in the series whose sum of empirical risk and VC confidence is minimal, as indicated in Fig.2.8.

2.5.3 Support Vector Machine

SVM is a promising pattern classification technique that was introduced by Vapnik and his colleague [113]. It can be seen as a new training algorithm for the

Figure 2.8: Relationship between ERM and SRM

traditional polynomial, radial basis function and multilayer perceptron classifier by defining relevant kernel functions. SVM maps input vectors into a high dimensional feature space by nonlinear transform function, where a maximal margin hyperplane is constructed. It is based on SRM principle and aims at minimizing an upper bound of the generalization error through maximizing the margin between the separating hyperplane and the data. It is necessary to find a class of functions whose capacity can be computed in order to design a learning machine from the empirical data to construct generalization separating hyperplanes. This algorithm is based on the fact that among all possible hyperplanes, there exists a unique one with maximum margin of separation from the two classes, i.e. optimal margin hyperplane. Fig.2.9 gives some geometrical interpretation of the reason that large margin is expected from a separating hyperplane, where better generalization capability is expected from (b).

Suppose it is a two-class classification problem, with output $y_i \in \{-1, 1\}, i = 1, \dots, l$. When the l samples are not linearly separable in the input space, they are transformed into a high dimensional feature space by nonlinear function ϕ , where they

Figure 2.9: a. A separating hyperplane with small margin. b. A separating hyperplane with larger margin.

are linearly separable. The margin in the feature space is defined by the distance between the separating hyperplane to the nearest positive and negative points, calculated as $2/\|w\|$. According to the SRM principle, the risk bound is minimized by formulating the optimization problem

$$\text{Minimize } \Phi(w) = \frac{1}{2}w^T w + C \sum_{i=1}^l \xi_i \quad (2.14)$$

$$\text{subject to } y_i f(\phi(\mathbf{x}_i)) \geq 1 - \xi_i \quad (2.15)$$

where w is weight vector, $\xi_i \geq 0$ is slack variables and C is penalty parameter. The support vectors are those particular points that satisfy the following equation

$$y_i f(\phi(\mathbf{x}_i)) = 1 - \xi_i \quad (2.16)$$

The decision function is

$$f(\phi(\mathbf{x}_i)) = w^T \cdot \phi(\mathbf{x}_i) + b \quad (2.17)$$

where b is bias. To solve this problem the Lagrangian is introduced

$$L(w, b, \xi, \alpha) = \frac{1}{2}w^T w - \sum_{i=1}^1 \alpha_i [y_i(w^T \cdot \phi(\mathbf{x}_i) + b) - 1 + \xi_i] \quad (2.18)$$

where $\alpha = (\alpha_1, \dots, \alpha_l)$ is the Lagrangian multipliers. The parameters that minimize the lagrangian must satisfy the conditions

$$\frac{\partial L(w, b, \xi, \alpha)}{\partial w} = 0 \quad (2.19)$$

$$\frac{\partial L(w, b, \xi, \alpha)}{\partial b} = 0 \quad (2.20)$$

The new dual form function is finally obtained

$$W(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \quad (2.21)$$

subject to the constraints

$$\sum_{i=1}^l y_i \alpha_i = 0 \quad (2.22)$$

$$0 \leq \alpha_i \leq C \quad (2.23)$$

In the solution, those points for which $\alpha_i > 0$ are support vectors. The dot product in the high dimensional space is equivalent to a kernel function in the input space, as $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$. It shows that the mapping function $\phi(\cdot)$ is not necessarily known. The dural form function finally takes the form

$$W(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \quad (2.24)$$

According Mercer's condition, there are many kernel functions can be used, the most common ones are: Gaussian radial basis function(RBF) and the polynomial

function, as

$$k(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2} \quad (2.25)$$

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d \quad (2.26)$$

the width σ^2 in Gaussian RBF kernel function and the power d in polynomial kernel function are specified *a priori* by the users.

In the field of pattern recognition, SVMs have found widely uses in face detection [23], text categorization [46], tracking [93], speech recognition [2], etc. In most of these cases, SVM generalization performance either matches or is better than that of the competing methods.

2.6 Concluding Remarks

This chapter presents a very general overview of the problems for breast mass detection in digitized mammograms from the medical point of view. The computer aided detection techniques by employing the clustering methods for medical image segmentation and statistical learning theory for classification are also reviewed. Based on the presented learning algorithms for image segmentation, DA method offers the global optima, where fuzzy C-means clustering is a local optima method. As a result, robust clustering algorithms based on DA optimization which is generated from information theory will also offer the properties as DA. For classification, statistical learning theory based SVM provides a new approach to the problem of mass pattern recognition. It is based on the assumption that data are identical and independent distribution(i.i.d) and the statistical learning theory thus links to the problems when the feature data are locally or vicinal i.i.d distribution.

Chapter 3

Deterministic Annealing Clustering Incorporating Circular Function for Mass Segmentation

3.1 Introduction

As presented in section 2.4, deterministic annealing method has three major advantages over the existing clustering algorithms, namely, the ability to avoid many poor optima, the applicability to various data structure, and the ability to minimize the right cost function. Including DA, most clustering methods perform image segmentation directly from the intensity (or color) space with an intensity filter to enlarge the difference between the mass region and normal breast tissues. The processing time is a prominent advantage of these algorithms. However, since the segmentation results often contain noisy points or regions and the mammographic images have low contrast, the intensity based methods cannot satisfactorily outline the boundary

of the mass region. Markov random field technique has been used in mass segmentation [31] to exploit the spatial continuity in order to improve the segmentation performance, which can reduce the segmentation error caused by intensity noise, but the computational cost is high. Leung *et al* [98] [7] proposed a fuzzy clustering algorithm incorporating an elliptic shape function for lip image segmentation. This algorithm takes both the color information and spatial distance in terms of an elliptic shape function into account. The algorithm is able to differentiate the pixels having similar color information but located in different regions.

In this chapter, the idea of incorporating spatial shape information into the objective function is taken into account in DA optimization, i.e. a new clustering objective function, which is composed of both intensity related feature and spatial shape information, is proposed for deterministic annealing clustering. A circular shape function is embedded in the dissimilarity measure of the objective function. The dominant dissimilarity measure of either intensity base feature or spatial information is controlled by two weighting parameters. As a result, pixels having similar intensity information but locating in different regions are differentiated. In addition, the convergent time of the clustering method decreases or keeps approximately constant as the difference value between the two weighting parameters increases. The algorithm is applied in circumscribed mass segmentation, which exhibits spherical shape property (or close to spherical shapes). Experimental results show that the proposed deterministic annealing incorporating circular function (DACF) algorithm works effective for circumscribed mass segmentation.

3.2 A Brief Review of the Mass Constrained Deterministic Annealing

The deterministic annealing (DA) algorithm, in particular, the mass constrained deterministic annealing, which aims to locate the mass centers of the given data set, is independent from the initialization of clustering procedure [49]. The prototype (or centroid) in mass constrained DA clustering algorithm is the mass center.

Suppose there are l input vectors $\mathbf{x}_i \in R^N$, $i = 1, \dots, l$, which are partitioned into K clusters with mass center at $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K\} \subset R^N$. The squared Euclidian distance is defined as the distance from any data to a cluster center

$$d(\mathbf{x}_i, \mathbf{c}_k) = \|\mathbf{x}_i - \mathbf{c}_k\|^2 \quad (3.1)$$

where $\mathbf{c}_k$, $k = 1, \dots, K$, is the prototype. The clustering cost function is the expectation of the average distortion

$$D(\mathbf{X}, \mathbf{C}) = \sum_{i=1}^l \sum_{k=1}^K p(\mathbf{x}_i, \mathbf{c}_k) d(\mathbf{x}_i, \mathbf{c}_k) = \sum_{i=1}^l p(\mathbf{x}_i) \sum_{k=1}^K p(\mathbf{c}_k | \mathbf{x}_i) d(\mathbf{x}_i, \mathbf{c}_k) \quad (3.2)$$

where $\mathbf{X}$ and $\mathbf{C}$ are the collections of input vectors and prototypes respectively. $p(\mathbf{x}_i, \mathbf{c}_k)$ is the joint probability distribution, $p(\mathbf{c}_k | \mathbf{x}_i)$ is the conditional probability. This is a *soft* or *fuzzy* clustering criterion function comparing to the *hard* criterion as in [103].

Minimizing D subject to a specified level of randomness, i.e. Shannon entropy, is reformulated as minimization of Lagrangian

$$F = D(\mathbf{X}, \mathbf{C}) - TH(\mathbf{X}, \mathbf{C}) \quad (3.3)$$

where T is the Lagrange multiplier (or temperature), and the joint entropy [87]

$$H(\mathbf{X}, \mathbf{C}) = H(\mathbf{X}) + H(\mathbf{C}|\mathbf{X}) \quad (3.4)$$

$H(\mathbf{X})$ is the source entropy, and $H(\mathbf{C}|\mathbf{X})$ is the conditional entropy, given by

$$H(\mathbf{C}|\mathbf{X}) = - \sum_{i=1}^l p(\mathbf{x}_i) \sum_{k=1}^K p(\mathbf{c}_k|\mathbf{x}_i) \log p(\mathbf{c}_k|\mathbf{x}_i) \quad (3.5)$$

Minimization of F with respect to the association probabilities $p(\mathbf{c}_k|\mathbf{x}_i)$ gives the Gibbs distribution

$$p(\mathbf{c}_k|\mathbf{x}_i) = \frac{\exp\left(-\frac{d(\mathbf{x}_i, \mathbf{c}_k)}{T}\right)}{\sum_{k=1}^K \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{c}_k)}{T}\right)}. \quad (3.6)$$

Since the conventional DA method can not give an optimal solution for unbalanced data set, so the optimization improvement will be achieved by placing an explicit constraint on the centroids. Let $p(\mathbf{c}_k)$ denotes the fraction of centroids which represent the effective cluster k and are coincident at $\mathbf{c}_k$, then the Gibbs distribution can be reformulated as the titled distribution

$$p(\mathbf{c}_k|\mathbf{x}_i) = \frac{p(\mathbf{c}_k) \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{c}_k)}{T}\right)}{\sum_{k=1}^K p(\mathbf{c}_k) \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{c}_k)}{T}\right)}. \quad (3.7)$$

The denominator is a normalization term. When the temperature is very high ($T \rightarrow \infty$), the conditional probability is uniform, all data points are equally associated with one cluster. As T is lowered, the conditional probabilities become more discriminating. As T is lowered close to zero, it becomes *hard* clustering problem with probability one to the nearest prototype and zero to other prototypes.

The optimization function is focused to maximize the conditional entropy $H(\mathbf{C}|\mathbf{X})$

due to the source entropy $H(\mathbf{X}) = -\sum_{i=1}^l \log \mathbf{p}(\mathbf{x}_i)$ is free from the clustering problem, as

$$F^* = -T \sum_{i=1}^l p(\mathbf{x}_i) \log \sum_{k=1}^K p(\mathbf{c}_k) \exp \left(-\frac{d(\mathbf{x}_i, \mathbf{c}_k)}{T} \right). \quad (3.8)$$

The free parameter $\mathbf{c}_k$ that optimizes the free energy F^* should satisfy

$$\frac{\partial}{\partial \mathbf{c}_k} F^* = 0 \quad (3.9)$$

with the constraint $\sum_k p(\mathbf{c}_k) = 1$, we can obtain [48]

$$\mathbf{c}_k = \frac{\sum_{i=1}^l \mathbf{x}_i p(\mathbf{x}_i) p(\mathbf{c}_k | \mathbf{x}_i)}{p(\mathbf{c}_k)}, \quad (3.10)$$

where

$$p(\mathbf{c}_k) = \sum_{i=1}^l p(\mathbf{x}_i) p(\mathbf{c}_k | \mathbf{x}_i). \quad (3.11)$$

As the temperature is lowered, the system undergoes a sequence of phase transitions which consist of natural cluster splitting. It is a coarse-to-fine algorithm. A cluster undergoes a phase splitting transition when the temperature reaches the critical point. The phase transition procedure can be illustrated in Fig.3.1. Note that six centers of the Gaussian distribution are approximately recovered as the temperature goes down and close to zero.

Figure 3.1: (a) One-dimensional input density of the data samples, the sample data are generated as 6 Gaussian distribution with the identical weights and identical variances, the density is a linear superposition of these 6 Gaussian distributions. (b) Computed cluster splitting and phase transition behavior as temperature T decreasing. Note that the six cluster centers are recovered gradually as temperature goes down.

3.3 Deterministic Annealing Clustering Incorporating Circular Function (DACF)

In a standard DA based image segmentation, the feature vector $\mathbf{x}_i$ present normally intensity (or color) based feature. The dissimilarity measure is defined merely based on Euclidean distances between the feature vectors and centroids without knowledge of the spatial shape information in the objective function. It works well for image segmentation when the image contrast is high with low noise level. The segmentation performance decreases when the image contrast is low. Solely using the intensity or intensity related information will generally create patches wherever the feature of normal breast tissue is close to the feature of mass body. The scattered regions may cause disturbances to the membership allocation and data partition. Large number of subregions in the same cluster that contains mass may result in heavy computation load. In addition to that, it is hard to find the fuzzy boundary when the image contrast is low. Therefore, a new dissimilarity measure for deterministic annealing clustering incorporating circular function (DACF) is proposed.

Consider an image with $M \times N$ pixels, which are located at positions denoted as (i, j) , where $i \in [1, M]$ and $j \in [1, N]$. Let $d_{k,i,j}$ stand for the Euclidean distance between feature vector $\mathbf{x}_{i,j}$ and the centroid $\mathbf{c}_k$ of the k th cluster, as in the original DA method, given as

$$d_{k,i,j} = \|\mathbf{x}_{i,j} - \mathbf{c}_k\|^2 \quad (3.12)$$

Let $f(k, i, j, \mathbf{s})$ represents the shape function, which is incorporated into the dissimilarity measure with a new dissimilarity measure $\hat{d}_{k,i,j}$, as

$$\hat{d}_{k,i,j} = d_{k,i,j} + \beta_k f(k, i, j, \mathbf{s}) \quad (3.13)$$

where β_k are predefined regulated parameter, and $f(k, i, j, \mathbf{s})$ is given as circular function

$$f(k, i, j, \mathbf{s}) = (i - x_c)^2 + (j - y_c)^2 \quad (3.14)$$

where $\mathbf{s} = \{x_c, y_c\}$ is a unique clique, and x_c, y_c are the physical X-Y coordinate of the center of a mass region. The dissimilarity measure $\hat{d}_{k,i,j}$ consists of a measure of the intensity dissimilarity between the (i, j) th pixel and the centroid $\mathbf{c}_k$ in the feature space and the spatial distance between the pixel and the center of the target image. With the inclusion of circular shape information, the pixels with similar intensity but located in disjointed region will be differentiated. The purpose of the inclusion of the shape function is to obtain a large membership for the cluster associated with mass region. In order to do so, the weighting parameter β_k is defined as the weight of the spatial distance against the intensity feature. According to the dissimilarity definition of the Euclidean distance, the closer a pixel belongs to a cluster, the smaller the distance is. Therefore, the shape distance between the location of a

pixel to a specific cluster center is small if the pixel belongs to the cluster, otherwise the distance is larger if it belongs to other clusters.

