UNIVERSITY OF SOUTHAMPTON

Automatic Reference Region Localisation in Positron Emission Tomography

by

Jun L. Chen

A thesis submitted in partial fulfilment for the degree of Doctor of Philosophy

in the
Faculty of Engineering and Applied Science
Department of Electronics and Computer Science

October 2003

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND APPLIED SCIENCE
DEPARTMENT OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

Automatic Reference Region Localisation in Positron Emission Tomography

by Jun L. Chen

Positron emission tomography (PET) is a functional imaging technique that enables brain function to be measured *in vivo*. PET is a challenging application area for data modelling with four dimensional data in space (3D) and time (1D). The spatio-temporal data sets are typically quantified using tracer kinetic models. An improper reference tissue input function will bias the modelling result. This thesis addresses the automatic extraction of a reference tissue region, devoid of receptor sites, which can then be used as an input for a reference tissue model, allowing for the quantification of receptor sites. It is shown that this segmentation can be determined from the time-activity curves associated with each voxel within the 3D volume, using modern machine learning methods.

Previously, supervised learning techniques have not been considered in PET reference region extraction. In this thesis, two new methods are proposed to incorporate expert knowledge and the image models with the data: a hierarchical method and a semi-supervised image segmentation framework. Markov random field (MRF) models are used as a stochastic image model to specify the spatial interactions. The first method uses a Bayesian neural network with a hierarchical Markov random field model. The second method advances the first method by employing a semi-supervised image segmentation framework to combine the fidelity of supervised data with the quantity of unsupervised data. This is realised by a three-level image model structure with probability distributions specifying the interconnections. This has the advantages for the generalisation performance and hence the reduction of bias in PET reference region extraction. An Expectation Maximisation based algorithm is proposed to solve this combined learning problem. The performance of unsupervised, supervised and semi-supervised classification in temporal models and spatio-temporal models are compared, using both simulated and $[^{11}C](R)$-PK11195 PET data. In conclusion, it shows that the inclusion of expert knowledge greatly reduces the uncertainty in the segmentation with the new semi-supervised framework achieving substantial performance gains over the other methods.

Contents

List of Figures

List of Tables

Nomenclature

B_{max}	total concentration of specific binding sites
BP	binding potential
C_1	reference region TAC
C_2	non-reference region TAC
C_*	concentration of $*$
$C_R(t)$	reference tissue time-activity curve
$C_T(t)$	target tissue time-activity curve
f_2	"free fraction" of the unbound radioligand in the tissue
φ	regularisation term
F_i	concentration of competing endogenous ligands
L	loss function
λ	physical decay constant
K	number of cluster centres
$K(\cdot, \cdot)$	kernel function
k_2	efflux rate constant from the target tissue
K_{D_i}	equilibrium disassociation constants of competing endogenous ligands
$K_{D_{tracer}}$	equilibrium disassociation constant of the radioligand
m	number of clusters
n	number of data examples
n^l	number of labelled data examples
$p(\cdot)$	probability function
$P(\cdot)$	prior
σ^2	variance
R	risk functional
R_{emp}	empirical risk functional
R_{reg}	regularised risk functional
$\mathbb{R}^d$	d-dimensional Euclidean space
R_I	ratio of the delivery in the target tissue to the reference tissue
S_k	data subset
w	parameter vector
$\boldsymbol{\mu}$	mean
Σ	diagonal standard deviation vector

σ	standard deviation scalar
$\mathbf{x}$	input vector
y	output
z	hidden variable or cluster label
$\phi(\mathbf{x})$	basis function
ξ	slack variable
Ξ	likelihood

Acknowledgements

I would like to express my sincere thanks to Dr Steve Gunn for his continuous guidance and support during the whole part of this work. He has been a consistent source of ideas and encouragement. I have truly enjoyed and benefited from working with him. I would also like to thank Dr Mark Nixon for his supervision and encouragement. I am grateful to Dr Roger Gunn for his guidance, provision of data and boundless knowledge on PET, which is of uncountable value for this research.