The expected distortion or objective function of the DACF incorporating spatial information is defined as

$$\hat{D} = \sum_{i=1}^M \sum_{j=1}^N \sum_{k=1}^K p(\mathbf{x}_{i,j}, \mathbf{c}_k) \hat{d}_{k,i,j} \quad (3.15)$$

$$= \sum_{i=1}^M \sum_{j=1}^N p(\mathbf{x}_{i,j}) \sum_{k=1}^K p(\mathbf{c}_k | \mathbf{x}_{i,j}) \hat{d}_{k,i,j} \quad (3.16)$$

$$= \sum_{i=1}^M \sum_{j=1}^N p(\mathbf{x}_{i,j}) \sum_{k=1}^K p(\mathbf{c}_k | \mathbf{x}_{i,j}) (d_{k,i,j} + \beta_k f(k, i, j, \mathbf{s})) \quad (3.17)$$

$$= D + D' \quad (3.18)$$

where D is the distortion measure as in the original DA method (3.2), and D' is the distortion measure of spatial information, as

$$D' = \sum_{i=1}^M \sum_{j=1}^N p(\mathbf{x}_{i,j}) \sum_{k=1}^K p(\mathbf{c}_k | \mathbf{x}_{i,j}) (\beta_k f(k, i, j, \mathbf{s})) \quad (3.19)$$

We recast the optimization problem as seeking the distribution which minimizes $\hat{D}$ subject to a specified level of randomness that is measured by Shannon entropy

$$H(X, C) = - \sum_{i=1}^M \sum_{j=1}^N \sum_{k=1}^K p(\mathbf{x}_{i,j}, \mathbf{c}_k) \log p(\mathbf{x}_{i,j}, \mathbf{c}_k) \quad (3.20)$$

The optimization is reformulated as minimization of the Lagrangian

$$\hat{F} = \hat{D} - TH \quad (3.21)$$

$$= D + D' - TH \quad (3.22)$$

$$= F + D' \quad (3.23)$$

Minimization of the Lagrangian $\hat{F}$ is focused on maximizing the conditional entropy

$$H(C|X) = - \sum_{i=1}^M \sum_{j=1}^N p(\mathbf{x}_{i,j}) \sum_{k=1}^K p(\mathbf{c}_k|\mathbf{x}_{i,j}) \log p(\mathbf{c}_k|\mathbf{x}_{i,j}) \quad (3.24)$$

According to the mass-constrained DA [48] method, minimizing $\hat{F}$ with respect to the probability of $p(C|X)$ leads to titled distribution

$$p(\mathbf{c}_k|\mathbf{x}_{i,j}) = \frac{p(\mathbf{c}_k) \exp(-\hat{d}_{k,i,j}/T)}{Z_{\mathbf{x}_{i,j}}}, \quad (3.25)$$

where the normalized factor is

$$Z_{\mathbf{x}_{i,j}} = \sum_{k=1}^K p(\mathbf{c}_k) \exp(-\hat{d}_{k,i,j}/T). \quad (3.26)$$

Taking the partial derivative on $\hat{F}$ with respect to C , we have

$$\frac{\partial \hat{F}}{\partial C} = \frac{\partial F}{\partial C} + \frac{\partial D'}{\partial C} \quad (3.27)$$

It can be seen that the partial derivative of the objective function with new dissimilarity measure with respect to C is identical to that of DA. Hence, the formula for computing centroids of DACF in the intensity feature space is the same as in DA, i.e.

$$\mathbf{c}_k = \frac{\sum_{i=1}^M \sum_{j=1}^N \mathbf{x}_{i,j} p(\mathbf{x}_{i,j}) p(\mathbf{c}_k|\mathbf{x}_{i,j})}{p(\mathbf{c}_k)}, \quad (3.28)$$

where

$$p(\mathbf{c}_k) = \sum_{i=1}^M \sum_{j=1}^N p(\mathbf{x}_{i,j}) p(\mathbf{c}_k|\mathbf{x}_{i,j}). \quad (3.29)$$

The partial derivative of $\hat{F}$ with respect to $\mathbf{s}$ is given by

$$\begin{aligned}\frac{\partial \hat{F}}{\partial \mathbf{s}} &= \frac{\partial D'}{\partial \mathbf{s}} \\ &= \sum_{i=1}^M \sum_{j=1}^N p(\mathbf{x}_{i,j}) \sum_{k=1}^K p(\mathbf{c}_k | \mathbf{x}_{i,j}) \beta_k \frac{\partial f(k, i, j, \mathbf{s})}{\partial \mathbf{s}}\end{aligned}\quad (3.30)$$

$$= 0 \quad (3.31)$$

i.e.

$$\frac{\partial \hat{F}}{\partial x_c} = \sum_{i=1}^M \sum_{j=1}^N p(\mathbf{x}_{i,j}) \sum_{k=1}^K p(\mathbf{c}_k | \mathbf{x}_{i,j}) \beta_k \frac{\partial f(k, i, j, \mathbf{s})}{\partial x_c} = 0 \quad (3.32)$$

$$\frac{\partial \hat{F}}{\partial y_c} = \sum_{i=1}^M \sum_{j=1}^N p(\mathbf{x}_{i,j}) \sum_{k=1}^K p(\mathbf{c}_k | \mathbf{x}_{i,j}) \beta_k \frac{\partial f(k, i, j, \mathbf{s})}{\partial y_c} = 0 \quad (3.33)$$

Taking (3.14) into (3.32) and (3.33), the spatial parameters are finally obtained as

$$x_c = \frac{\sum_{i=1}^M \sum_{j=1}^N i p(\mathbf{x}_{i,j}) \sum_{k=1}^K \beta_k p(\mathbf{c}_k | \mathbf{x}_{i,j})}{\sum_{i=1}^M \sum_{j=1}^N p(\mathbf{x}_{i,j}) \sum_{k=1}^K \beta_k p(\mathbf{c}_k | \mathbf{x}_{i,j})} \quad (3.34)$$

$$y_c = \frac{\sum_{i=1}^M \sum_{j=1}^N j p(\mathbf{x}_{i,j}) \sum_{k=1}^K \beta_k p(\mathbf{c}_k | \mathbf{x}_{i,j})}{\sum_{i=1}^M \sum_{j=1}^N p(\mathbf{x}_{i,j}) \sum_{k=1}^K \beta_k p(\mathbf{c}_k | \mathbf{x}_{i,j})} \quad (3.35)$$

The titled distribution (3.25) is the membership of each pixel belonging to different clusters. Generally, the intensities of the center part of a mass region are higher than that of locating outside of mass region. For pixels inside a mass region, the intensity dissimilarity is in dominant position, while spatial information plays a major role in dissimilarity measure for pixels outside the mass region. Therefore, the pixels with the same intensity values but locate in different positions in an image will be differentiated, which makes DACF yield better performance for both mass and non-mass related regions.

The implementation of DACF is summarized as follows:

1. Set the maximum number K of clusters, the minimum temperature T_{min} , and cooling rate α .
2. Initialize: $T > 2\lambda_{max}(C_x)$, $k = 1$, $p(\mathbf{c}_1) = 1$, where $\lambda_{max}(C_x)$ is the maximum eigenvalue of the input vector covariance. Initialize the spherical center (x_c, y_c) .
3. Calculating the distance in (3.13).
4. Update $\mathbf{c}_k$, $p(\mathbf{c}_k|\mathbf{x}_{i,j})$, $p(\mathbf{c}_k)$ according to (3.28), (3.25) and (3.29) for $k=1,...,K$.
5. Update x_c and y_c according to (3.34) and (3.35) for $k=1,...,K$.
6. Test whether $p(\mathbf{c}_k|\mathbf{x}_{i,j})$ is convergent, if not satisfied go to 3.
7. Check if $k = K$ or $T \leq T_{min}$, if yes then stop.
8. If $T > T_{min}$, set cooling step: $T \leftarrow \alpha T$.
9. If $k < K$, check condition for phase transition for $k = 1, ..., K$. If the critical temperature T (the maximum eigenvalue of the input vector covariance) is reached for the respective cluster k , add a new term $\mathbf{c}_{k+1} = \mathbf{c}_k + \delta$, $p(\mathbf{c}_{k+1}) = p(\mathbf{c}_k)$, and increase k .
10. Go to 3.

3.4 Experimental Results

There are twenty-two mammograms that contain twenty-four circumscribed masses in MiniMIAS database [44]. Among these masses, four are malignant cases and twenty are benign ones. The mammogram mdb005 and mdb132 each contains two mass regions. The two mass regions in mammogram mdb005 are overlapped, which are processed together as a single one. The two mass regions in mdb132 are processed

independently, as illustrated in Fig.3.2. Hence, twenty-three regions of interest (ROIs) with circumscribed mass are analyzed. Although well-defined masses have high tendency for the benign class, there are still certain mass regions belonging to malignant class. The cost will be very high if a false negative occurs.

Mass segmentation is normally carried out within a ROI, which is taken from the global mammographic image based on the information provided by the database. All the ROIs that contain well-defined masses in the MiniMias database are illustrated in Fig.3.2. Gaussian low pass filter is chosen to phase out noisy points with high frequency in the image preprocessing step.

3.4.1 Segmentation Results

Three clustering methods of DACF, DA and fuzzy c-means(FCM) are implemented for the twenty-three ROI segmentation to show the good performance of DACF algorithm for mammographic mass segmentation. The decision to choose the cluster that contains mass region is based on the assumption that the suspicious mass area is brighter than its surrounding breast tissues, which is valid for most of the real applications [31]. For the illustration purpose, the clustering results are transformed into binary images, where pixels with gray value 255 belong to the suspicious cluster, and pixels with gray value 0 belong to other clusters.

In the experiment, the cluster number is set as two for DACF and two to six for DA and FCM. It is one of the advantages of DACF to use the fixed cluster number (two in our experiment). The predefined weighting parameter $\beta_1 = 1/4$ and $\beta_2 = 5 \sim 60$ are selected proportionally to the complexity of ROIs through experiments. Fig.3.3 shows the segmentation results by the DACF algorithm for all the twenty-three ROIs, where pixels with high intensity (255, for example) belongs to the suspicious cluster. It can be seen that the mass region has the largest area in each ROI.

Therefore, the mass region in each ROI is identified as the one with the maximum number of pixels in the suspicious cluster. Fig.3.4 shows the segmentation results by DA, and the segmentation results by FCM are illustrated in Fig.3.5.

In order to evaluate the segmentation performance, a quantitative technique is applied for the clustering algorithms of mammographic masses. There are several methods to evaluate the quality of image segmentation algorithms which can be cataloged into two groups, i.e. supervised and unsupervised algorithms. Unsupervised evaluation does not depend on a correct segmentation [61]. While in supervised evaluation, the difference between a reference segmentation and the output of a segmentation algorithm is computed. This study chooses the supervised evaluation method and manually draws the boundary of the masses. The probability of segmentation error (PSE) is given by [106]

$$PSE = P(O)P(B|O) + P(B)P(O|B) \quad (3.36)$$

where $P(O)$ and $P(B)$ are the priori probability of the object (mass region) or background (non-mass region) respectively. $P(B|O)$ is the probability of classifying objects as background, and $P(O|B)$ is the probability of classifying background as object. Suppose the pixel number of the mass region in the reference segmentation image is $N_{trueobj}$, and the pixel number of the mass region in the segmented image done by DACF, DA or FCM is N_{calobj} , then the probabilities are defined as

$$P(B|O) = (N_{trueobj} - R_{trueobj} \cap R_{calobj}) / N_{trueobj} \quad (3.37)$$

$$P(O|B) = (N_{calobj} - R_{trueobj} \cap R_{calobj}) / N_{calobj} \quad (3.38)$$

where $R_{trueobj}$ is the region of mass in reference image, and R_{calobj} is the calculated mass region by DACF, DA and FCM. Therefore, $R_{trueobj} \cap R_{calobj}$ represent the number of pixels in the overlapped mass region between the reference image and

calculated image. The computed PSE for DACF, DA and FCM are shown in table 3.1. We can see that DACF performs better than DA algorithm for 15 out of 23 and FCM for 16 out of 23 ROIs, which have comparatively low contrast compared with the rest eight (or seven) images (see the original images in Fig.3.2). Due to the incorporation of spherical shape information, pixels belonging to the same intensity feature cluster while locating in different positions can be differentiated to certain degrees by DACF, as illustrated in the segmentation results of Fig.3.3.

Image	PSE of DACF (%)	PSE of DA (%)	PSE of FCM (%)
mdb001	11.17	10.77	10.21
mdb002	27.14	34.98	34.93
mdb005	7.87	11.37	14.81
mdb010	6.48	6.53	5.94
mdb012	4.93	6.06	6.20
mdb015	8.22	8.82	6.91
mdb017	12.18	39.18	38.13
mdb019	13.77	15.01	30.27
mdb021	9.85	8.64	13.68
mdb023	5.03	8.27	5.18
mdb025	5.88	4.36	4.14
mdb028	11.82	11.64	12.93
mdb069	15.18	24.04	25.35
mdb080	26.72	18.46	21.47
mdb091	13.16	13.76	10.0
mdb132	7.76	10.25	10.6
mdb132	1.82	4.20	2.06
mdb141	12.54	5.04	4.85
mdb142	4.07	12.8	7.05
mdb244	21.07	48.33	53.76
mdb270	8.97	10.59	11.24
mdb290	15.34	16.32	16.19
mdb315	7.08	10.77	8.55
Average of PSE	11.2	14.59	14.98

Table 3.1: The probability of segmentation error (PSE) by DACF, DA and FCM.

3.4.2 Weighting Parameter Analysis

The weighting parameter β in (3.13) controls the influence of the physical distance and the intensity feature in the dissimilarity measure. It is desirable that the membership of a pixel is close to one if it is located near the center of a mass. For a pixel far away from the center of the mass, its membership to the cluster should be close to zero. Suppose that the two clusters in DACF algorithm are cluster 1 and cluster 2, where cluster 1 represents the mass region and cluster 2 represents the non-mass region. For an intensity based algorithm, like standard DA or FCM, the membership of a pixel in non-mass region to cluster 1 may approximately equal to one if it has the same intensity value as the ones in mass region. In this situation, it is hoped that the spatial information will be dominant in the objective function of DACF, i.e. the weighting parameter of cluster 2 should be large enough such that the influence of intensity-based feature is reduced significantly. For the weighting parameter of cluster 1, it is hoped that the β value should be small enough such that the intensity-based feature is dominant in the objective function.

However, it is difficult to analyze the two weighting parameters separately since the incorporation of different values of β_1 and β_2 will lead to different dissimilarity measures. According to the simulation results, we find that the influence of weighting parameter on the segmentation results depends on the image content of each ROI. Experimental results show that better segmentation results are obtained when the value of β_1 is smaller than one, and the value of β_2 is relative large and proportional to the complexity of each ROI.

In order to illustrate the relationship between the values of the weighting parameters and the segmentation results, experiments are carried out on four ROIs that contain mass with different kinds of background tissues. The segmentation results are illustrated from Fig.3.6 to Fig.3.9. The PSE is provided to evaluate the segmentation results under different values of weighting parameters, as shown in table 3.2.

It can be seen that the values of $\beta_1 = \frac{1}{4}$ and $\beta_2 = 10$ are optimal values for these four ROIs.

SE (%)	$\beta_2 = 1$	$\beta_2 = 5$	$\beta_2 = 10$	$\beta_2 = 20$
$\beta_1 = 1$	34.73	18.02	7.14	11.83
$\beta_1 = \frac{1}{2}$	34.56	15.48	7.07	12.09
$\beta_1 = \frac{1}{4}$	34.55	14.59	7.03	12.23
$\beta_1 = \frac{1}{8}$	33.89	14.34	7.04	12.29

Table 3.2: Average segmentation error (PSE) of DACF with different weighting parameter β_1 and β_2 .

3.4.3 Convergence Time Analysis

The main parameter that affects the convergence time in DACF is the weighting parameter due to the cluster number is fixed as two. Through regulating the values of weighting parameters, DACF can obtain good segmentation results with cluster number of two even if the intensities in ROI belong to more than two clusters, as shown in Fig.3.7. For DA and FCM, the relative segmentation results for some ROIs can be obtained with wide range of cluster number from two to six. Larger number of clusters leads to longer convergent time. Table 3.3 illustrates the convergence time of DACF and FCM for the segmentation results in Fig.3.3 and Fig.3.5 respectively. It can be seen that the optimal results are obtained with less convergence time for DACF than FCM.

The convergence speed of DACF is not affected significantly by changing the values of weighting parameters. In order to present a detailed illustration of the relationship between the convergence time of DACF and the values of weighting parameters, experiments have been carried out for the four ROIs in the last subsection. The calculated convergence time with different weighting parameters is shown in table 3.4 to table 3.7. Clearly, we can see that the convergence time decreases or keeps

Image	DACF(sec)	FCM (sec)	Image	DACF (sec)	FCM (sec)
mdb001	10.813	36.703	mdb069	2.625	21.50
mdb002	1.078	9.938	mdb080	2.156	9.625
mdb005	1.500	3.640	mdb091	1.046	3.766
mdb010	0.437	2.297	mdb132	0.812	2.265
mdb012	1.235	2.625	mdb132	1.234	7.937
mdb015	1.703	1.329	mdb141	0.625	9.187
mdb017	1.531	2.750	mdb142	1.016	6.531
mdb019	1.812	16.093	mdb244	2.109	5.078
mdb021	5.343	19.562	mdb270	1.843	4.797
mdb023	1.375	8.672	mdb290	1.328	12.171
mdb025	3.594	8.312	mdb315	21.00	9.421
mdb028	2.250	5.829			

Table 3.3: The convergence time of DACF and FCM for the segmentation results in Fig.3.3 and Fig.3.5. The number of cluster is two for DACF, and two to six for FCM.

approximately constant as the difference between the two weighting parameters increases. Through experiments, it can be seen that the relatively higher difference between the two weighting parameters makes the DACF capable of handling images with more complicated content. Under this situation, larger number of clusters is necessary for both DA and FCM to obtain optimal segmentation results.

3.5 Conclusion

In this chapter, a new DACF clustering algorithm is proposed for mass segmentation based on the deterministic annealing incorporating with circular shape function. It offers two main advantages. First, the ability for mass segmentation is improved. The average probability of segmentation error is 11.2% for DACF, which is smaller than that of DA and FCM, whose errors are 14.59% and 14.98%, respectively. In addition, the number of patched regions in the cluster of mass for DACF is less than that of DA and FCM, as illustrated in Fig.3.3, Fig.3.4 and Fig.3.5. Second,

the convergence time is reduced. In DACF, the cluster number is fixed as two, the optimal segmentation results are obtained by regulating the weighting parameters. Table 3.4 to table 3.7 illustrate that the convergence time is reduced as the difference between the weighting parameters increases. While, in both DA and FCM, the optimal segmentation results are obtained by regulating (or increasing) the cluster number, which leads the convergence time to be increased. Due to the inclusion of the weighting parameters in DACF, the convergence time of twenty-one cases are much less than that of FCM for the segmentation results in Fig.3.3 and Fig.3.5, as illustrated in table 3.3.

In a word, DACF is robust against noisy regions and computational efficient with fixed cluster number and relative large difference of weighting parameters. Unlike classical clustering methods for image segmentation, the objective function of DACF contains both intensity based information and circular shape function as a means to improve the image data partitions. Experimental results show that the proposed DACF works more effective for mammographic image segmentation compared with DA and FCM methods.

Figure 3.2: ROIs that contain masses with circumscribed shape in MiniMIAS database. The two ROIs in mdb132 are processed independently.

Figure 3.3: Segmentation results by DACF for all the ROIs in Fig.3.2.