Chapter 1

Introduction

1.1 Background

The human brain is amazing, enabling complex tasks to be performed in a diversity of environments. It is now possible to examine its anatomical structure *in vivo* using techniques such as X-rays, computerised tomography (CT) and magnetic resonance imaging (MRI). There are many circumstances where dynamical information is also needed, for example, in study of brain function and in the diagnosis of benignancy or malignancy of a brain tumour. The measurement of the electrical signals on the scalp using electroencephalography (EEG), opens up new possibilities in studying brain function. However, signals recorded in the scalp may not represent the activity in the underlying cortex. The advent of functional imaging modalities of single photon emission computerised tomography (SPECT), positron emission tomography (PET), functional magnetic resonance imaging (fMRI), and magnetoencephalography (MEG) over the last twenty years has led to a new era in the study of brain function.

Medical image processing benefits greatly from computer processing. In early stage, image processing was limited to 2D images. In 1987, the first attempts at fully automatic computer-aided diagnosis of X-ray mammograms were proposed (Chan et al. 1987). The advancement in computing alongside the development of new imaging and modelling techniques, has enabled medical image processing to be extended from 2D to 3D. The appearance of functional images gives a challenge to data modelling, as the data are often in 4D (3D in spatial domain and 1D in time domain). This new type of data is highly multivariate: there are typically 100,000-400,000 voxels in each observation and the observation varies with time. Statistical tools have been developed to analyse functional image data, such as statistical parametric mapping (SPM) (Friston et al. 1995). Recently, more scientists have begun to use Bayesian techniques in modelling functional image data (Genovese 2000; Svensen et al. 2000).

The earliest experiments to measure cerebral blood flow were performed in 1948 by Kety

and Schmidt (1948). The development of computer tomography in the 1970's allowed mapping of the distribution of the radioisotopes in the brain, and led to the development of SPECT imaging (Kuhl and Edwards 1963). The radiotracer used in SPECT emits gamma rays, as opposed to the positron emitters used in PET. The first PET scanner appeared in 1975. Positron emission tomography has two major advantages over SPECT, namely better spatial resolution and greater sensitivity (Fox et al. 1984). PET imaging involves injecting the human body with a radiolabelled compound and measuring the distribution of the radiotracer. The distribution of the radiolabelled compound measures how the body functions with that compound. Dynamic imaging involves acquiring a sequence of images that measure some temporal property of an object of interest. These changes relate to the biochemical and physiological interactions of a particular radiotracer within the human body. Using this imaging modality, information can be gathered about processes that occur over time within the human body. This information can then be used to study how organs function in normal and diseased states, and under external influences such as drugs or other therapeutic actions.

1.2 Motivation of This Research

The problem under investigation in this research can be stated as follows: given 4D PET dynamic images of a subject, how can the image to be segmented into regions based on their different dynamic behaviours? Apart from the application of various image segmentation techniques to PET, this thesis also investigates methods for increasing the segmentation accuracy by improving the generalisation in learning. Specifically, the combination of expert knowledge with the image data is investigated.

In 4D spatio-temporal PET data, the "intensity" of each voxel is a signal over time, representing the changing of tissue concentration. Signal modelling can often be categorised as parametric modelling and non-parametric modelling. In parametric modelling, an explicit model based on prior knowledge of the system, using a set of biological or physiological meaningful parameters which describe the process, is used. In non-parametric modelling, no assumption is made about the data generating system.

The reference region model (Lammertsma and Hume 1996) is a widely used parametric model in PET to produce the parametric images of binding potential and relative delivery. The model requires the reference region time-activity curve (TAC) as an input. The reference region TAC is often obtained from a non-parametric technique or from the co-registered Magnetic Resonance Image (MRI). Currently, most non-parametric reference region segmentation methods (Ashburner et al. 1996; Yap et al. 1996) used in PET belong to the class of unsupervised techniques, where no *a priori* knowledge is involved in the segmentation process. Both the number of underlying patterns and the final discrimination need to be determined manually. Another option, the co-registration with

MRI, means that an extra MRI is needed in every PET experiment, which is expensive and time-consuming. Additionally, this image co-registration between different modalities can be difficult (Kiebel et al. 1997). The two modalities may image very different properties: e.g. analytical information and functional information. There is no proved evidence that any links between the regions in MRI and the reference region in PET exist. Accordingly an efficient methodology to segment PET images into a reference region and a non-reference region is needed.

The work described in this thesis investigates the integration of knowledge from different sources to enhance the accuracy of PET image segmentation. Supervised learning techniques are applied to the characterisation of dynamic PET images, in combination with expert knowledge. A supervised neural network (Haykin 1999; Patterson 1996) is configured to learn the distinction between reference and non-reference regions using examples labelled by an expert. The resulting classifier is used to select the reference curve used in the reference model, producing an automatic system, which enables the generation of parametric maps of binding potential without human intervention and MRI co-registration.

Ideally an image segmentation method should be able to use knowledge from various available sources: expert knowledge and the knowledge embedded in the image to be segmented. By integrating knowledge across scans, we may enhance the robustness of the segmentation process and hence increase confidence in the extraction of the reference regions.

1.3 Contributions

The main contributions of this work are the application of statistical classification techniques to PET modelling and the proposition of two new methods to incorporate knowledge from different sources (such as expert knowledge and the image models) into the PET reference region extraction process: a hierarchical supervised method and a semi-supervised image segmentation framework. This work first introduces supervised and semi-supervised learning into PET modelling. Previously, most learning techniques considered in PET reference region segmentation belongs to unsupervised learning. The theoretical analysis is supported with experimental comparison of the classification and image segmentation techniques with simulated and real PET data showing performance gains of the proposed methods over previous methods. The main contributions are detailed below:

- The use of statistical classification techniques to PET modelling is investigated. Various data classification techniques such as supervised, unsupervised and semi-supervised methods are formulated. Several classification techniques such as Bay-

esian neural networks, Gaussian mixture modelling are applied to simulated and real PET data segmentation.

- Statistical classification techniques are used in image segmentation with the use of stochastic models - Markov random fields (Geman and Geman 1984). Markov random field models are very suited for modelling the image's pixel (voxel) local connections. The use of Markov random field models in classification enables image pixel correlations to be considered instead of the unrealistic independence assumption. The Markov random field model can be combined naturally into the unsupervised Gaussian mixture modelling, with optimisation methods such as simulated annealing and mean field annealing being used to find a global or approximated solution to the problem. The combination of a Markov random field model with supervised classification techniques can be carried out in a hierarchical manner.

- A new semi-supervised image segmentation framework is developed. Both the labelled data and unlabelled data are used in the learning phase, so that the classifier uses knowledge in the labelled samples as well as additional knowledge of the data distribution from the unlabelled data, which is very important when labelled data are sparse. Each image is modelled at three different levels: the observed image, the "mixture label" image and the "class label" image, where connections between different levels are described by probability distributions, enabling a posterior probability distribution of the data to be recovered. A Markov random field model is used to incorporate neighbourhood information into the learning process. All information is considered in an integrated framework instead of hierarchically.

- Segmentation techniques are applied to simulated and real PET data. In simulations, performance assessment is measured by the estimation error as the ground truth is known. These simulations confirm that expert knowledge can be integrated successfully with the learning, reducing the uncertainty in the segmentation, and improving segmentation accuracy. In real PET data, a test-retest scheme is used to compare different segmentation techniques.

The work in this thesis has contributed to the following publications:

- Jun L. Chen, Steve R. Gunn and Mark S. Nixon, Markov random field models for segmentation of PET images, Proceedings of 17th International conference on information processing in medical imaging (IPMI), Davis, USA, June, 2001, pp.468-474.

- Jun L. Chen and Steve R. Gunn, A model-based image segmentation framework using labeled and unlabeled data, Proceedings of Advanced Concepts for Intelligent Vision Systems, Germany, July, 2001, pp.112-126.

- Jun L. Chen, Steve R. Gunn, Mark S. Nixon, Ralph P. Myers and Roger N. Gunn, A Supervised Method for PET Reference Region Extraction, Proceedings of Medical Image Understanding and Analysis, London, UK, July, 2001, pp.179-182.

1.4 Scope of Thesis

The research undertaken as part of this thesis is concerned with the development and application of machine learning techniques to PET image segmentation to enhance the segmentation accuracy. The thesis first gives an introduction of PET imaging and the classification problem, followed by applying classification techniques to PET temporal reference region extraction. Then the new image segmentation techniques combined with expert knowledge are proposed with the application to PET spatio-temporal reference region extraction. The thesis is structured as follows:

Chapter 2: PET Dynamic Image Modelling

This chapter gives an introduction to PET imaging and describes the problem tackled in the thesis. The physical principles, data acquisition and data reconstruction processes are described. This gives a general idea of the PET data generation process and the quantity of data generated from a PET scan. The factors that limit spatial resolution of PET data are discussed. Relevant PET modelling techniques including compartmental models and reference region models, are discussed. The reference region model is of central interest since the segmentation of PET images to extract these reference regions is the central aim of this research.