Figure 3.4: Segmentation results by DA for all the ROIs in Fig.3.2.

Figure 3.5: Segmentation results by FCM for all the ROIs in Fig.3.2.

Time(sec)	$\beta_2 = 1$	$\beta_2 = 5$	$\beta_2 = 10$	$\beta_2 = 20$	$\beta_2 = 30$	$\beta_2 = 50$	$\beta_2 = 60$
$\beta_1 = 1$	3.56	2.82	1.54	1.01	0.87	0.79	0.76
$\beta_1 = \frac{1}{2}$	3.96	2.39	1.39	1.10	0.92	0.79	0.84
$\beta_1 = \frac{1}{4}$	3.93	3.36	1.56	1.03	0.92	0.90	0.73
$\beta_1 = \frac{1}{8}$	4.39	2.50	1.57	1.18	0.82	0.79	0.81

Table 3.4: The convergence time (second) of DACF for mdb002 mass segmentation with different weighting parameter β_1 and β_2 .

Time(sec)	$\beta_2 = 1$	$\beta_2 = 5$	$\beta_2 = 10$	$\beta_2 = 20$	$\beta_2 = 30$	$\beta_2 = 50$	$\beta_2 = 60$
$\beta_1 = 1$	1.97	2.68	1.36	1.03	1.61	1.65	1.01
$\beta_1 = \frac{1}{2}$	1.80	3.03	1.42	1.17	1.96	1.62	1.02
$\beta_1 = \frac{1}{4}$	2.10	2.37	1.25	1.29	1.60	1.42	1.17
$\beta_1 = \frac{1}{8}$	1.56	2.64	1.48	1.11	1.59	1.62	1.15

Table 3.5: The convergence time (second) of DACF for mdb023 mass segmentation with different weighting parameter β_1 and β_2 .

Time(sec)	$\beta_2 = 1$	$\beta_2 = 5$	$\beta_2 = 10$	$\beta_2 = 20$	$\beta_2 = 30$	$\beta_2 = 50$	$\beta_2 = 60$
$\beta_1 = 1$	5.31	4.62	3.36	3.15	2.87	2.71	2.61
$\beta_1 = \frac{1}{2}$	4.04	4.09	3.64	2.78	2.98	2.53	2.61
$\beta_1 = \frac{1}{4}$	3.82	3.79	3.32	2.73	2.50	2.60	2.62
$\beta_1 = \frac{1}{8}$	3.71	3.75	2.92	2.82	2.82	2.59	2.65

Table 3.6: The convergence time (second) of DACF for mdb025 mass segmentation with different weighting parameter β_1 and β_2 .

Time(sec)	$\beta_2 = 1$	$\beta_2 = 5$	$\beta_2 = 10$	$\beta_2 = 20$	$\beta_2 = 30$	$\beta_2 = 50$	$\beta_2 = 60$
$\beta_1 = 1$	14.01	21.30	21.80	12.40	7.53	5.93	5.92
$\beta_1 = \frac{1}{2}$	14.64	23.31	21.31	11.26	7.43	5.89	5.90
$\beta_1 = \frac{1}{4}$	15.11	24.89	20.90	11.26	7.54	6.04	5.87
$\beta_1 = \frac{1}{8}$	14.31	25.70	20.83	11.42	7.48	5.93	5.87

Table 3.7: The convergence time (second) of DACF for mdb315 mass segmentation with different weighting parameter β_1 and β_2 .

Figure 3.6: Segmentation results by DACF with different parameter values of β_1 and β_2 for the case of mdb002. (a) $\beta_1 = 1$, $\beta_2 = 1$. (b) $\beta_1 = \frac{1}{2}$, $\beta_2 = 5$. (c) $\beta_1 = \frac{1}{4}$, $\beta_2 = 10$. (d) $\beta_1 = \frac{1}{8}$, $\beta_2 = 20$.

Figure 3.7: Segmentation results by DACF with different parameter values of β_1 and β_2 for the case of mdb023. (a) $\beta_1 = 1$, $\beta_2 = 1$. (b) $\beta_1 = \frac{1}{2}$, $\beta_2 = 5$. (c) $\beta_1 = \frac{1}{4}$, $\beta_2 = 10$. (d) $\beta_1 = \frac{1}{8}$, $\beta_2 = 20$.

Figure 3.8: Segmentation results by DACF with different parameter values of β_1 and β_2 for the case of mdb025. (a) $\beta_1 = 1$, $\beta_2 = 1$. (b) $\beta_1 = \frac{1}{2}$, $\beta_2 = 5$. (c) $\beta_1 = \frac{1}{4}$, $\beta_2 = 10$. (d) $\beta_1 = \frac{1}{8}$, $\beta_2 = 20$.

Figure 3.9: Segmentation results by DACF with different parameter values of β_1 and β_2 for the case of mdb315. (a) $\beta_1 = 1$, $\beta_2 = 1$. (b) $\beta_1 = \frac{1}{2}$, $\beta_2 = 5$. (c) $\beta_1 = \frac{1}{4}$, $\beta_2 = 10$. (d) $\beta_1 = \frac{1}{8}$, $\beta_2 = 20$.

Chapter 4

Robust Information Clustering for Automatic Mass Detection

4.1 Introduction

The shapes and margins of masses have complex variations. As indicated in chapter 3, the circumscribed masses have relative well-defined margins and close to round shapes. Masses with irregular, spiculated margins, architectural distortion and asymmetry in shapes have high tendency to be malignant [75], which are even hard to be detected, especially when they are surrounded by dense parenchymal breast tissues with pretty low contrast. Moreover, the digitized mammographic images are not clean in terms of noises and artifacts which can lead the segmentation results deviated from the optimal partitions.

It is true that the image segmentation needs robust algorithms, which implies that the performance of an algorithm should not be affected significantly by small deviations from the assumed model and it should not deteriorate drastically due to

noise and outliers [92]. Recent years have seen increasing interests in developing robust clustering algorithms for the applications in computer vision [43] [32], image segmentation [117] [18] [51] and pattern recognition [4]. Many robust algorithms have been proposed for fuzzy clustering and a comprehensive review can be found in [92]. Dave [91] proposed robust fuzzy clustering algorithm by introducing the concept of noise cluster in addition to c clusters. A noise prototype is defined such that it is equi-distant to all the points in the data set, i.e. the noise cluster is determined based on distance measure. In the same year, Krishnapuram [84] published a probabilistic approach for c -means clustering. The membership of each data is a probabilistic membership. A data can be considered as an outlier if its memberships to all these clusters are small. Many other robust clustering algorithms can be found in the literature, such as [83] [56]. Wu *et al* [47] proposed a new metric to replace the Euclidean norm in c -means clustering procedure which is robust to noisy data and outliers. The more recent work of robust clustering is presented in [45]. This approach identifies and prunes the outliers based on the sizes and shapes of the clusters. It adopts c thresholds to distinguish the noisy points from the c clusters.

In this chapter, a robust information clustering (RIC) algorithm is proposed for automatic mammographic mass detection for the concern of mammographic screening. Pixels on the fuzzy margin of a mass and noisy data are identified by the RIC based on minimax of the mutual information. Similar to the general statistical learning theory in [113], the RIC algorithm is considered as a machine learning problem with the limited training data rather than based on the infinite number of training data. It is a two-phase minimax optimization procedure. Through a minimizing mutual information scheme and starting from one, the number of clusters is increased gradually, which minimizes the free energy function, known as the Jaynes's maximum entropy principle [48], from high to low temperature. Starting from an initial single cluster of all input patterns, a new cluster is added each time as the temperature is lowered to a critical point. The robust estimation phase is carried out to maximize

the mutual information and estimate the outliers. The algorithm is robust in the sense that both peak and valley of image intensity histogram are estimated by the RIC algorithm simultaneously.

In order to achieve a high detection rate for the automatic mass detection in mammograms, a bottom-up image interpretation strategy is employed, which includes three major steps: preprocessing, segmentation and false positive (FP) reduction. In the preprocessing step, a gain adaptive contrast enhancement (GACE) algorithm is proposed to enhance the contrast of the mammograms. In the second step, the raw regions of interest (ROIs) are first extracted from global mammograms by two steps of adaptive thresholding. Then RIC algorithm is applied for further segmentation. The segmentation scheme is based on the assumption that masses are brighter than their surrounded breast tissues. In the FP reduction step, pixels number is accounted for each subregion in the suspicious cluster resulted from the RIC segmentation, and a binary decision algorithm is used as an area filter to reduce the number of FPs.

Instead of constructing an artificial system as a high-tech radiologist, the objective of this research intends to serve as the first vision in the process of mammographic screening. It aims to locate the suspicious regions of mass candidates in the mammograms that will be further examined by other diagnosis techniques or by radiologists. The detection method has been verified with 54 mammograms in the MIAS database. Each mammogram contains at least one mass case that is embedded in glandular or dense glandular breast tissues. The verification results of the proposed algorithm and the detection system are presented.

As illustrated by the flowchart in Fig.4.1, the detection procedure of the proposed technique consists of the following stages: GACE enhancement, raw ROI extraction, RIC segmentation and FP reduction. RIC clustering algorithm is used here as a technique to partition the entire ROI into distinctive meaningful regions based on

the statistical probability distribution of intensity feature.

The rest of the chapter is organized as follows. Section 4.2 describes the GACE method. Section 4.3 illustrates the two steps of adaptive thresholding technique for raw ROIs extraction. The proposed RIC algorithm is described in section 4.4. Section 4.5 gives the experimental results. Section 4.6 is the discussion, and the conclusion comes finally.

Figure 4.1: Flowchart of the automatic mass detection system for digitized mammograms.

4.2 Gain Adaptive Contrast Enhancement

Although the intensity of mass region is relative high and brighter than its surrounding background, while it is still a challenging and hard task to detect mass due to the complicated background structure and low contrast of the mammographic image, particularly the signal-to-noise is pretty low around the boundary of mass region. As a help of improving the masses detection performance, a gain adaptive contrast enhancement (GACE) technique is first proposed to preprocess the image before image segmentation. The block diagram of GACE is given in Fig.4.2.

Figure 4.2: The block diagram of the GACE preprocessing procedure, where $I_{CO}(x, y)$ is the output of GACE.

The GACE is an expansion of the density-weighted contrast enhancement filter of N. Petrick [69]. The aim of GACE in this study is to enhance global mammograms and local ROIs that contain mass body according to the information provided by the database. The original image is denoted as $I(x, y)$, which will first pass a low pass Gaussian filter in order to remove certain noises.

$$I_G(x, y) = G(x, y) * I(x, y) \quad (4.1)$$

where $G(x, y)$ is a Gaussian smoothing filter with zero mean and standard deviation, $\sigma = 0.6$ is used in this study. Then the image will be scaled between 0.0 and 1.0, called normalized image $I_N(x, y)$. Each pixel in the normalized image will multiply a gain factor which adjusts the intensity at each pixel in the final image

$$I_C = K_{IA}(x, y) \times I_N(x, y) \quad (4.2)$$

The essence of the GACE technique is the definition of $K_{IA}(x, y)$, which is a regulation parameter according to each pixel. GACE algorithm aims to enhance the image contrast, particularly for mass body and boundary where the intensities are high. The gain factor will be low if the intensity of a pixel is low. Otherwise the gain factor will be high if the intensity of a pixel is high, i.e. the gain value to each pixel is adaptive according to its intensity. Inspired by the sloping rate is increasing of quadratic equation, the gain factor here is defined as

$$K_{IA}(x, y) = I_N(x, y) \times I_N(x, y) \quad (4.3)$$

Figure 4.3: Plot of the gain value function of GACE filter. It is a positive quadratic function, where the slop is increasing.

The plot of gain factor (K_{IA}) curve is shown in Fig.4.3. The application result

of GACE algorithm for a local region in a mammogram is shown in Fig.4.4. The histogram of the original region and the histogram of the region after GACE are given in Fig.4.5. It can be seen that the image histogram with GACE enhancement expands and presents the image content more detailed.

Figure 4.4: Original and the enhanced images by GACE. a) Original image. b) Enhanced image by GACE.

Figure 4.5: Histograms of the original and enhanced images. a) Histogram of original image. b) Histogram of the image enhanced by GACE.

4.3 Raw ROI Extraction

The RIC is applied within certain raw ROI. In this study, two steps of adaptive thresholding are proposed for ROI extraction. There are two main areas in a mammographic image: breast region and non-breast region. In general, the non-breast region of the digitized magmmograms has a very low intensity level and a maximum peak in the histogram. The intensities of the breast region, however, may vary with different patients and depend on the mammography conditions. Particularly, the background objects with high intensity will affect the extraction result of ROI. Therefore, the breast region is first extracted from digitized mammograms by adaptive thresholding technique. The threshold level for the breast region is determined as the mean of the intensity of the window around a pixel. Considering the processing speed, a window of 4×4 is chosen instead of each pixel. Some images in MiniMIAS database contain certain background objects with intensity higher than the breast region, even higher than mass region, as shown in Fig.4.6. If they are segmented together with the breast region, they can deteriorate the results of raw ROI extraction and final segmentation. Due to the breast region is the major region in the mammogram that occupies the largest area compared with the regions of other segmented objects, the segmented region with largest area is selected as breast region.

Based on breast region extraction, the second step of adaptive thresholding is carried out for the extraction of ROI. The threshold for each pixel in the breast region is computed as $Threshold = mean + \frac{std}{factor}$, where std represents the standard deviation. The $mean$ and standard deviation are computed for all the pixels in the breast region. A $factor$ of 10.0 is chosen empirically. There are a few sub-regions after the second step of adaptive thresholding, including some isolated points and very small regions, which may not be part of a mass. These subregions will be dilated around their boundary for eight nearest neighborhoods and then pass through an

Figure 4.6: Illustration of components in a mammogram. There are two major areas in a mammogram, breast region and non-breast region. The breast region contains pectoral muscle, breast tissues or mass. The non-breast region contains dark background. For some mammograms there are some background objects with high intensities, as indicated in the figure.

area filter. The sub-regions with areas larger than certain threshold will be remained as part of the interested region for further processing, otherwise they will be removed. A raw ROI is formulated finally (see Fig.4.7), based on which RIC is used for further segmentation.

4.4 The Proposed Robust Information Clustering (RIC) Algorithm

4.4.1 Basics of Information Theory

Suppose that there are two random data sets $\mathbf{x}_i \in \mathbf{X}$ $i = 1, \dots, l$ and $\mathbf{w}_k \in \mathbf{W}$ $k = 1, \dots, K$ with the same dimension, where $\mathbf{x}_i$ represent the training data points and $\mathbf{w}_k$ represent the code vectors in the feature space. K is the maximum number of pattern partitions. Corresponding to information theory, $\mathbf{X}$ is the information source and $\mathbf{W}$ is the prototypes, or the transmission channel. Suppose the probability distributions of $\mathbf{x}_i$ and $\mathbf{w}_k$ are $p(\mathbf{x}_i)$ and $p(\mathbf{w}_k)$ respectively, and conditional probability distributions are $p(\mathbf{x}_i|\mathbf{w}_k)$ and $p(\mathbf{w}_k|\mathbf{x}_i)$. The mutual information $I(\mathbf{X}, \mathbf{W})$, introduced by Shannon, is the fundamental concept of evaluating rate-distortion function and channel capacity in information theory. It can be written in the form of Kullback-Leibler [38] measure, i.e.,

$$I(\mathbf{X}, \mathbf{W}) = \sum_{i=1}^l \sum_{k=1}^K p(\mathbf{x}_i, \mathbf{w}_k) \ln \frac{p(\mathbf{x}_i, \mathbf{w}_k)}{p(\mathbf{x}_i)p(\mathbf{w}_k)} \quad (4.4)$$

Since

$$p(\mathbf{x}_i, \mathbf{w}_k) = p(\mathbf{x}_i)p(\mathbf{w}_k|\mathbf{x}_i), \quad (4.5)$$

and

$$p(\mathbf{w}_k) = \sum_{i=1}^l p(\mathbf{x}_i)p(\mathbf{w}_k|\mathbf{x}_i), \quad (4.6)$$

the mutual information is obtained as

$$I(\mathbf{X}, \mathbf{W}) = \sum_{i=1}^l \sum_{k=1}^K p(\mathbf{x}_i) p(\mathbf{w}_k | \mathbf{x}_i) \ln \frac{p(\mathbf{w}_k | \mathbf{x}_i)}{\sum_{i=1}^l p(\mathbf{x}_i) p(\mathbf{w}_k | \mathbf{x}_i)}. \quad (4.7)$$

Mutual information is related to entropy by the equations

$$I(\mathbf{X}, \mathbf{W}) = H(\mathbf{X}) + H(\mathbf{W}) - H(\mathbf{X}, \mathbf{W}) \quad (4.8)$$

$$= H(\mathbf{X}) - H(\mathbf{X} | \mathbf{W}) \quad (4.9)$$

$$= H(\mathbf{W}) - H(\mathbf{W} | \mathbf{X}) \quad (4.10)$$

where $H(\mathbf{X})$ is the entropy of source information. It is a measure of the amount of uncertainty about the random source viable $\mathbf{X}$. $H(\mathbf{W})$ is the entropy of target information. $H(\mathbf{X}, \mathbf{W})$ is the joint entropy. $H(\mathbf{X} | \mathbf{W})$ and $H(\mathbf{W} | \mathbf{X})$ are the conditional entropy of $\mathbf{X}$ given $\mathbf{W}$ and $\mathbf{W}$ given $\mathbf{X}$. It is the amount of uncertainty left in $\mathbf{X}$ when knowing $\mathbf{W}$ or vise verse. $I(\mathbf{X}, \mathbf{W})$ in (4.9) is the reduction in the uncertainty of the source information by knowing of code vector $\mathbf{W}$. The entropies are expressed in terms of probability by equations

$$H(\mathbf{X}) = - \sum_{i=1}^l p(\mathbf{x}_i) \ln p(\mathbf{x}_i) \quad (4.11)$$

$$H(\mathbf{X}, \mathbf{W}) = - \sum_{i=1}^l \sum_{k=1}^K p(\mathbf{x}_i, \mathbf{w}_k) \ln p(\mathbf{x}_i, \mathbf{w}_k) \quad (4.12)$$

$$H(\mathbf{X} | \mathbf{W}) = - \sum_{i=1}^l \sum_{k=1}^K p(\mathbf{x}_i | \mathbf{w}_k) \ln p(\mathbf{x}_i | \mathbf{w}_k) \quad (4.13)$$