Chapter 3: Data Classification

This chapter provides a short introduction to learning and pattern recognition. The fundamental goal of learning techniques is to maximise the generalisation performance of the learning machine. The mathematical formulations of unsupervised, supervised and semi-supervised classification methods are given. Various efficient classification techniques are illustrated. In unsupervised classification, the Gaussian mixture modelling method often outperforms other methods as it provides a tractable probabilistic representation. In supervised classification, neural networks and kernel machines (such as support vector machine) are often used. As both unsupervised and supervised classification uses only the knowledge in the labelled examples or the knowledge in the unlabelled data, semi-supervised classification provides methods for including knowledge from both sources. The methodology to combine both forms of knowledge remains an active research topic. The concept of induction and transduction learning is distinguished within both supervised and semi-supervised classification.

Chapter 4: Temporal PET Reference Region Extraction

This chapter deals with using unsupervised, supervised and semi-supervised pattern recognition techniques in PET image segmentation with the assumption of data independence. The PET reference region localisation problem is addressed using both simulated and real PET data. The segmentation results with different pattern recognition techniques are compared. In real PET data, a test-retest scheme is used to estimate performance. Parametric images of binding potential are generated by using a simplified reference region model. Additionally, as the subject's age is increasingly being recognised as an important factor influencing the brain's function, the binding potential's correlation with age is investigated to find possible connections between them.

Chapter 5: Image Segmentation

A Markov random field model is introduced to model the image's pixel interactions. The introduction of Markov random field models enables regularisation in low-level image segmentation. When the model parameters are known, the iterated continual mode (ICM) method can be used as an *ad hoc* iterative optimisation method. When joint image segmentation and parameter estimation is needed, Markov chain Monte Carlo, Mean field annealing and EM methods can be used to deal with the optimisation problem in unsupervised image segmentation.

Apart from modelling the image segmentation problem in an unsupervised manner, this chapter introduces the new idea of learning in image segmentation, i.e., combination of image models with expert knowledge in the image segmentation process. Two new methods are proposed to realise this. One method is to hierarchically use a supervised neural network and statistical image models to model image's pixel correlations. A new combined learning framework is also proposed to solve this difficult problem with improved accuracy. The new semi-supervised image segmentation scheme uses both labelled and unlabelled data as well as imposing local constraints on image pixels in the learning process, so that the classifier uses knowledge in the labelled samples as well as additional knowledge of the data distribution from the unlabelled data, which is important when labelled data are sparse. Each image is modelled at three different levels: the observed image, the "mixture label" image and the "class label" image, where connections between different levels are described by probability distributions, enabling a posterior probability distribution of the data to be recovered. A Markov random field model is used to incorporate neighbourhood interaction into the learning process. All information is considered in an integrated framework instead of hierarchically. A method based on Expectation-Maximisation is also presented to solve the difficult optimisation problem.

Chapter 6: Spatio-Temporal PET Reference Region Analysis

This chapter applies the spatio-temporal segmentation techniques to PET reference region extraction, with comparisons between unsupervised image segmentation, hierarchical supervised image segmentation and semi-supervised image segmentation on simu-

lated and real PET data. The segmentation results are compared with each other as well as the temporal segmentation results in chapter 4. The experimental results show that the use of expert knowledge and image's pixel local connections reduces the uncertainty in segmentation and improves the segmentation performance. Similar to chapter 4, parametric images of binding potential and the binding potential's correlation with age are also given.

Chapter 7: Conclusions and Future Work

The final chapter summaries the theoretical and experimental results and suggests directions for future work.

Chapter 2

PET Dynamic Image Modelling

As a dynamic imaging technique, PET enables investigating human body *in vivo*. This chapter gives a brief review of the physical principles of PET imaging, including the imaging and reconstruction process. The parametric models such as the simpilified reference region model will be described for analysing PET images.

2.1 Physical Principles and Data Acquisition

2.1.1 Physical Principles

Phelps et al. (1986) give a rigorous treatment of the theoretical and practical issues related to PET imaging. PET uses radio tracers to image human biological and chemical processes *in vivo* (Phelps and Gambhir 1993). A tracer is an analogue of a biologically active compound in which one of the atoms has been replaced by a radioactive atom. When the tracer is introduced into the body, its spatio-temporal distribution can be located by means of the radioactive atom. PET requires an on-site cyclotron to produce the short-lived radioisotopes.