For the image feature $\mathbf{x}_i$ (scalar if only intensity) of a pixel in ROI in a mammogram, image segmentation requires that a decision be made about the type of the pixel, i.e. which intensity groups (or code vectors) it belongs to. It is an observation at

each pixel and describes the relation between source vector $\mathbf{x}_i$ and pattern class $\mathbf{w}_k$ (scalar if intensity only), which can also be stated as partitioning a given set of pixels into subgroups, each of which should be as homogeneous as possible within group, and should be as inhomogeneous as possible between groups. It is an unsupervised clustering problem. The dissimilarity measure between the information source and target patterns is stated by the least mean square distance in Euclidean space

$$\min_{k,i} d(\mathbf{x}_i, \mathbf{w}_k) = \frac{1}{2} \|\mathbf{w}_k - \mathbf{x}_i\|^2 \quad (4.14)$$

Based on the least mean square measure, some of the popular clustering algorithms have been developed like the basic k-means, fuzzy c-means and improved robust versions, like robust noise [92] and possibility clustering algorithm [67]. According to Huber [76], a robust procedure can be characterized as first, it should have a reasonably good accuracy at the assumed model; second, small deviations from the model assumptions should impair the performance only by a small amount; and third, larger deviations from the model assumptions should not cause a catastrophe. The first requirement is the consideration whether data are clean. It plays great importance in statistics and engineering since systems have to be designed such that their performance yields accurate estimation. The second property gains much attentions since the data is not clean in engineering practice, i.e. small deviations should not have a significant negative effect on the performance. The third requirement is related to the concept of breakdown point. According to these properties, the most important one is the robustness to initialization, namely, stable performance. Another common scheme of robustness is to reject or ignore a subset of the input patterns by evaluating the membership function, which can be viewed as a robust weight function to phase out outliers. The robust performance of the fuzzy algorithms in [91] is explained in a sense that it involves the memberships of all input patterns of a given cluster, but not the membership of a particular pattern

in all clusters.

The ROI normally contains subregions of fat, parenchymal tissues or masses. Each region corresponds to one group of intensity level. The mean intensity of each group varies according to their surrounding tissues. Under some conditions, the mean intensity of parenchymal tissue is higher than the mean intensity of the mass region in the same mammogram. The proposed method to segment mass region is based on the assumption that the mean intensity of mass region is brighter than its surrounding tissues, but the fuzzy characters make it hard to find the boundary of a mass. Moreover, artifacts or noisy intensities may deviate the target patterns from the optimal positions, therefore, robust clustering algorithm is required for the segmentation of masses.

4.4.2 Robust Information Clustering Algorithm

Considering the randomized partition of the expected distortion

$$D(\mathbf{X}, \mathbf{W}) = \sum_{i=1}^l \sum_{k=1}^K p(\mathbf{x}_i) p(\mathbf{w}_k | \mathbf{x}_i) d(\mathbf{x}_i, \mathbf{w}_k) \quad (4.15)$$

Following the information theory, the optimal objective function is defined based on the objective function [88]

$$F = D(\mathbf{X}, \mathbf{W}) + TI(\mathbf{X}, \mathbf{W}) \quad (4.16)$$

$$\begin{aligned} &= \sum_{i=1}^l \sum_{k=1}^K p(\mathbf{x}_i) p(\mathbf{w}_k | \mathbf{x}_i) d(\mathbf{x}_i, \mathbf{w}_k) + TI(\mathbf{X}, \mathbf{W}) \\ &= \sum_{i=1}^l \sum_{k=1}^K p(\mathbf{x}_i) p(\mathbf{w}_k | \mathbf{x}_i) d(\mathbf{x}_i, \mathbf{w}_k) + T(H(\mathbf{X}) - H(\mathbf{X} | \mathbf{W})) \end{aligned} \quad (4.17)$$

where $p(\mathbf{w}_k | \mathbf{x}_i)$ and $p(\mathbf{x}_i)$ are the conditional probability mass function *pmf* and unconditional *pmf*, respectively. In general, clustering algorithm (4.15) is a soft

dissimilarity measure among different clusters based on statistic theory. Since $H(\mathbf{X}) = -\sum p(\mathbf{x}) \log p(\mathbf{x})$ is the source entropy, which is independent of clustering, therefore, minimization of (4.15) with respect to the free *pmf* parameter $p(\mathbf{w}_k|\mathbf{x}_i)$ leads to the deterministic annealing (DA) clustering algorithm (4.16) [48]. The DA recasts the clustering optimization problem that of seeking the distribution *pmf* which minimizes (4.14) subject to a specified level of randomness, which can be measured by the mutual information. It is a global optimization clustering algorithm in terms of insensitive to the volume of input patterns in the respective clusters, which avoids the local minima of the clustering like k-means and provides natural clusters in the annealing procedure.

To gain robustness, we focus on the property of density estimation $p^*(\mathbf{x}_i) \in p^*(\mathbf{X})$. Instead of the fixed equal distribution $p(\mathbf{x}_i) = 1/l$ in the DA algorithm, maximization of the mutual information is considered to evaluate different distributions of all input patterns of a given cluster, in particular, seeking the outliers with $p^*(\mathbf{x}_i) = 0$ and eliminating them from the data set for the final data partitions. According to information theory, the mutual information can also be written in the following form

$$I(\mathbf{X}, \mathbf{W}) = \max_{p^*(\mathbf{x}_i|\mathbf{w}_k)} \sum_{i=1}^l \sum_{k=1}^K p^*(\mathbf{x}_i) p(\mathbf{w}_k|\mathbf{x}_i) \ln \frac{p(\mathbf{x}_i|\mathbf{w}_k)}{p^*(\mathbf{x}_i)} \quad (4.18)$$

The capacity of a channel (or prototypes $\mathbf{w}_k$) expresses the maximum rate at which information can be reliably conveyed by the channel, defined as [80]

$$C = \max_{p^*(\mathbf{x}_i)} I(\mathbf{X}, \mathbf{W}) \quad (4.19)$$

$$= \max_{p^*(\mathbf{x}_i)} \max_{p(\mathbf{x}_i|\mathbf{w}_k)} \sum_{i=1}^l \sum_{k=1}^K p^*(\mathbf{x}_i) p(\mathbf{w}_k|\mathbf{x}_i) \ln \frac{p(\mathbf{x}_i|\mathbf{w}_k)}{p^*(\mathbf{x}_i)} \quad (4.20)$$

According to the *Theorem 1* in [80], for a given $p(\mathbf{w}_k|\mathbf{x}_i)$, the capacity is maximized

by

$$p^*(\mathbf{x}_i) = \frac{\exp \left(\sum_{k=1}^K p(\mathbf{w}_k|\mathbf{x}_i) \ln p(\mathbf{x}_i|\mathbf{w}_k) \right)}{\sum_{i=1}^l \exp \left(\sum_{k=1}^K p(\mathbf{w}_k|\mathbf{x}_i) \ln p(\mathbf{x}_i|\mathbf{w}_k) \right)}, \quad (4.21)$$

where $p(\mathbf{w}_k|\mathbf{x}_i)$ is the titled (or conditional probability) distribution obtained in the DA step, and the backward transition probability $p(\mathbf{x}_i|\mathbf{w}_k)$ is obtained through the Bays formula

$$p(\mathbf{x}_i|\mathbf{w}_k) = \frac{p^*(\mathbf{x}_i)p(\mathbf{w}_k|\mathbf{x}_i)}{\sum_{i=1}^l p^*(\mathbf{x}_i)p(\mathbf{w}_k|\mathbf{x}_i)} \quad (4.22)$$

Note that the mutual information $I(\mathbf{X}, \mathbf{W})$ is not negative. However, the individual item in the sum of the capacity can be negative. If the i th pattern $\mathbf{x}_i$ is taken into account and $p(\mathbf{w}_k|\mathbf{x}_i) < \sum_{i=1}^l p^*(\mathbf{x}_i)p(\mathbf{w}_k|\mathbf{x}_i)$, then the probability of the k th code vector (mass center) is decreased by the observed pattern and gives a negative information about pattern $\mathbf{x}_i$. Then the particular input pattern may be considered as an unreliable pattern (outlier) and its negative effect must be off set by other input patterns.

Therefore, the maximization of the mutual information provides a robust density estimation of the noisy pattern (outlier) in term that the average information is over all clusters and input patterns. The robust density estimation and optimization is now to maximize the mutual information against the *pmf* $p^*(\mathbf{x}_i)$ and $p(\mathbf{x}_i|\mathbf{w}_k)$, i.e, the maximization of mutual information in (4.19) is a double maximization for *a priori* $p^*(\mathbf{x}_i)$ via $p(\mathbf{x}_i|\mathbf{w}_k)$ in (4.20) [80]. For any value of i , if $p(\mathbf{x}_i|\mathbf{w}_k) = 0$, then $p^*(\mathbf{x}_i)$ should be set equal to zero in order to obtain the maximum, such that a corresponding training pattern $\mathbf{x}_i$ can be deleted and dropped from further consideration in the optimization procedure as an outlier.

The RIC is basically a two-phase iteration clustering algorithm, i.e. deterministic

annealing clustering to minimize the objective function (4.14) and the robust estimation of the outliers to maximize (4.16). For more details, comprehensive illustration and proof are in [78] [80]. The implementation of RIC is summarized as follows.

1. Do DA clustering [48], obtain the optimal title distribution of $p(\mathbf{w}_k|\mathbf{x}_i)$.
2. Initialize the prior probability of $p^*(\mathbf{x}_i) = 1/l$, where l is the number of the input patterns.
3. Set

$$c_i = \exp \left(\sum_{k=1}^K p(\mathbf{w}_k|\mathbf{x}_i) \ln \frac{p(\mathbf{w}_k|\mathbf{x}_i)}{\sum_{i=1}^l p^*(\mathbf{x}_i) p(\mathbf{w}_k|\mathbf{x}_i)} \right) \quad (4.23)$$

4. If

$$\ln \sum_{i=1}^l p^*(\mathbf{x}_i) c_i - \ln \max_{i=1, \dots, l} c_i < \varepsilon \quad (4.24)$$

go to step 5, where $\varepsilon > 0$, otherwise, set

$$p^*(\mathbf{x}_i) = p^*(\mathbf{x}_i) \frac{c_i}{\sum_{i=1}^l p^*(\mathbf{x}_i) c_i} \quad (4.25)$$

go to step 3.

5. The input patterns with $p^*(\mathbf{x}_i) = 0$ are estimated as outliers.
6. Do DA clustering for input patterns of all $\mathbf{x}_i$ with $p^*(\mathbf{x}_i) > 0$, stop.

The $\varepsilon > 0$ is an accuracy control parameter. At the final point of estimation, pixels will be deemed as outliers whose intensity probability $p(\mathbf{x}_i) = 0$. Until now all the pixels in ROI belong to two clusters: outliers and non-outliers. Label zero is assigned to pixels that belong to outliers, and label one to pixels that belong to the

non-outliers. Then the cluster centers and memberships of pixels with label one will be recalculated by minimization of mutual information.

4.5 Experimental Results

In this section, the experimental results for breast mass detection are presented. A testing data set of 54 mammograms are used to test and evaluate the performance of the proposed algorithms. Each of the mammograms contains at least one mass surrounded by glandular or dense glandular breast tissues. The total data set includes 30 benign and 24 malignant masses.

4.5.1 The Performance of RIC for Outliers Estimation and Mammographic Image Segmentation

Before applying RIC for image segmentation, let us first explain how RIC behaves and performs for outliers estimation. Fig.4.8 shows a simulated data set with two clusters that are overlapped. The points around the overlapped region between the two clusters may be misclassified due to the distances to the two cluster centers are not differentiated compared with other data points to the two cluster centers. The misclassification can make the cluster centers deviate from the optimal positions. As a result, more data points can be clustered wrongly. It is ideal to recognize these points and leave them out for further clustering. Incidentally, the proposed RIC algorithm has the inherent property to identify these noisy and boundary points, as illustrated in Fig.4.8(b). On the contrary, the standard clustering algorithms, such as DA, have no such character. Fig.4.8(c) shows the clustering results by standard DA method. Clearly, certain data in the narrow band of the overlapped region between the two clusters are misclassified.

An example of RIC for outliers identification and image segmentation for mammographic image is provided in the following paragraph and shown in Fig.4.9. The maximum cluster number is chosen as two. Assume there are three true clusters after RIC clustering, namely, normal breast tissues, mass body, and the outlier cluster of the regions between the two clusters. After RIC clustering, the intensity of pixels belonging to the cluster of outliers are set to zero, and the intensities of the pixels in the other two clusters are set to 255. Fig.4.9 shows the outlier estimation and image segmentation for mammographic image. Fig.4.9(a) is the original ROI that contains mass. Its corresponding histogram is illustrated in Fig.4.9(d). During RIC clustering, the whole image data is first partitioned into two clusters: outliers and non-outliers data. As shown in Fig.4.9(b), the estimated outliers are represented by intensity value of zero, i.e. dark points in the image. Where bright points are the normal data points. It can be seen that the outliers represent pixels on the boundary around the mass region and some disconnected noisy points. The three clusters after RIC clustering, i.e. outliers, normal breast tissue and mass region, are illustrated in Fig.4.9(c). The outlier-free image data are partitioned into two clusters, each of which has distinctive intensity. The histogram of the original image with outliers removed by RIC algorithm is illustrated in Fig.4.9(e). Fig.4.9(f) shows the segmentation result by DA.

Fig.4.9 demonstrates that a portion of gray levels in the proximity of the valley between the two peaks in histogram are identified automatically by RIC, rather than one by the standard clustering method. Because of the estimation of outliers, the suspicious cluster that contains mass has less number of subregions than that obtained by standard DA, which will reduce the complexity of further analysis.

4.5.2 FP Reduction

FP reduction is performed during the two steps of segmentation, i.e. adaptive thresholding and RIC, due to many subregions are segmented together with ROIs and masses, which may lead further analysis more complex. Following each step of segmentation, region growing is performed for each subregion that belongs to the suspicious group, and the area of each subregion is calculated. In the first step of adaptive thresholding, the object with the largest area is selected as breast region since it occupies the major structure in the mammogram. The raw ROI is extracted within the segmented breast structure, in which two main regions that may contain mass have high intensity: the pectoral muscle and the glandular breast tissues. The pectoral muscle represents a predominant density region in most medio-lateral oblique (MLO) views. The segmentation results will be affected or biased if it is included in ROI for intensity-based image processing methods. Ferrari *et al* [82] proposed a new method for the identification of the pectoral muscle based upon a multi-resolution technique using Gabor wavelets. In this study, the pectoral muscle is identified as the largest object among the first ten subregions with area larger than 200 pixels according to the searching method. Raw ROI is thus formulated based on the pectoral muscle and noise points removal.

After RIC segmentation, all the pixels in ROI are grouped into K clusters. The mean intensities of each cluster are different, where the mean intensity of the cluster that may contain mass has the highest value (according to assumption), which is called suspicious cluster. There will be some subregions or isolated points in the suspicious cluster. The final suspicious cases are identified as the subregions with the area larger than certain threshold. According to the information of the database, the minimum radius of the mass region is 17 pixels. An area of 200 pixels is chosen as the threshold for the FP reduction in order to ensure that no mass will be lost during this stage.

4.6 Detection Results

The purpose of final FP reduction is to increase the detection rate. For the complicated diagnosis system, it can also decrease the computation complexity and allows more computationally intensive classifiers to be applied to as few objects as possible. In this step, a simple area threshold is used to filter subregions with area less than 200 pixels in the suspicious cluster. Corresponding to the target of mammography screening, the purpose of this study is to locate the suspicious areas that may contain mass, therefore, the detection results are evaluated by terms of sensitivity and the number of false positives per image.

For the purpose of comparison, both DA and FCM for mass segmentation were also implemented based on the same detection system as RIC, i.e. the image segmentation by RIC, DA and FCM are based on the same ROIs. By regulating the parameters of threshold in the step of ROI selection and the cluster number in image data clustering, different detection results are obtained. In the experiments, the number of cluster for RIC is regulated from two to three, and the number of cluster for DA and FCM is from two to six.