All radioisotopes used with PET decay by positron emission. Positrons are positively charged electrons. They are emitted from the nucleus of some radioisotopes that are unstable because they have an excessive number of protons and hence a positive charge. Positron emission stabilises the nucleus by removing this positive charge through the conversion of a proton into a neutron.

A positron emitted from a decaying nucleus travels a short distance before colliding with an electron of a nearby atom. When a positron comes in contact with an electron, the two particles annihilate turning their mass to energy (via Einstein's equation the total energy released is $E = 2m_0c^2$). Conservation of energy yields two 511-keV gamma-rays that are emitted in opposite directions (see Fig. 2.1).

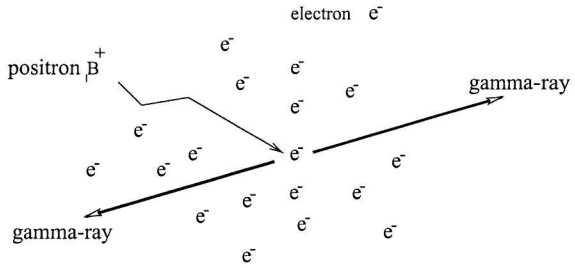


FIGURE 2.1: Positron and electron annihilation

2.1.2 Data Acquisition

The Gamma-rays which escape from the human body can be recorded by external detectors, as shown in Fig. 2.2. The PET detector is set up in such a way as to accept events in which both annihilation photons are detected in coincidence. Typically, two photons are identified as coming from a single event if they arrive at detectors within about 15ns of one another. Events are recorded between the currently many millions of detector pairs which represent line integrals or projections at different angles around the subject. The raw data set is termed a sinogram (a matrix of angles vs. projection).

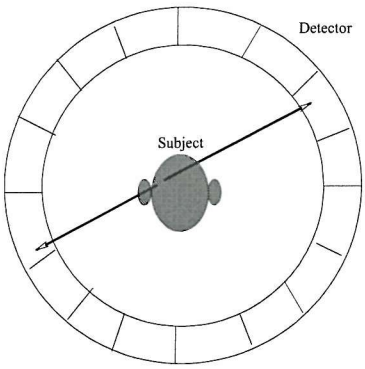


FIGURE 2.2: Detection of Gamma-rays

2.2 Reconstruction

From the raw data (sinogram) acquired by the tomography, it is desired to reconstruct the regional decay rates, as it indicates the tracer concentration. Image reconstruction is a rich research area where a large number of algorithms are available. Filtered back-projection (FBP) (Kak and Slaney 1988) is the reconstruction method routinely used in most PET scans. Accurate correction for scatter and attenuation is essential for the production of quantitative images. Each reconstructed image shows the spatial measure of tracer concentration at a certain time frame. In a PET scan, the images at different

time frames are reconstructed to form a dynamic image volume. Thus the information obtained from a PET scan is a spatio-temporal measure of the tracer concentration in the subject. Region of Interest (ROI) analysis or voxel level analysis of dynamic images can be used to produce tissue time-activity curves (TACs). These curves represent the counts/second/voxel (or kBq/ml) of the tracer concentration. TACs provide the key information for PET modelling and diagnosis as the kinetics of the tracer in tumours and normal tissues are significantly different.

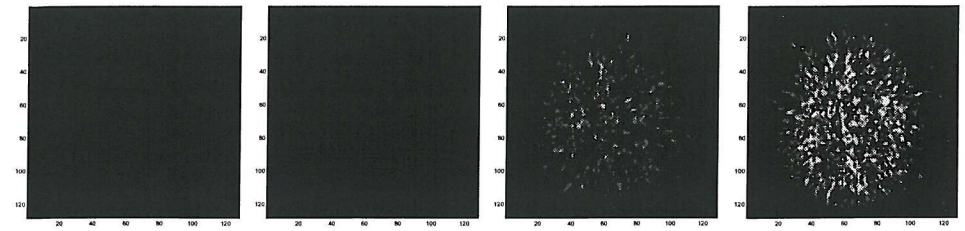
The spatial resolution of PET images is limited, mainly as a result of the following factors:

- A positron travels a long distance (a few mm) before annihilating with an electron.
- Absorption and scatter of the gamma rays occur in the tissue before detection in the detector rings.
- The size of the detector ring limits the spatial resolution of the reconstructed distributions.
- The decay rate of the isotope and radioactive safety levels limits the temporal resolution

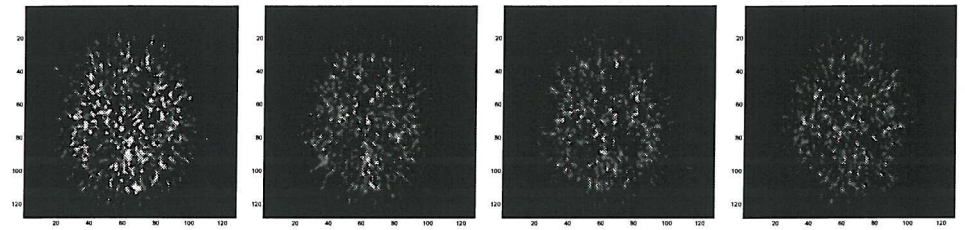
The random gamma-ray photons and the scattered photons recorded by the detector need to be removed to provide quantitative images.

A dynamic PET scan yields 4-D data (in space 3-D and time 1-D) which quantifies the distribution of the tracer over the period of scanning (typically 1-2 hours). Fig. 2.3 illustrates the reconstructed dynamic images for one slice from a PET scan. Each subfigure represents the tracer concentration in the slice during a short time period. In functional imaging techniques such as PET, the interest is in the dynamic image or correspondingly, for each voxel, the interest is the tracer concentration's change over time, which constitutes a time-activity curve. Fig. 2.4 shows an example of a TAC for one voxel.

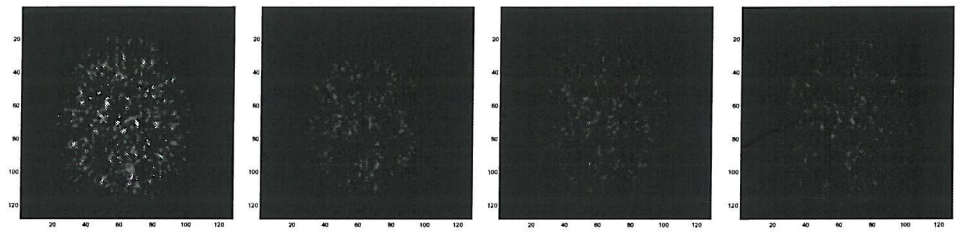
A PET experiment produces a large amount of data, with reasonably high noise, as shown in Fig. 2.4. Thus efficient modelling techniques are necessary for PET image analysis. The raw detector data gives little direct insight of what is happening inside the human body. This data must be reconstructed, de-noised and pre-processed to produce an interpretable form as in Fig. 2.3. Statistical modelling techniques can then be applied to enable biochemical or physiological meaningful parameters to be extracted.



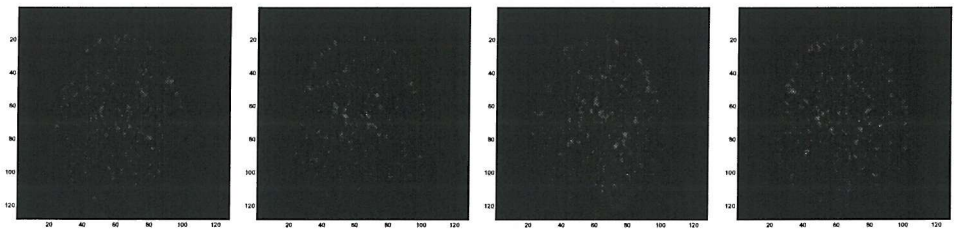
(a) 0-30s (b) 30-45s (c) 45-50s (d) 50-60s



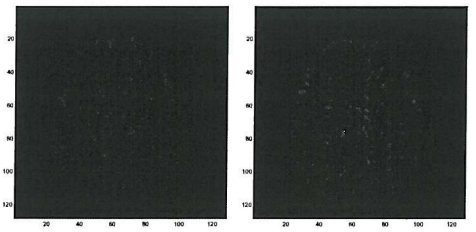
(e) 60-90s (f) 90-150s (g) 150-210s (h) 210-270s



(i) 270-330s (j) 330-630s (k) 630-930s (l) 930-1230s



(m) 1230-1530s (n) 1530-1830s (o) 1830-2130s (p) 2130-2430s



(q) 2430-3030s (r) 3030-3630s

FIGURE 2.3: $[^{11}C](R)$-PK11195 PET dynamic images at different time frames (one slice of PET scan No. n03578)