Among the 54 mammograms, thirty masses are embedded in glandular breast tissues, and twenty-four masses are within dense glandular breast tissues. The numbers of suspicious subregions that are considered as mass candidates for the sampled mammograms are listed in table 4.1. The detected suspicious ROIs in the four images, mdb124, mdb125, mdb163 and mdb193, contain no true positive (TP) for these three methods, i.e. the number of true positive is the same for RIC, DA and FCM. While the number of FPs are quite different. From this table we can see that RIC based detection system generates quite less number of FPs for 45 images compared with the detection results by DA and FCM. Only four images, mdb083, mdb099, mdb188 and mdb199, have a little higher number of FPs for RIC than DA

Mammogram	RIC	DA	FCM	Mammogram	RIC	DA	FCM
mdb001	1	5	5	mdb117	3	3	3
mdb002	1	1	1	mdb120	2	3	4
mdb013	7	7	7	mdb121	2	1	2
mdb015	8	12	12	mdb127	4	7	7
mdb017	3	4	4	mdb145	3	5	4
mdb019	7	6	11	mdb165	1	1	1
mdb021	7	8	7	mdb170	2	2	3
mdb023	4	8	7	mdb171	1	2	2
mdb030	10	15	14	mdb175	1	2	1
mdb032	10	17	10	mdb178	1	2	2
mdb058	4	7	7	mdb179	1	2	1
mdb063	2	3	2	mdb181	1	1	1
mdb072	2	2	2	mdb186	1	4	4
mdb081	1	1	1	mdb188	2	1	1
mdb083	9	6	10	mdb190	1	1	1
mdb090	5	5	6	mdb191	3	5	6
mdb099	14	9	10	mdb198	2	5	5
mdb102	3	3	3	mdb199	7	5	6
mdb104	5	5	4	mdb202	1	7	7
mdb105	4	3	4	mdb207	1	1	1
mdb107	6	7	7	mdb244	4	5	5
mdb110	3	3	3	mdb264	1	1	1
mdb111	2	3	3	mdb265	3	4	3
mdb115	1	1	2	mdb315	1	1	1
mdb163	2	5	5	mdb270	9	8	9
mdb193	2	8	7	mdb290	1	1	1
mdb124	2	3	2	mdb125	4	4	4

Table 4.1: The detection results of the number of suspicious regions per image.

and FCM. The proposed detection system incorporating RIC algorithm obtains a sensitivity of 91.7% at the false positive of 2.56 per image. The detection result of the detection system incorporating with DA is 3.56 FPs per image and 3.59 FPs per image for the incorporation of FCM.

In order to illustrate the the systematic comparison based on RIC, DA and FCM methods, the free-response ROC (FROC) is given as shown in Fig.4.10. From this figure, we can see that at the same TP fraction, the number of FPs by RIC is less

than that of DA and FCM. From another point of view, at the same number of FPs per image, the TP fraction by RIC is higher than that being produced by DA and FCM. When the number of false positives is larger than certain value, as 3.5 in this study, the TP fraction of these three methods are the same. Due to this chapter focuses on image segmentation instead of further suspicious ROI analysis, thus the FROC curve within the region that the number of FPs per image is less than one is not included.

Because of the estimation of outliers, RIC can segment mass body from its background even when it is embedded in the dense glandular breast tissue with relatively less number of FPs compared with DA and FCM methods, i.e. RIC has robust ability in data splitting and estimating the noisy points. In some cases, RIC can separate even close image data, like the case in mdb099, which is the reason that more FPs are generated. The experiments also observed that the segmented mass region by DA and FCM includes large area of background breast tissues which leads to the number of FPs is relatively less in some images. Based on the experimental analysis we can see that the detection system incorporating RIC works more effective in false positive reduction than the systems based on DA and FCM.

The detection results by RIC for four different kind of masses embedded in glandular or dense glandular breast tissues are illustrated in Fig.4.11 and Fig.4.12. Fig.4.11 shows the final FP reduction results of the case mdb181, which contains spiculated mass within glandular breast tissue. Fig.4.12 is the detection results of the case mdb198 that has spiculated mass in dense breast tissue.

4.7 Discussion

The two steps of adaptive thresholding are found effectively for the extraction of raw ROIs that may contain masses in mammograms. Some problems are observed

during the ROI extraction because of the highly complicated conditions of digitized mammograms. The second step of adaptive thresholding is to find ROI within breast structure, where the dense pectoral muscle is a problem faced by most image segmentation techniques for mass detection. In this study, it leads to a relatively high threshold value, which results in the missing of certain subregions that may contain mass with lower intensity. In addition, the pectoral muscle is much denser than the mass. The mass region may also be missed during suspicious cluster selection if the breast muscle is included in the raw ROI. One possible solution is to remove this part from ROI. It was observed that the adaptive thresholding had problems when the mass region is too close to the region of pectoral muscle, i.e. the mass region can be removed together with the pectoral muscle. One such case is shown in Fig.4.13(a).

The aim of RIC segmentation algorithm is to find the location of suspicious masses with high sensitivity and less number of false positive. The raw ROI after initial segmentation is composed of a large compact region or some regions located randomly without any location information of masses. The pixels within a particular ROI are clustered into two groups (without consideration of outliers) by RIC algorithm when the number of cluster is two. One group represents regions that may contain mass. This group is called suspicious cluster. The other group contains pixels of normal tissues. The suspicious cluster selection is based on the assumption that the intensity of the mass region is higher than that of its surrounding tissues. While the assumption does not work for certain number of mammograms, for example, the intensities of some mass regions are even lower than that of their surrounded glandular tissues, which means that the mean intensity of the suspicious cluster is not the highest among the mean intensities of all the clusters. One malignant mass and two benign masses are missed since they are observed to lie in the non-suspicious cluster, as shown in Fig.4.13(b) to Fig.4.13(d).

4.8 Conclusion

This chapter focuses on a robust information clustering algorithm for the automatic mass detection in digitized mammograms. The scheme employs RIC algorithm for the image segmentation based on the initial segmentation results of two steps of adaptive thresholding. RIC has the properties of identification of outliers automatically. The cluster centers and the memberships of pixels in the non-outlier cluster are relocated using the minimax of mutual information. Because of the identification of pixels around the boundary of mass body and pixels in some noisy regions, the mass region can be differentiated from non-mass region with relatively less number of clusters and false positive. Experimental results of breast mass detection show that the automatic detection system incorporating RIC algorithm works more effectively than that of the incorporation of DA and FCM algorithms. Other methods for raw ROI extraction and relative algorithms based on RIC are also under consideration that may further improve the detection system for large database.

Figure 4.7: Raw ROI extraction. It is the case of mdb111 in MiniMIAS that contains an asymmetry malignant mass surrounded by dense glandular breast tissues. (a) Original image. The circled region contains mass. In addition to the breast region, there are two more other objects in the background. (b) GACE enhancement. (c) Breast region, obtained by first step of thresholding. (d) Raw ROI, extracted by second step of adaptive thresholding.

Figure 4.8: Clustering performance of RIC and DA. (a). The distribution of original data. It includes two clusters with overlapped data. Different symbols refer data to different clusters. (b). Clustering result of RIC. The symbols 'o' represent data that belong to outlier cluster. (c). Clustering result of DA.

Figure 4.9: The performance of RIC on outliers identification and image segmentation for mammographic image data. The cluster number is two. (a) Original ROI that contains a case of mass. (b) Outliers (dark points) that are estimated by RIC. (c) Segmentation result by RIC, where dark points represent outliers. (d). Histogram of the original ROI. (e) Histogram of the original image with outliers removed. (f) Segmentation result by DA, cluster number is two.

Figure 4.10: FROC analysis for the detection systems based on RIC, DA and FCM.

Figure 4.11: Detection results for mdb181. (a) Original image. The circled region contains mass. (b) Detection result of RIC.

Figure 4.12: Detection results for mdb198. (a) Original image. The circled region contains mass. (b) Detection result of RIC.

Figure 4.13: Masses that are missed. (a) The mass region is removed together with pectoral muscle. (b)-(d) The masses are missed during suspicious cluster selection.

Chapter 5

Vicinal Support Vector Machine for Mass Pattern Analysis

5.1 Introduction

Mammography screening is a technique that is used widely for breast cancer detection. Studies have shown that mammography is sensitive in screening and diagnosis of breast cancer, but with a high false positive. Considering the suffering of patients and cost of biopsy, it is desirable to develop computer-based methods to distinguish more accurately between normal and abnormal, benign and malignant masses, which may help the radiologists in analyzing breast cancer in initial reading.

Normally, most benign masses possess well-defined and less fuzzy boundary compared with the fuzzy and ill-defined boundaries possessed by most of malignant masses. Studies [72] have shown that the difference between benign and malignant masses are not so clear, i.e. certain benign masses may also possess the poorly defined margins. In other word, some benign masses appear with irregular and spic-

ulated boundaries and some malignant masses possess well-defined margins. The classification of breast masses based on shape features are effective in separating circumscribed benign masses from spiculated malignant cases, but not so effective for benign and malignant masses that both have irregular boundaries. Furthermore, the mass boundaries are blurry, which make it difficult for automatic mass detection.

Figure 5.1: Mammograms that contain normal and abnormal cases. The abnormal cases in (b) and (c) are illustrated in small windows. (a) Mammogram with normal case. (b) Mammogram with benign mass. (c) Mammogram with malignant mass.

In this chapter, a new classifier of vicinal support vector machine (VSVM) is proposed for semi-automatic mass pattern analysis. The term of semi-automatic means that the suspicious regions are selected by users but not by computer automatically. The analysis scheme includes two steps. First, one-class support vector machine (SVM) is applied for the abnormal case analysis, where only normal cases are served as training samples. Then VSVM is proposed and investigated for the malignant mass identification. The aim of this step is to decide whether an identified abnormal case is benign or malignant.

There are two main reasons to address the particular application of using VSVM. First, outliers exist between the two classes. Second, the data distribution is unknown. It should be noted that one of the main assumption of the standard SVM

is that all samples in the training set are independent and identically distributed (i.i.d). For this consideration, Song *et al* [79] proposed a robust support vector machine. While in many practical engineering applications, the obtained training data set is often subject to different vicinities. The vicinal SVM was first proposed according to the vicinal risk minimization (VRM) principle by Vapnik [112], which introduces a new class of support vector machine by defining the relevant vicinal kernel functions. It assumes that for a given set of training data, the unknown density function is smooth in a vicinity of any point and the function minimizing is also smooth and symmetric in the vicinity of any point. The main motivation of the vicinal SVM is to take care of different probability density functions respect to different vicinities of the training data set. It derives a hyperplane by maximizing the margin between two classes under the vicinal risk function. However, Vapnik derived the vicinity using a simple algorithm of the input data space and the partition ability of the vicinities is rather limited. This chapter extends the concept of the deterministic annealing clustering method proposed by K.Rose [48] in input space to feature space, as a general clustering algorithm to partition the whole training samples into different soft vicinal areas by kernel method for the application of VSVM.

The proposed VSVM approach for semi-automatic mass analysis is tested by using a database of 118 clinical mammograms for abnormal cases identification and 56 clinical mammograms for malignant masses recognition. All these images are taken from MIAS database. Nine texture features are analyzed, seven out of these nine features are selected for the benign or malignant case analysis. For the first step, 27 normal cases were used as training samples, and 91 cases are used as testing samples. For the second step, 29 cases were used as training samples and 27 cases for testing. Leave one out method is adopted during feature selection and VSVM testing stage.

5.2 One-Class Support Vector Machine

Traditionally, both positive and negative examples are necessary in training phase for two-class classification problems. While what this study concerns about in the first step is when only one class information is available, e.g., only positive examples are available for training. SVM methodology to handle training problems using only one class information has been proposed by Schölkopf *et al* [9], called *one-class* SVM, which has found a wide range of real-world applications [55] [20]. It assumes that the origin is the only member representing the second class in the feature space. The decision function f takes the value $+1$ in a *small* region capturing most of the second class data points, and -1 elsewhere. For a new point $\mathbf{x}$, the value $f(\mathbf{x})$ is determined by evaluating which side of the hyperplane it falls on. Vapnik *et al* [1] expanded this idea into clustering method by transforming the data from input space into a high dimensional feature space using nonlinear transformation to look for the smallest enclosing sphere of radius, i.e. to look for the smallest radius of a ball in the feature space that encloses all the data belonging to the positive class, and negative patterns are elsewhere. This chapter uses the *balls* to describe the data in feature space.

The abnormal cases identification is first treated as an unsupervised learning problem. The main reasons to address one-class classification in abnormal cases recognition lie in two facets. First, large amount of information on normal cases are available and easier to be obtained compared with the information on abnormal cases in the real world applications. Second, in medical applications, features of normal cases are mostly common while features of abnormal cases vary a lot. The one-class classification problem assumes that only one class information is available in the training phase and no other class information is present to provide some constraints on the decision boundary. In this study, only normal cases are used as training samples. The task is to predict a boundary around the class of normal cases

Figure 5.2: Data distribution in the feature space. The hypersphere containing most of the data in one class described by center a and radius R . Data on the boundary are support vectors, data that are outside boundary have $\xi_i > 0$.

so that it can accept as many data of the normal objects as possible, meanwhile, it minimizes the chance of accepting data representing abnormal objects.

Assume there are l training samples $\mathbf{x}_i$, where $i \in [1, \dots, l]$, using a nonlinear transformation ϕ to transform the samples from input space to a high dimensional feature space, we look for the smallest enclosing sphere with radius R , such that

$$\min_{R \in \mathbb{R}, \xi \in \mathbb{R}^l, C \in F} R^2 + C \sum_i \xi_i \quad (5.1)$$

$$\text{subject to } \|\phi(\mathbf{x}_i) - a\|^2 \leq R^2 + \xi_i \quad (5.2)$$

where C is the penalty term, $\xi_i \geq 0$ is the slack variable, a is the center of the training samples in feature space, R is the radius of the circle that contains all the training samples in feature space. After introducing Lagrangian L and the derivative

L with respect to R , a and ξ_i , respectively, it leads to the dual form function

$$\min_{\beta} \sum \beta_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j) - \sum_i \beta_i k(\mathbf{x}_i, \mathbf{x}_j) \quad (5.3)$$

subject to

$$0 \leq \beta_i \leq C, \quad (5.4)$$

$$\sum_i \beta_i = 1 \quad (5.5)$$

where β_i are Lagrange multipliers. The solution of the sphere center

$$a = \sum_i \beta_i \phi(\mathbf{x}_i) \quad (5.6)$$

The radius of the sphere is:

$$R = \{R(\mathbf{x}_i) \mid \mathbf{x}_i \text{ is the support vector}\}. \quad (5.7)$$

Finally the decision function obtained

$$f(x) = \text{sign} \left(R^2 - \sum_{ij} \beta_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j) + 2 \sum_i \beta_i k(\mathbf{x}_i, \mathbf{x}) - k(\mathbf{x}, \mathbf{x}) \right). \quad (5.8)$$

The kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$, which implicitly mapping the input vector into a high-dimensional feature space through an inner product between feature vectors, plays a crucial role in SVM algorithms. These kernel parameters selection results will directly affect the performance of SVMs.

5.3 The Proposed Vicinal Support Vector Machine (VSVM)

The VSVM is originally proposed by Vapnik with the input space partition according to the VRM principle, which aims at using the SVM [112] technique to obtain solutions in the form called vicinal kernels corresponding to different vicinities of the training points. In this part, a short introduction of VSVM for input space partition is first present according to Vapnik. Following is the detailed illustration on the proposed VSVM in the feature space. Finally, the kernel based deterministic annealing (KBDA) method for the application of VSVM is given.

5.3.1 VSVM with Input Space Partition

Considering the input-output training data pairs

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l), y_k \in \{-1, 1\}, k = 1, \dots, l \quad (5.9)$$

These training data pairs are normally assumed to be generated i.i.d according to an unknown probability distribution $p(\mathbf{x}, y)$. However, in practical world, it is possible that the probability distribution can be multi-modal with different vicinities of the training data pairs.

According to Vapnik, the vicinity functions $v(\mathbf{x}_k)$ of sample $\mathbf{x}_k$ are constructed if the training samples satisfy two assumptions:

1. The unknown density function is smooth in a vicinity of any point $\mathbf{x}_k$.
2. The function minimizing the risk functional is also smooth and symmetric in the vicinity of any point $\mathbf{x}_k$.

Then Vapnik proposed the following optimization problem based on VRM principle [73], named linear vicinal SVM

$$\text{Minimize } \Phi(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{k=1}^l \xi_k \quad (5.10)$$

subject to the risk minimization constraint

$$y_k \int_{v(\mathbf{x}_k)} ((\mathbf{x}, \mathbf{w}) + b) p(\mathbf{x}|v(\mathbf{x}_k)) d\mathbf{x} \geq 1 - \xi_k \quad (5.11)$$

where $\mathbf{w}$ is the weight, C is a penalty constant for the slack variable $\{\xi_k\}$, b is the offset, $v(\mathbf{x}_k)$ is the vicinity associated with the training vector $\mathbf{x}_k$, and $p(\mathbf{x}|v(\mathbf{x}_k))$ is the conditional probability of the respective vicinity in the input space.

It can be shown that the optimization problem of the VSVM is equivalent to the standard SVM solution if the training vector $\mathbf{x}_k$ is the center of mass of the respective vicinity, i.e.

$$\mathbf{x}_k = \int v(\mathbf{x}_k) p(\mathbf{x}|v(\mathbf{x}_k)) d\mathbf{x} \quad (5.12)$$

Then the vicinal SVM solution coincides with the standard SVM solution

$$y_k \int_{v(\mathbf{x}_k)} ((\mathbf{x}, \mathbf{w}) + b) p(\mathbf{x}|v(\mathbf{x}_k)) d\mathbf{x} = (\mathbf{x}_k, \mathbf{w}) + b \quad (5.13)$$

There are two kinds of vicinity functions: hard vicinity function and soft vicinity function. For soft vicinity functions, all points of the space may belong to a vicinity of the sample $\mathbf{x}_k$, while they have different measures. For hard vicinity functions, only the points in a vicinity area contribute to the sample $\mathbf{x}_k$. It can be seen that Vapnik proposed a simple scheme to partition the input space into different vicinities according to different metrics under certain limitations. He proposed Laplacian type

kernel as vicinal kernel for hard vicinity functions, and Gaussian distribution is used for soft vicinity of the input space to derive the soft vicinal kernel.