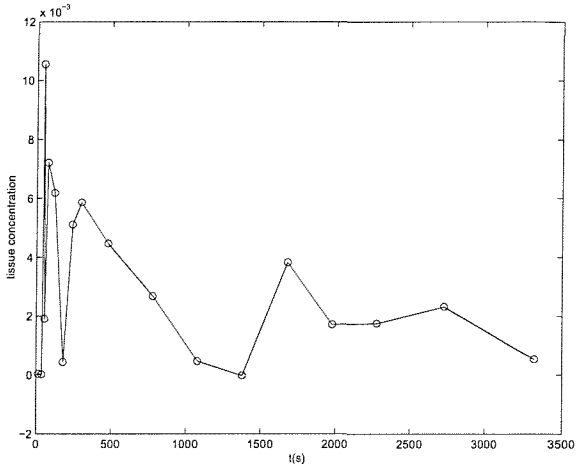


FIGURE 2.4: An example of a time-activity curve

2.3 Modelling

A goal in PET is to produce functional parametric images of the biological parameters of interest, and as such it is not the "raw" images that we are interested in, but rather some function of them. Different TACs correspond to different properties of the underlying tissue. Often a compartmental model (Gunn et al. 2001; Lammertsma and Hume 1996) is used to set up the relation between the TAC and the biological parameters. For each parameter, a 3-D (or 2-D) parametric image can be obtained to view the parameter for each voxel (or pixel).

Kinetic modelling is a very useful and widely used tool for analysis of living systems. In kinetic modelling, the system to be observed is modelled as a set of macroscopic subsystems, called compartments. It is assumed that each compartment is homogeneous and interacts with each other and the environment. Compartmental analysis forms the basis for tracer kinetic modelling in PET. A comprehensive analysis of compartmental models can be found in Jacquez (1985).

In the circumstance that the experiment conditions can be tightly controlled, very detailed models can be applied, for example in physiology experiments. However, as a nuclear imaging techniques, safety factors llmit the injected radioactive doses and hence the temporal resolution of PET imaging. Furthermore, the relative low temporal resolution and the noise in the PET data means that only a coarse model can be applied.

There are various parametric PET models in the literature, e.g., one tissue compartment models, two tissue compartment models with various number of input or unknown model parameters, three tissue compartment models, etc. These compartmental models are used for the quantification of blood flow (Kety and Schmidt 1948), cerebral metabolic rate of glucose (Sokoloff et al. 1977) and for neuroreceptor ligand binding (Mintun et al. 1984) etc. For neuroreceptor ligand modelling, compartmental models can be placed into

two groups: those requiring an arterial blood or plasma input function and "reference region models" which requires no blood sampling. Gunn et al. (2001) have developed a general theory for compartmental models in PET using state space representation, for both plasma input models and reference tissue models. Among PET models, the reference region model is the main interest in this thesis as it needs no blood sampling. The reference region model has been validated and applied successfully at the region of interest level and voxel level for other PET studies (Lammertsma and Hume 1996; Gunn et al. 1996).

2.3.1 Radioligand Binding

A radioligand is a radioactively labelled drug that binds with a receptor, transporter, enzyme, or any site of interest. Binding occurs when the ligand and receptor collide due to diffusion, and when the collision has the correct orientation and enough energy. Measuring the rate and extent of binding provides information on the number of binding sites, and their affinity and accessibility for various drugs.

Analysis of radioligand binding experiments is based on a simple model, called the law of mass action. This model assumes that binding is reversible, that is, the ligand and receptor are the same after dissociation as they were before binding (Yamamura 1990),

$$\boxed{\text{Receptor}} + \boxed{\text{Ligand}} \underset{K_{off}}{\overset{K_{on}}{\rightleftharpoons}} \boxed{\text{Receptor+ligand}}$$

where K_{on} is the on-rate (association rate) constant and K_{off} is the off-rate (dissociation rate) constant. Equilibrium is reached when the rate at which new ligandreceptor complexes are formed equals the rate at which the ligandreceptor complexes dissociate:

$$C_{Ligand} \cdot C_{Receptor} \cdot K_{on} = C_{Complex} \cdot K_{off}, \tag{2.1}$$

where C_* is the concentration of $*$. The equilibrium dissociation constant K_D is defined as

$$K_D = \frac{K_{off}}{K_{on}} = \frac{C_{Ligand} \cdot C_{Receptor}}{C_{Complex}} \tag{2.2}$$

A low K_D indicates the receptor has a high affinity for the ligand and it will take a low concentration of ligand in the experiment.