5.3.2 VSVM with Feature Space Partition

However, for a general nonlinear case like the kernel based SVM, application of the VRM principle with the input space partition may not be straightforward because there is nonlinear mapping to transform the input training data into the feature space. We extend Vapnik's idea to the clustering methods in feature space. For instance, there are l training samples with K vicinities, where $K \leq l$ and kth vicinity consists of N_k data pairs

$$(\phi_k(\mathbf{x}_i), y_k), i = 1, \dots, N_k, k = 1, \dots, K. \quad (5.14)$$

K vicinities here are constructed through clustering method, i.e. clustering the samples into K clusters in feature space, each cluster represents one vicinity respectively. Let's first consider soft vicinity, the mass center $\phi_k(\mathbf{x})$ of the kth vicinity in the feature space is

$$\phi_k(\mathbf{x}) = \sum_{i=1}^l \gamma_{ki} \phi(\mathbf{x}_i), k = 1, \dots, K \quad (5.15)$$

with

$$l = \sum_{k=1}^K N_k \quad (5.16)$$

where γ_{ki} is the parameters to be decided by the clustering algorithm in the feature space. For hard vicinity, the mass center $\phi_k(\mathbf{x})$ of the kth vicinity in the feature

space is

$$\phi_k(\mathbf{x}) = \sum_{i=1}^{N_k} \gamma_{ki} \phi(\mathbf{x}_i), k = 1, \dots, K \quad (5.17)$$

Considering $\gamma_{ki} = 0$, when samples $\phi(\mathbf{x}_i)$ is not within the vicinity, the equation (5.17) is used to represent both soft vicinity and hard vicinity mass center.

Contrast to vicinities in the input space, it is natural to consider the partition in feature space since SVM can be generalized as a linear classifier and is native in the feature space. Therefore, we propose a similar optimization problem as in [112] but with the feature space partition according to the VRM principle.

$$\text{Minimize } \Phi(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{k=1}^K \xi_k \quad (5.18)$$

subject to

$$y_k \int_{v(\phi_k(\mathbf{x}))} f(\mathbf{x}, \mathbf{w}, b) p(\phi(\mathbf{x}) | \phi_k(\mathbf{x})) d\phi(\mathbf{x}) \geq 1 \quad (5.19)$$

it is equal to

$$y_k \int_{v(\phi_k(\mathbf{x}))} ([(\phi(\mathbf{x}), \mathbf{w}) + b] p(\phi(\mathbf{x}) | \phi_k(\mathbf{x}))) d\phi(\mathbf{x}) \geq 1 \quad (5.20)$$

where $v(\phi_k(\mathbf{x}))$ represents the k th vicinity associated with the mass center $\phi_k(\mathbf{x})$ in the feature space, and $p(\phi(\mathbf{x}) | \phi_k(\mathbf{x}))$ is the conditional probability of the respective vicinity in the feature space.

It can be proved that the constraint of the VSVM in the feature space can be

presented as

$$\begin{aligned} & y_k \int_{v(\phi_k(\mathbf{x}))} ([(\phi(\mathbf{x}), \mathbf{w}) + b] p(\phi(\mathbf{x}) | \phi_k(\mathbf{x}))) d\phi(\mathbf{x}) \\ &= y_k [(\sum_{i=1}^l \gamma_{ki} \phi(\mathbf{x}_i), \mathbf{w}) + b] \geq 1, k = 1, \dots, K \end{aligned} \quad (5.21)$$

Then the one vicinal kernel is defined as

$$L_k(\mathbf{x}) = \sum_{i=1}^l \gamma_{ki} K(\mathbf{x}, \mathbf{x}_i), k = 1, \dots, K \quad (5.22)$$

and the two vicinal kernel is

$$M_{km}(\mathbf{x}) = \sum_{i=1}^l \sum_{j=1}^l \gamma_{ki} \gamma_{mj} K(\mathbf{x}_i, \mathbf{x}_j) \quad (5.23)$$

where $m, k = 1, \dots, K$. Then the VSVM with the feature space partition has the solution

$$f(\mathbf{x}) = \sum_{k=1}^K \beta_k y_k L_k(\mathbf{x}) + b \quad (5.24)$$

where β_k is the coefficient which minimizes the functional

$$W(\beta) = \sum_{k=1}^K \beta_k - \frac{1}{2} \sum_{k,m=1}^K \beta_k \beta_m y_k y_m M_{km}(\mathbf{x}) \quad (5.25)$$

Subject to the constraints

$$\sum_{k=1}^K y_k \beta_k = 0 \quad (5.26)$$

$$\beta_k \geq 0 \quad (5.27)$$

To obtain a sparse solution at the cost of the extra clustering procedure, a good selection of the cluster number with class information is a most challenging problem and may lead to a better understanding of the new VSVM.

5.3.3 Kernel Based Deterministic Annealing for VSVM

In order to obtain the soft partition scheme of samples in feature space, the concept of the mass constrained deterministic annealing (DA) clustering method proposed by Rose [48] is extended into feature space as a general soft clustering algorithm, particularly, for the application of the VSVM. Mass constrained DA has the ability to avoid local minima of non-convex optimization problems. The square distance between the mass center and the training vector of feature space can be defined in the implicit form of kernel functions. Assume that all the patterns in feature space is partitioned into K clusters, where the mass center of each cluster is represented by $\phi_k(\mathbf{x})$, $k = 1, \dots, K$. According to the partition of (5.17), γ_{ki} is the fuzzy membership parameter decided by kernel based deterministic annealing (KBDA) algorithm. It shows in (5.17) that all the patterns in the feature space have relationship with all the K cluster centers.

Then the squared Euclidian distance between the center and the training vector of feature space can be defined in the implicit form of kernel functions

$$\begin{aligned}
 d_k(\phi(\mathbf{x})) &= \|\phi(\mathbf{x}_i) - \phi_k(\mathbf{x})\|^2 \\
 &= \left\| \phi(\mathbf{x}_i) - \sum_{i=1}^l \gamma_{ki} \phi(\mathbf{x}_i) \right\|^2 \\
 &= k(\mathbf{x}_i, \mathbf{x}_i) - 2 \sum_{i=1}^l \gamma_{ki} k(\mathbf{x}, \mathbf{x}_i) \\
 &\quad + \sum_{i=1}^l \sum_{j=1}^l \gamma_{ki} \gamma_{kj} k(\mathbf{x}_i, \mathbf{x}_j)
 \end{aligned} \tag{5.28}$$

where $k(\mathbf{x}, \mathbf{x}) = (\phi(\mathbf{x}), \phi(\mathbf{x}))$ is the kernel function. The kernel trick provides us a tool of implicitly calculating the dot product in input space by using kernel function instead of directly calculating the dot product in feature space, which is a high dimension space that will be impossible solved directly.

In this study, a kernel based optimization scheme is proposed by using deterministic annealing for the parameters estimation. Suppose the patterns in kernel space, the cost function is stated as

$$\begin{aligned} D &= \sum_{i=1}^l \sum_{k=1}^K p(\phi(\mathbf{x}_i), \phi_k(\mathbf{x})) d_k(\phi(\mathbf{x}_i)) \\ &= \sum_{i=1}^l p(\phi(\mathbf{x}_i)) \sum_{k=1}^K p(\phi_k(\mathbf{x}) | \phi(\mathbf{x}_i)) d_k(\phi(\mathbf{x}_i)) \end{aligned} \quad (5.29)$$

The optimization problem can be reformulated as minimization of the Lagrangian

$$F = D - TH \quad (5.30)$$

where T is the Lagrange multiplier called temperature. H is the Shannon entropy in feature space. To determine the parameter $\gamma(\cdot)$ in (5.17), minimizing the free energy function F with respect to the association probability $p(\phi_k(\mathbf{x}) | \phi(\mathbf{x}_i))$ is straightforward and gives the so-called titled distribution, which is linked to the Gibbs distribution, as

$$p(\phi_k(\mathbf{x}) | \phi(\mathbf{x})) = \frac{p_k \exp(-d_k(\phi(\mathbf{x}))/T)}{Z_\phi(\mathbf{x})} \quad (5.31)$$

where p_k is the fraction of the center of the k th cluster in the kernel space, and the normalized factor is

$$Z_\phi(\mathbf{x}) = \sum_{k=1}^K p_k \exp(-d_k(\phi(\mathbf{x}))/T) \quad (5.32)$$

Plugging (5.31) into (5.30), the free energy function in feature space is presented as

$$\begin{aligned}
 F^* = & -T \sum_{i=1}^l p(\phi(\mathbf{x}_i)) \log \sum_{k=1}^K p_k \exp(-d_k(\phi(\mathbf{x}_i))/T) \\
 & + \lambda \left(\sum_{k=1}^K p_k - 1 \right)
 \end{aligned} \tag{5.33}$$

where λ is the Lagrangian multiplier. Minimizing the free energy of F^* against $\phi_k(\mathbf{x})$ we have

$$\frac{\partial(F^*)}{\partial(\phi_k(\mathbf{x}))} = 0 \tag{5.34}$$

i.e.

$$\sum_{i=1}^l p(\phi(\mathbf{x}_i)) p_k \exp(-d_k(\phi(\mathbf{x}_i))/T) [\phi(\mathbf{x}_i) - \phi_k(\mathbf{x})] = 0 \tag{5.35}$$

Divide the above equation by the normalized factor leads to

$$\begin{aligned}
 & \frac{p(\phi(\mathbf{x}_i)) p_k \exp(-d_k(\phi(\mathbf{x}_i))/T)}{Z_\phi(\mathbf{x})} \phi(\mathbf{x}_i) \\
 = & \frac{p(\phi(\mathbf{x}_i)) p_k \exp(-d_k(\phi(\mathbf{x}_i))/T)}{Z_\phi(\mathbf{x})} \phi_k(\mathbf{x})
 \end{aligned} \tag{5.36}$$

Use (5.31) the above equation can be rewritten as

$$\begin{aligned}
 & \sum p(\phi(\mathbf{x}_i)) p(\phi_k(\mathbf{x}) | \phi(\mathbf{x}_i)) y_i \phi(\mathbf{x}_i) \\
 = & \sum_{i=1}^l p(\phi(\mathbf{x}_i)) p(\phi_k(\mathbf{x}) | \phi(\mathbf{x}_i)) \phi_k(\mathbf{x})
 \end{aligned} \tag{5.37}$$

i.e.

$$\begin{aligned}\phi_k(\mathbf{x}) &= \frac{p(\phi(\mathbf{x}_i))p(\phi_k(\mathbf{x})|\phi(\mathbf{x}_i))\phi(\mathbf{x}_i)}{p(\phi(\mathbf{x}_i))p(\phi_k(\mathbf{x})|\phi(\mathbf{x}_i))} \\ &= \sum_{i=1}^l \gamma_{ki}\phi(\mathbf{x}_i)\end{aligned}\tag{5.38}$$

where

$$\gamma_{ki} = \frac{p(\phi(\mathbf{x}_i))p(\phi_k(\mathbf{x})|\phi(\mathbf{x}_i))}{\sum_{i=1}^l p(\phi(\mathbf{x}_i))p(\phi_k(\mathbf{x})|\phi(\mathbf{x}_i))}\tag{5.39}$$

Note that (5.31) and (5.38) are the clustering algorithm by KBDA algorithm. The implementation procedure is much the same as the implementation procedure of DA algorithm in input space, where fixed point iteration is adopted. The different update equations are (5.29), (5.31) and (5.38) in KDBA algorithm.

5.4 Data Acquisition

5.4.1 Region of Interest

Individual mass is localized in a mammogram according to the information provided by the database. Therefore, it is sufficient to examine the image content within a small neighborhood around the location to detect whether a mass is present at a given location. The region of interest(ROI) is defined as a small window with $2R \times 2R$ pixels centered at the mass center, where R is the radius of the area that contains mass. For normal cases, the ROIs are randomly selected in mammograms.

Figure 5.3: Selected ROI. (a)Original mammographic image from MIAS. (b) Region of interest, the inner square area is the bounding box of mass region.

5.4.2 Feature Extraction

Features are used to describe the character of an object. It is a key step in most pattern analysis tasks. An artificial system can identify the segmented suspicious area and make final decision based on certain features of the mass. Unlike the much more sophisticated process of human observer to identify a mass, the machine observers make decision with limited features. The extracted features to identify an object should have four characteristics: discrimination, reliability, independence and small number. Discrimination means that the features represent objects in different classes should have significantly different values. The reliability refers to features that describe the objects in the same class should have much similar values. Features should be independent, i.e. they should not be strongly correlated to each other. It does not mean that the large the number of the features, the better they

present an object. Acturally, some redundant features can even make the decision worse. It is a problem of curse of dimension. Thus, a small number of features that can fully represent the character of mass in different class are desirable.

Based on the radiologists' experience, the primary features that indicate malignancy of masses have been suggested by previous research work in mass detection [72]. The main concerns are tumor's intensity, geometric and texture features, which are calculated within or around the boundary of a mass. Though these features work effectively in some mass detection system, while they cannot capture all of the capable aspects of human perception nature. This study focuses on the gradient-based and texture-based features, which are computed based on gray-level co-occurrence matrices (GLCM).

Statistical image texture analysis, which has been described by Raeth *et al* [111], is used to improve the ability of distinguishing normal from abnormal, benign from malignant cases. Nine features are extracted within ROIs. Among these features, two are the first order texture features, i.e. the mean of gray level and gray variance that describe the hight and shape of the gray level histogram. The other seven features are second order texture features. Two features are calculated from the first order gradient distribution: mean gradient and variance of gradient. Five features are calculated from the gray level co-occurrence matrix: namely, energy, inertia, entropy, homogeneity and correlation. The co-occurrence matrices are taken in the East-West directions at pixels spacing of 1. The features employed in this study are illustrated as below.

Suppose there are altogether N pixels in a ROI, the gray value of each pixel is represented by $g(i, j)$, then

Mean Gray Level, a measure of average brightness in ROI is given by

$$G_{avg} = \frac{1}{N} \sum_{i,j} g(i,j) \quad (5.40)$$

Variance of Gray Level, a description of the spread of gray level histogram, is given by

$$G_{var} = \frac{1}{N} \sum_{i,j} (g(i,j) - G_{avg})^2 \quad (5.41)$$

Mean of Gradients, a measure of average magnitudes of brightness transitions, and *Variance of Gradients* are given in the same form as G_{avg} and G_{var} respectively, except that the gradient at each pixel is instead of gray level $g(i,j)$.

For the features related with Co-occurrence matrix, suppose $p(i,j)$ is the (i,j) th element of a normalized co-occurrence matrix, then

Energy, a measure of sum probability of gray levels i, j in two pixels a fixed distance apart, is given by

$$Eg = \sum_{i,j} p(i,j)^2 \quad (5.42)$$

Inertia or *Contrast*, a measure of the frequency of large gray level changes, is given by

$$Con = \sum_{i,j} (i-j)^2 p(i,j) \quad (5.43)$$

Entropy, a measure of texture coarseness, is given by

$$Etr = \sum_{i,j} p(i,j) \log p(i,j) \quad (5.44)$$

Homogeneity, a measure of local homogeneity, is given by

$$H = \sum_{i,j} \frac{p(i,j)}{1 + (i-j)^2} \quad (5.45)$$

Correlation, a measure of linear dependency of brightness, is given by

$$Cor = - \sum_{i,j} \frac{(i - \mu_x)(j - \mu_y)}{\sqrt{\sigma_x \sigma_y}} p(i,j) \quad (5.46)$$

where $\mu_x = \sum_i i \sum_j p(i,j)$, $\mu_y = \sum_j j \sum_i p(i,j)$, $\sigma_x = \sum_i (i - \mu_x)^2 \sum_j p(i,j)$, and $\sigma_y = \sum_j (j - \mu_y)^2 \sum_i p(i,j)$

5.5 Experimental Results

The one-class SVM and the proposed VSVM software are programmed based on Steven Gunn's Matlab SVM software [97].

5.5.1 Abnormal Cases Detection by One-Class SVM

The training samples for one-class SVM are normal cases from 27 mammograms that do not contain masses. These mammograms have background with fat, glandular and density tissues respectively. The study has tested 63 cases with 7 normal cases, 29 benign and 27 malignant masses. In this step, Gaussian RBF kernel function is used.

Two parameters will affect the generalization performance in one-class SVM with RBF kernel, i.e. C and σ . As indicated in [1], the number of outliers are controlled by the parameter C . The purpose is to find a sphere in the feature space that includes all samples belonging to normal cases. This study shows that outliers are not an

important problem for the first step analysis. We then set $C = 1$ and regulate the kernel parameter σ . Fig.5.4 illustrates the generalization ability versus kernel parameter σ . The experimental results show that the sensitivity is best obtained when $\sigma > 0.2$, which equal to 1. When $\sigma \geq 2$, sensitivity as well as specificity is well obtained. The experimental results demonstrate that the texture information is effective for abnormal step detection, i.e. no abnormal case has been omitted.

Figure 5.4: The performance of one-class SVM for abnormal cases analysis versus Gaussian RBF kernel width σ .

5.5.2 Malignant Cases Detection by VSVM

In this step, fifty-six mammographic images that were identified as abnormal cases by the first diagnosis step are taken into account for the malignant cases recognition, i.e. to classify the abnormal cases into benign or malignant cases. Feature selection is first done before the final generalization performance is obtained.

Feature selection is important in many machine learning problems as a means of increasing generalization performance and reducing running time. It also helps to give constraints and interpretational issues imposed by the problem itself. This study only concerns about feature selection result on the generalization performance of the

proposed VSVM classifier. Nine features are calculated as illustrated in the previous section. In order to check whether there exists some redundant features that make the performance worse, leave one out method is used in feature selection. Seven features are observed to contribute more on the classification performance, namely, energy, inertia, entropy, homogeneity, correlation, gradient mean and gradient variance. Other two features make the performance even worse.

Based on feature selection, KBDA is first run within each class, i.e. unsupervised clustering method is performed for samples in the same class. The same number of cluster is allocated to the two classes. For example, if the number of cluster in positive class is two, then number of cluster in negative class is also set as two. Then VSVM is carried out to classify the malignant cases into benign and malignant masses. The classification performance is also studied by leave one out method. Gaussian RBF kernel is used.

The performance of VSVM with different cluster number is first studied. We mainly investigated the three representing cluster number of two, five and ten. For these fifty-six data set, the best benign versus malignant classification accuracy are the same as 84% for VSVM with these three different cluster number. The performance of VSVM is further validated by using the receiver operating characteristics (ROC) curve methodology. We found that VSVM with cluster number of five performs best with largest area under curve. ROCKIT 0.9B [10] was used to fit ROC curves to the continuous data obtained from the output of the classifier.

Method	Accuracy	Sensitivity	Specificity	Area under Curve(Az)
VSVM	84%	93%	74%	0.89
SVM	82%	86%	78%	0.86

Table 5.1: The classification results by SVM and VSVM.

For the purpose of comparison, the performance of standard SVM is also studied. The best classification accuracy of standard SVM is 82% when $C = 10$ and kernel

width equals 8. These texture features result in sensitivity of 86% with specificity of 78%. From table 5.1 we can see that the accuracy of VSVM with cluster number of five is 84%, with the sensitivity of 93% and specificity of 74%. The area under SVM and VSVM ROC curve (A_z) in Fig.5.5 are 0.86 and 0.89, respectively.

Figure 5.5: ROC curve for VSVM and standard SVM for malignant masses classification. The performance of VSVM with cluster number $K=2, 5, 10$ are present.

From these experimental analysis, we can see that VSVM with cluster 5 performs best for this data set.

The study also shows that slack variable C in VSVM acts different as it is in standard SVM. In standard SVM, C is a pre-selected parameter that controls the penalty on errors, a larger value of C corresponding to assign a higher penalty to errors. Adaptive margin is obtained in non-separable case through the term $C(\sum_i \xi_i)^k$. In VSVM, the adaptive margin is obtained through kernel based deterministic annealing process. As illustrated in last section, vicinal kernel (5.24) and (5.25) are associated with parameter γ which is obtained through KBDA. In mass-constrained deterministic annealing, as the temperature is lowered, the system undergoes a sequence of *phase transitions* that consists of natural cluster splittings, with high probability to their close cluster and low probability to farther cluster. Then the new clusters will move to the higher points density. Experiments indicates that the temperature T is

a parameter acts as C in VSVM that affects the value of γ . The value of γ will be larger if it is closer to the cluster center. The adaptive margin is obtained through annealing process.

5.6 Conclusion

In this study, a new classifier of VSVM is proposed for a semi-automatic scheme of mass pattern analysis. The effort is particularly focused on KBDA method for the application of VSVM in abnormal cases classification. There are basically two learning phases in the proposed VSVM algorithm: clustering and SVM learning. The targets of VSVM is to minimize the vicinal risk function under the constraints of respective vicinal areas pre-defined by KBDA. Experimental results show that the two-step detection scheme works effective for mass pattern identification and VSVM with cluster number of five performs best for this data set. It also shows that the VSVM is a promising classifier that is adaptive to different data structures. We note that the performance of the proposed classifier needs to be further improved, e.g. the relationship between cluster number and the classifier performance, the relationship between the kernel parameter in KBDA and the kernel parameter in SVM learning. It is hoped that this method can provide a more useful tool for the clinical application in mass pattern diagnosis.

Chapter 6

Conclusion and Recommendations

6.1 Conclusion

This thesis provides an in-depth investigation to develop advanced machine learning algorithms for automatic breast mass detection in digitized mammograms. The work consists of the establishment of software system to process the digitized mammographic images automatically. According to the character of masses and the background breast tissue in digitized mammograms, two image segmentation algorithms based on information theory and a new classifier based on statistical learning theory are proposed. The main contributions of this thesis include: the proposal DACF method in the segmentation of circumscribed mass, the investigation of RIC algorithm in the segmentation of masses that are embedded in glandular or dense glandular breast tissue, the study of VSVM for mass pattern analysis that are embedded in fat, glandular or dense glandular breast tissue with various shapes by a semi-automatic approach. The performance of the proposed algorithms is presented in table 6.1.

Accurate detection of the masses could provided valuable information to the clini-

Features	Number of tested images	Probability of segmentation error	Sensitivity	False positive per image
Circumscribed mass segmentation	23	11.2%	-	-
Dense mass detection	54	-	91.7%	2.56
Mass pattern analysis	63	-	93%	-

Table 6.1: Performance of the proposed algorithms.

cians. The DACF segmentation method offers a valuable tool to the location of the circumscribed masses. The establishment of automatic detection system based on RIC algorithm makes the automatic analyzing system become more reliable for the masses embedded in dense breast tissues, and the VSVM provides a new tool for mass pattern analysis that is valuable in mammographic diagnosis. The proposed algorithms and the results are summarized in the following.

6.1.1 Segmentation of Circumscribed Mass

The circumscribed mass is one type of masses that have relative well-defined margin and close to round shape. While it is still a challenging task to segment such kind of masses with relatively high accuracy. A new algorithm, called DACF, is proposed for circumscribed masses segmentation based on deterministic annealing by maximizing Shannon entropy. Different from the mass constrained deterministic annealing and other related clustering methods, the proposed algorithm takes both the intensity information and spatial distance into account. The new dissimilarity measure integrates the intensity dissimilarity and spatial information in terms of a circular shape function. Because of the presence of the circular shape function, the new dissimilarity measure is able to differentiate the pixels that have similar intensity information but located in different regions in the mammographic images.

The proposed DACF algorithm is compared with DA and FCM methods for circumscribed mass segmentation. It shows that the DACF algorithm can give more accurate mass localization, especially when the margins of masses are fuzzy. Twenty-three mammographic images that contain all circumscribed masses in MiniMIAS are tested by the proposed algorithm, and the average of the probability of segmentation error is 11.2%, which is close to 3.5% less than the segmentation results by DA and FCM. In addition to that, the convergence speed of DACF is faster than that of DA and FCM for the optimal segmentation results.

6.1.2 Detection of Masses Embedded in Glandular or Dense Glandular Breast Tissue

Masses with irregular, spiculated margins, architectural distortion and asymmetry in shapes have high tendency to be malignant. Particularly, it is hard to detect masses embedded in dense glandular breast tissue. In order to detect such kind of masses automatically, a robust information clustering algorithm (RIC) is proposed for for automatic mammographic mass detection. RIC is based on the minimax of the mutual information. The minimization of mutual information of interclusters leads to the deterministic annealing. The maximization of the mutual information provides a robust density estimation of probability soft margin to identify outliers. Pixels on the boundary of mass and noisy points are identified in the robust estimation. The purpose to address RIC in automatic mass detection is to locate the mass region with relatively high detection rate and small number of false positive since some of the noisy regions are identified and rejected during RIC robust estimation. RIC is introduced together with other image processing techniques for the mass detection in mammograms. The detection method has been verified with 54 mammograms in the MIAS database. Each mammogram contains at least one

mass case embedded in glandular or dense breast tissues. The verification results show that the proposed algorithm has a sensitivity of 91.7% and the number of false positive per image is 2.56. Experiments based on the mammograms from MIAS database demonstrate that the RIC is robust to differentiate the mass region from its background, and good detection rate is obtained.

6.1.3 Mass Pattern Analysis

The background tissues that surround breast masses are fatty, fatty-glandular and dense glandular breast tissues. In order to distinguish more accurately between normal and abnormal, benign and malignant masses within different types of breast tissues, a new classifier of vicinal support vector machine (VSVM) is proposed for semi-automatic mass analysis. The main motivation of the VSVM is to take care of different probability density functions respect to different vicinities of the training data set since the difference between benign and malignant masses are not so clear. The concept of the deterministic annealing clustering in input space is extended into feature space, as a general clustering algorithm to partition the whole training samples into different soft vicinal areas for the application of VSVM. The semi-automatic term means that the suspicious regions are selected by users instead of by computer automatically. The diagnosis system includes two steps. First, one-class support vector machine (SVM) is applied for the abnormal cases detection, where only normal cases are served as training samples. Then VSVM is investigated for the malignant cases identification.

The system has been tested by using 118 mammograms for abnormal cases identification and 56 mammograms for malignant masses recognition. The best benign versus malignant classification accuracy is 84% for VSVM, with a sensitivity of 93% and a specificity of 74%. The area under the ROC curve (A_z) is 0.89 for VSVM and 0.86 for standard SVM. Experimental results show that the VSVM algorithm is a

promising classifier for mass pattern analysis.

6.2 Recommendations

The work done in this thesis demonstrates that it is promising to develop machine learning algorithms of image segmentation and pattern recognition for the automatic detection of breast masses in digital mammograms. Further research work should be carried out to make the automatic detection and diagnosis system more acceptable to clinical applications. Possible areas of future work that emerge from this thesis are briefly discussed in this subsection.

6.2.1 Incorporation of the Spatial Information

The RIC algorithm is observed to work effectively in differentiating the mass body from its background breast tissues. It should be noted that part of the boundary information of masses may be lost during the robust estimation of the outliers. The further diagnosis of mass pattern, i.e. to classify the detected suspicious objects into normal and abnormal, benign and malignant masses, will be carried out based on the features extracted within or around the boundary of the masses. It is hoped that the boundary information is rich enough for accurate classification. Therefore, future work for the application of RIC in automatic mass detection should consider to incorporate spatial information either in the dissimilarity measure or during the restoration of the outliers into respective clusters.

6.2.2 Multi-modality Images Registration

Mammographic images can not provide enough information for masses that are embedded in very dense glandular tissues, or masses that are spiculated surrounded by glandular tissues. The further discrimination of benign and malignant masses also needs more information. Fortunately, different image modalities, such as MRI and ultrasound, are based on different physical principles and capture different and often complementary information. Many applications in clinical practice of breast cancer analysis benefit from the integrated visualization and combined analysis of such multi-modal images. Therefore, multi-modality image registration technique should be taken into account by using the criteria of maximization of mutual information to provide more information for mammographic mass detection and analysis.

6.2.3 Multi-layer Agent Detection System

The composition of the normal breast tissues vary from women to women, and even various for a woman at her different stages of ages. The fat breast tissues are transparent to X-ray, while the glandular and dense glandular tissues are even opaque. Stand-alone detection systems have often been proven insufficient to detect masses in all these kinds of mammograms. Plus, the breast muscle is a key region that can affect the final mass segmentation and detection results. Therefore, a multi-layer agent mass detection system is preferred to be set up that are robust to large database. The system should be have high-level knowledge-based control over low-level image segmentation algorithms. It is a hybrid detection system which is composed of both bottom-up and top-down control strategies for breast mass detection. The father-agents control the high knowledge of the images, which contain fat, glandular, or dense glandular breast tissues respectively. Each son-agent is responsible for one object in the mammograms and communicates with other agents. The

son-agents also control low-level image segmentation of the corresponding object.

6.2.4 Database Setup

A large database should be set up to describe the content of digital mammograms. The database can store: 1) as much as possible of the information of a patient, and 2) as many as possible of the number of patients. It is hoped to provide more objective information to aid both radiologists and computer detection systems to manage a large amount of mammograms. The progress of the disease of a patient can be diagnosed by comparing the mammographic images at different stages. In addition, the diagnosis can be obtained by the means of image retrieval to find the similar mass cases in the database once enough information of the masses are accumulated in the database.

Author's Publications

1. A.Z.Cao, Q.Song, Y.Shui, "Mammographic Mass Segmentation by Deterministic Annealing Clustering Incorporating Circular Function", Submit to *IEEE Transaction on Medical Imaging*, Dec 2004.
2. A.Z.Cao, Q.Song, Z.M.Wang, "Automatic Mass Detection in Digital Mammograms by Robust Information Clustering", Submit to *Computer Vision and Image Understanding*, Dec 2004.
3. A.Z.Cao, Q.Song, X.L.Yang and L.Wang, "Breast Mss Segmentation based on Information Theory" *17th International Conference on Pattern Recognition (ICPR)*, August 23-26, Cambridge, UK, 2004.
4. A.Z.Cao, Q.Song, X.L.Yang and S.Liu, C.Y.Guo, "Mammographic Mass Detection by Vicinal Support Vector Machine," in *Proceedings of IEEE on International Joint Conference on Neural Networks (IJCNN'04)*, July 25-29, Budapest, Hungary, 2004.
5. A.Z.Cao, Q.Song, X.L.Yang and S.Liu, "Breast Mass Segmentation on Digital Mammograms by a Combined Deterministic Annealing Method", *IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI'04)*, April 15-18, Arlington, VA, US, 2004.
6. X.L.Yang, A.Z.Cao, Q.Song, and S.Liu., "Robust C-Shells Based Deterministic Annealing Clustering Algorithm", in *Proceedings of IEEE International Conference on Fuzzy Systems (FUZZ'04)*, July 25-29, Budapest, Hungary, 2004.
7. X.L.Yang, A.Z.Cao, Q.Song, "Clustering Nonlinearly Separable and Unbalanced Data Set", *IEEE International Conference on Intelligent Systems (IS'04)*, June 22-24, Varna, Bulgaria, 2004.

8. S.Liu, Q.Song, W.J.Hu, A.Z.Cao, "Disease Classification Using Support Vector Machine (SVM)", in *Proceedings of 9th International Conference on Neural Information Processing (ICONIP'02), computational intelligence for the E-age*, volume 2, pp.760-763, Nov 18-22, Singapore, 2002.

Bibliography

- [1] A.B-Hur, D.Horn, H.T.Siegelmann, and V.Vapnik. Support vector clustering. *Journal of Machine Learning Research* 2, pages 125–137, 2001.
- [2] A.Ganapathiraju, J.E.Hamaker, and J.Picone. Applications of support vector machines to speech recognition. *IEEE Trans. on Signal Processing*, 52(8):2348–2355, 2004.
- [3] A.K.Jain, M.N.Murty, and P.J.Flynn. Data clustering: a review. *ACM Computing Survey*, 31(3):264–323, 1999.
- [4] A.K.Jain, P.W.Duin, and J.Mao. Statistical pattern recognition: a review. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22:4–37, 2000.
- [5] A.Petrosian, H.P.Chan, M.A.Helvie, M.M.Goodsitt, and D.D.Adler. Computer-aided diagnosis in mammography: classification of mass and normal tissue by texture analysis. *Physics in Medicine and Biology*, 39:2273–2288, 1994.
- [6] A.R.Rao and R.Jain. Knowledge representation and control in computer vision systems. *IEEE Expert*, (1):64–79, 1988.
- [7] A.W-C.Liew, S.H.Leung, and W.H.Lau. Segmentation of color lip images by spatial fuzzy clustering. *IEEE Trans. Fuzzy Systems*, 11:542–549, 2003.
- [8] A.W.C.Liew, S.H.Leung, and W.H.Lau. Fuzzy image clustering incorporating spatial continuity. In *IEE Proceedings- Vision, Image and Signal Processing*, volume 147, pages 185–192, 2000.
- [9] B.Scholkopf, J.C.Platt, J.S-Taylor, and A.J.Smola. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.
- [10] C.E.Metz. *ROC software*. <ftp://random.bsd.uchicago.edu/roc/ibmpc/>.

- [11] C.J.C.Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.
- [12] C.Kimme, B.J.O’Loughlin, and J.Sklansky. *Data structures, computer graphics, and pattern recognition*. Academic, New York, 1977.
- [13] C.M.Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, New York, 1995.
- [14] C.Vyborny. Can computers help radiologists read mammograms. *Radiology*, pages 315–317, 1994.
- [15] C.Xu and J.L.Prince. Snakes, shapes, and gradient vector flow. *IEEE Trans. on Image Processing*, 7(3):359–369, 1998.
- [16] D.Brzakovic, X.M.Luo, and P.Brzakovic. An approach to automated detection of tumors in mammograms. *IEEE Transactions on Medical Imaging*, pages 233–241, 1990.
- [17] D.L.Pharm, C.Y.Xu, and J.L.Prince. A survey of current methods in medical image segmentation. *Ann. Rev. Biomed. Eng*, 2, 2000.
- [18] D.L.Pharm and J.L.Prince. An adaptive fuzzy c-means algorithm for image segmentation in the presence of intensive inhomogeneities. *Pattern Recognition Letter*, 20:57–68, 1999.
- [19] D.L.Pharm and J.L.Prince. Adaptive fuzzy segmentation of magnetic resonance images. *IEEE Trans. Medical Imaging*, 18(9):737–752, 1999.
- [20] D.M.J.Tax and R.P.W.Duin. Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research* 2, pages 155–173, 2001.
- [21] D.Wei, H.P.Chan, M.A.Helvie, B.Sahiner, N.Petrick, D.D.Adler, and M.M.Goodsitt. Classification of mass and normal breast tissue on digital mammograms: multiresolution texture analysis. *Medical Physics*, 22:1501–1513, 1995.
- [22] E.Gokcay and J.C.Principe. Information theoretic clustering. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(2):158–170, 2002.

- [23] E.Osuna, R.Freund, and F.Girosit. Training support vector machines: an application to face detection. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, pages 130–136, 1997.
- [24] F.F.Yin, M.L.Giger, C.J.Vyborny, K.Doi, and R.A.Schmidt. Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses. *Invest Radiol.*, pages 473–481, 1993.
- [25] F.Yin, M.L.Giger, K.Doi, C.E.Metz, C.J.Vyborny, and R.A.Schmidt. Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images. *Medical Physics*, 18:955–963, 1991.
- [26] F.Yin, M.L.Giger, K.Doi, C.J.Vybomy, and R.A.Schmidt. Computerized detection of masses in digital mammograms: automated alignment of breast images and its effect on bilateral-subtraction technique. *Medical Physics*, 21(3):445–452, 1994.
- [27] G.D.Tourassi and R.Vargas-Voracek. Computer-assisted detection of mammographic masses: a template mathing scheme based on mutual information. *Medical Physics*, 30(8):2123–2130, 2003.
- [28] G.J.Mclachln and K.E.Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York, 1988.
- [29] G.M.T.Brake and N.Karsseneijer. Single and multiscale detection of masses in digital mammograms. *IEEE Transactions on Medical Imaging*, 18:628–639, 1999.
- [30] H.-P.Chan, N.Petrack, and B.Sahiner. Computer-aided breast cancer diagnosis. *Artificial Intelligence Techniques in Breast Cancer Diagnosis and Prognosis*, 20:1215–1227, 2001.
- [31] H.D.Li, M.Kallergi, L.P.Clark, V.K.Jain, and R.A.Clark. Markov random field for tumor detection in digital mammography. *IEEE Transactions on Medical Imaging*, 14(3):565–576, 1995.
- [32] H.Frigui and R.Krishnapuram. A robust competitive clustering algorithm with applications in computer vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21:450–465, 1999.
- [33] H.Kobatake, M.Murakami, H.Takeo, and S.Nawano. Computerized detection of malignant tumors on digital mammograms. *IEEE Transactions on Medical Imaging*, pages 367–379, 1999.

- [34] H.Kobatake and Y.Yoshinaga. Detection of spicules on mammogram based on skeleton analysis. *IEEE Transactions on Medical Imaging*, 15:235–245, 1996.
- [35] H.Li, Y.Wang, K.J.Ray Liu, Shih-Chung B.Lo, and M.T.Freedman. Computerized radiographic mass detection-part ii: Decision support by featured database visualization and modular neural networks. *IEEE Transactions on Medical Imaging*, 20:302–313, 2001.
- [36] H.Li, Y.Wang, K.J.Ray Liu, S.C.B.Lo, and M.T.Freedman. Computerized radiographic mass detection-part i: lesion site selection by morphological enhancement and contextural segmentation. *IEEE Transactions on Medical Imaging*, pages 289–300, 2001.
- [37] I.Christoyianni, E.Dermatas, and G.Kokkinakis. Fast detection of masses in computer-aided mammography. *IEEE Signal Processing Magazine*, pages 54–64, 2000.
- [38] I.Vajda. *Theory of Statistical Inference and Information*. Kluwer, Dordrecht, the Netherlands, 1989.
- [39] J.C.Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.
- [40] J.C.Bezdek and N.R.Pal. Some new indexes or cluster validity. *IEEE Trans. on Fuzzy Systems*, 28(3):301–303, 1998.
- [41] J.Kilday, F.Palmieri, and M.D.Fox. Classifying mammographic lesions using computerized image analysis. *IEEE Transactions on Medical Imaging*, 12:664–669, 1993.
- [42] J.K.Udupa and S.Samarasekera. Fuzzy connectedness and object definition: thory, algorithm and applications in image segmentation. *Graph. Models Image Processing*, 58(3):246–261, 1996.
- [43] J.M.Jolion, P.Meer, and S.Bataouche. Robust clustering with applications in computer vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(8):791–802, 1991.
- [44] J.Suckling, J.Parker, D.R.Dance, S.Astley, I.Hutt, C.R.M.Boggis, I.Ricketts, E.Stamatakis, N.Cerneaz, S.L.Kok, P.Taylor, D.Betal, and J.Savage. The mammographic image analysis society digital mammogram databse. In

- A.G.Gale, S.M Astley, D.R.Dance, and A.Y Cairns, editors, *Proc. the 2nd International Workshop on Digital Mammography*, pages 375–378, 1994.
- [45] J.S.Zhang and Y.W.Leung. Robust clustering by pruning outliers. *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, 33(6):983–999, 2003.
- [46] K.I.Kim, K.Jung, and J.H.Kim. Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(12):1631–1639, 2003.
- [47] K.L.Wu and M.S.Yang. Alternative c-means clustering algorithms. *Pattern Recognition*, 35:2267–2278, 2002.
- [48] K.Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, 1998.
- [49] K.Rose, E.Gurewitz, and G.C.Fox. Constrained clustering as an optimization method. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15:785–794, 1993.
- [50] K.Woods. *Automated image analysis techniques for digital mammography*. Ph.D Dissertation, University of South Florida, Dec,1994.
- [51] L.Cinque, G.Foresti, and L.Lombardi. A clustering fuzzy approach for image segmentation. *Pattern Recognition*, 37(9):1797–1807, 2004.
- [52] L.Hadjiiski, H.Chan, B.Sahiner, N.Petrick, and M.A.Helvie. Automated registration of breast lesions in temporal pairs of mammograms for interval change analysis-local affine transformation for improved localization. *Medical Physics*, 28(6):1070–1079, 2001.
- [53] L.Kalisher. Factors influencing false negative rates in xer-mammography. *Radiology*, pages 297–301, 1979.
- [54] L.M.Bruce and R.R.Adhami. Classifying mammographic mass shapes using the wavelet transform modulus-maxima method. *IEEE Transactions on Medical Imaging*, 18:1170–1177, 1999.
- [55] L.M.Manevitz and M.Yousef. One-class svms for document classification. *Journal of Machine Learning Research* 2, pages 139–154, 2001.

- [56] L.N.F.Ana and A.K.Jain. Robust data clustering. *Proceedings on Computer Vision and Pattern Recognition*, 2:II-128-II-133, 2003.
- [57] L.V.Ackerman. *Computer classification of radiographs and xerograms of the breast*. PhD thesis, 1971.
- [58] L.Zheng and A.K.Chan. An artificial intelligent algorithm for tumor detection in screening mammogram. *IEEE Transactions on Medical Imaging*, 20:559–567, 2001.
- [59] M.A.Kupinski and M.L.Giger. Automated seeded lesion segmentation on digital mammograms. *IEEE Trans. Medical imaging*, 17:510–517, 1998.
- [60] Mammography. <http://www.radiologyinfo.org/content/mammogram.htm>.
- [61] M.D.Levine and A.M.Nazif. Dynamic measurement of computer generated image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 7:155–164, 1985.
- [62] M.Giger. *Digital Mammography*. Chicago, IL, USA, 1996.
- [63] M.J.Broeders and A.L.Verbeek. Breast cancer epidemiology and risk factors. In *Radiological Diagnosis of Breast Diseases*, pages 1–12, Berlin, Germany, 1997. Springer-Verlag.
- [64] M.J.Homer. *Mammographic Interpretation: A Practical Approach*. McGraw-Hill, Boston, MA, 1997.
- [65] M.N.Ahmed, S.M.Yamany, N.Mohamed, and A.A.Farag. A modified fuzzy c-means algorithm for bias field estimation and segmentation of mri data. *IEEE Trans. on Medical Imaging*, 21(3):193–199, 2002.
- [66] M.Sameti and R.K.Ward. A fuzzy segmentation algorithm for mammogram partitioning. *Digital Mammography*, pages 471–474, 1996.
- [67] M.Shen and K.L.Wu. A similarity-based robust clustering method. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(4):434–448, 2004.
- [68] N.Karssemeijer and G.M.te Brake. Detection of stellate distortions in mammogram. *IEEE Transactions on Medical Imaging*, pages 611–619, 1996.
- [69] N.Perick, H-P.Chan, B.Sahiner, and D.T.Wei. An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection. *IEEE Trans. On Medical Imaging*, 15:59–67, 1996.

- [70] N.Perick, H-P.Chan, B.Sahiner, and M.A.Helvie. Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. *Medical Physics*, 26:1642–1654, 1999.
- [71] N.R.Mudigonda and R.M.Rangayyan. Detection of breast masses in mammograms by density slicing and texture flow-field analysis. *IEEE Trans. on Medical Imaging*, 20:1215–1227, 2001.
- [72] N.R.Mudigonda, R.M.Rangayyan, and E.L.Desautels. Gradient and texture analysis for the classification of mammographic masses. *IEEE Transactions on Medical Imaging*, 19(10):1032–1043, 2000.
- [73] O.Chapelle, J.Weston, L.Bottou, and V.Vapnik. Vicinal risk minimization. *NIPS 2000*, pages 416–422, 2000.
- [74] Pictures of breast cancer. <http://www.breastcancer.org>.
- [75] P.Gamagami. *Atlas of Mammography: New Early Signs in Breast Cancer*. Blackwell Science, Cambridge, Mass., USA, 1996.
- [76] P.J.Huber. *Robust Statistics*. Wiley, New York, 1981.
- [77] P.Sajda, C.Spence, and J.Pearson. Learning contextual relationships in mammograms using a hierarchical pyramid neural network. *IEEE Transactions on Medical Imaging*, 21(3):239–250, 2002.
- [78] Q.Song. A robust information clustering algorithm. Technical report, Nanyang Technological University, Singapore, 2004.
- [79] Q.Song, W.J.Hu, and W.F.Xie. Robust support vector machine with bullet hole image classification. *IEEE Transactions on System, Man and Cybernetics*, 2002, In Press.
- [80] R.E.Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. on Information Theory*, IT-18(4):460–473, 1972.
- [81] R.Guliano, R.M.Rangayyan, W.A.Carnielli, J.A.Zuffo, and J.E.L.Desautels. Segmentation of breast tumors in mammograms by fuzzy region growing. In *Proc. 20th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society*, pages II:1002–1004, 1998.

- [82] R.J.Ferrari, R.M.Rangayyan, J.E.L.Desautels, R.A..Borges, and F.Frère. Automatic identification of the pectoral muscle in mammograms. *IEEE Trans. on Medical Imaging*, 23(2):232–245, 2004.
- [83] R.J.Hathaway and J.C.Bezdek. Fuzzy c-means clustering of incomplete data. *IEEE Trans. on Systems, Man, And Cybern.B*, 31:735–744, 2001.
- [84] R.Krishnapuram and J.M.Keller. A possibilistic approach to clustering. *IEEE Trans. on Fuzzy Systems*, 1(2):98–110, 1993.
- [85] R.Krishnapuram, O.Nasraoui, and H.Frigui. The fuzzy c spherical shells algorithm: a new approach. *IEEE Trans. on Neural Networks*, 3(5):663–671, 1992.
- [86] R.L.Egan. *Breast Imaging: Diagnosis and Morphology of Breast Diseases*. Saunders, Philadelphia, PA, 1988.
- [87] R.M.Gray. *Entropy and Information Theory*. Springer-Verlag, New York, 1990.
- [88] R.M.Gray. *Source Coding Theory*. Kluwer Academic Publisher, 1990.
- [89] R.M.Rangayyan, N.M.El-Faramawy, J.E.L.Desautels, and O.A.Alim. Measures of acutance and shape for classification of breast tumours. *IEEE Transactions on Medical Imaging*, 16:779–810, 1997.
- [90] R.M.Rangayyan, N.R.Mudigonda, and J.E.L.Desautels. Boundary modelling and shape analysis methods for classification mammographic masses. *Medical and Biological Engineering*, 38:487–496, 2000.
- [91] R.N.Dave. Robust fuzzy clustering algorithms. In *Second IEEE International Conference on Fuzzy Systems*, pages 2:1281–1286, 1993.
- [92] R.N.Dave and R.Krishnapuram. Robust clustering methods: a unitied view. *IEEE Trans. Fuzzy Systems*, 5:270–293, 1997.
- [93] S.Avidan. Support vector tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(8):1064–1072, 2004.
- [94] S.C.B.Lo, H.Li, Y.Wang, L.Kinnard, and T.Freedman. A mutiple circular path convolution neural network system for detection of mammographic masses. *IEEE Transactions on Medical Imaging*, 21(2):150–158, 2002.

- [95] S.C.Chen and D.Q.Zhang. Robust image segmentation using fcm with spatial constraints based on new kernel-induced distance measure. *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, 34(4):1907–1916, 2004.
- [96] S.Geman and D.Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
- [97] S.Gunn. *Matlab SVM toolbox*. <http://www.kernel-machines.org/>.
- [98] S.H.Leung, S.L.Wang, and W.H.Lau. Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. *IEEE Trans. Image Processing*, 13:51–62, 2004.
- [99] S.Kirkpatrick, C.D.Gelatt, and M.P.Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.
- [100] S.K.Kinoshita, P.M.A.Marques, A.F.F.Slates, H.R.C.Marana, R.J.Ferrari, and R.L.Villela. *Proc. 4th Int. Workshop Digital Mammography*. Nijmegen, The Netherlands, 1998.
- [101] S.Mitra, R.Castellanos, and S.Joshi. An adaptive deterministic annealing approach for medical image segmentation. *Fuzzy Information Processing Society, NAFIPS, 19th International Conference*, 2000.
- [102] S.M.Lai, X.Li, and W.F.Bischof. On techniques for detecting circumscribed masses in mammograms. *IEEE Trans. On Medical Imaging*, (4):377–386, 1989.
- [103] S.Selim and M.Ismail. K-means-type algorithms: a generalized convergence theorem and characterization of local optimality. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15, 1983.
- [104] S.Shoham. Robust clustering by deterministic agglomeration em of mixtures of multivariate t -distributions. *Pattern recognition*, 35(5):1127–1142, 2002.
- [105] Structure. <http://www.oncologychannel.com/breastcancer/breastanatomy.shtml>.
- [106] S.U.Lee, S.Y.Chung, and R.H.Park. A comparative performance study of several global thresholding techniques for segmentation. *Computer Vision, Graphics, Image Processing*, 52:171–190, 1990.

- [107] S.V.Engeland, P.Snoeren, J.Hendriks, and N.Karssemeijer. A comparison of methods for mammogram registration. *IEEE Transactions on Medical Imaging*, 22(11):1436–1444, 2003.
- [108] T.Hofmann and J.M.Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(1):1–14, 1997.
- [109] T.Hofmann, J.Puzicha, and J.M.Buhmann. Unsupervised texture segmentation in a deterministic annealing framework. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(8):803–818, 1998.
- [110] T.K.Lau and W.F.Bischof. Automated detection of breast tumors using the asymmetry approach. *Computers and Biomedical Research*, 24:273–295, 1991.
- [111] U.Raeth, D.Schlaps, B.Limberg, I.Zuna, A.Lorenz, G.V.Kaick, W.J.Lorenz, and B.Kommerell. Diagnostic accuracy of computerized b-scan texture analysis and conventional ultrasonography in diffuse parenchymal and malignant liver disease. *Journal Clinical Ultrasound* 13, pages 87–99, 1985.
- [112] V.N.Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.
- [113] V.N.Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
- [114] V.N.Vapnik. An overview of statistical learning theory. *IEEE Trans. on Neural Networks*, 10(5):988–999, 1999.
- [115] W.E.Polakowski, D.A.Cournoyer, S.K.Rogers, M.P.DeSimio, D.W.Ruck, J.W.Hoffmeister, and R.A.Raines. Computer-aided breast cancer detection and diagnosis of masses using difference of gaussians and derivative-based feature saliency. *IEEE Trans. Medical imaging*, 16:811–819, 1997.
- [116] W.Qian, L.P.Clark, M.Kallergi, and R.A.Clark. Tree-structured nonlinear filters in digital mammography. *IEEE Expert*, (1):25–36, 1994.
- [117] X.Q.Shen, M.Spann, and P.Nacken. Segmentation of 2d and 3d images through a hierarchical clustering based on region modelling. *Pattern Recognition*, 31(9):1295–1309, 1998.
- [118] Y.A.Tolias and S.M.Panas. On applying spatial constraints in fuzzy image clustering using a fuzzy rule-based system. *IEEE Signal Processing Letter*, 5(10):245–247, 1998.

-
- [119] Y.Yao, L.Chen, and Y.Q.Chen. Associative clustering for clusters of arbitrary distribution shapes. *Neural Processing Lett*, 14(3):169–177, 2001.
- [120] Z.Huo, M.L.Giger, C.J.Vyborny, U.Bick, P.Lu, D.E.Wolverton, and R.A.Schmidt. Analysis of spiculation in the computerised classification of mammographic masses. *Medical Physics*, 22:1569–1579, 1995.
- [121] Z.Tu and S.C.Zhu. Image segmentation by data-driven markov chain monte carlo. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24:657–673, 2002.