In addition to binding to receptors of interest, radioligands may also bind to other sites. Binding to the receptor of interest is called *specific binding*, while binding to the other sites is called *nonspecific binding*.

2.3.2 Reference Region Models and the Basis Function Method

A reference tissue model based on compartmental structures has been derived (Lammertsma and Hume 1996) and applied successfully on a region of interest (ROI) level. Consequently the model has been successfully used to quantify ligand-receptor binding at the voxel level (Gunn et al. 1996), using a basis function method.

2.3.2.1 A Two-Compartment Example

The reference region refers to the region in the subject which shows no binding, devoid of specific receptor sites. The compartmental structure for a simplified reference tissue model is shown in Fig. 2.5. The model describes the relation between the reference tissue TAC and the target tissue TAC. Given a reference tissue TAC, the model can be used to estimate the biological parameters associated with each target tissue TAC.

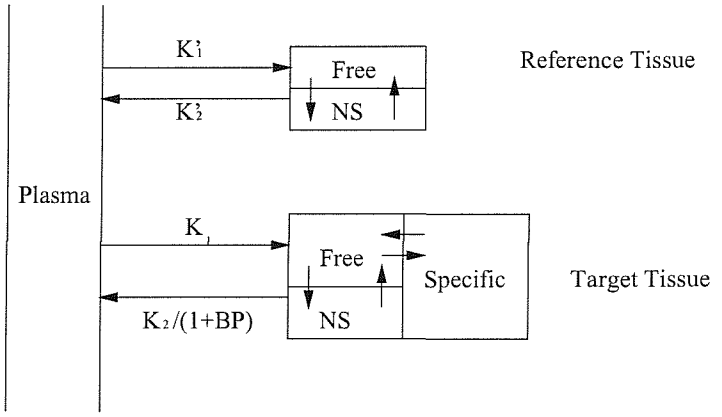


FIGURE 2.5: The compartmental structure for the reference tissue model (NS: nonspecific)

The model structure shown in Fig. 2.5 is based on the following assumptions for the radioligand: (1) there exists a reference region that is devoid of specific binding, (2) labelled metabolites of the parent tracer do not cross the blood-brain barrier, (3) the degree of nonspecific binding and the volume of distribution of the free/nonspecific compartment is the same in the reference and target tissues, and (4) the exchange between free/nonspecific and specific compartments is rapid such that a single compartment can approximate their behaviours. Under these assumptions the target tissue concentration may be expressed as,

$$C_T(t) = R_I C_R(t) + \left(k_2 - \frac{R_I k_2}{1+BP}\right) C_R(t) \otimes e^{-\left(\frac{k_2}{1+BP}+\lambda\right)t} \tag{2.3}$$

where $C_R(t)$ is the TAC in the reference tissue, $C_T(t)$ is the TAC in the target tissue, R_I is the ratio of the delivery in the target tissue to the reference tissue, k_2 is the efflux

rate constant from the target tissue, BP is the binding potential, λ is the physical decay constant of the isotope, and $\otimes$ is the convolution operator.

The model allows for parameter estimates of relative delivery and binding potential. Relative delivery, R_I, is defined as

$$R_I = \frac{K_1}{K_1'} = \frac{F(1 - e^{-PS/F})}{F'(1 - e^{-P'S'/F'})} \tag{2.4}$$

where F and F' are the blood flows in the target tissue and reference tissue, respectively, and PS and $P'S'$ are the permeability surface area products in the target tissue and the reference tissue respectively. Binding potential (BP) (Mintun et al. 1984) describes the potential of a specific membrane to interact with either a specific radioligand or a specific neurotransmitter. BP is defined here as

$$BP = \frac{k_3}{k_4} = \frac{B_{max} f_2}{K_{D_{Tracer}}(1 + \sum_i \frac{F_i}{K_{D_i}})} \tag{2.5}$$

where B_{max} is the total concentration of specific binding sites, $K_{D_{tracer}}$ is the equilibrium disassociation constant of the radioligand, f_2 is the "free fraction" of the unbound radioligand in the tissue, and F_i and K_{D_i} are the concentration and equilibrium disassociation constants of competing endogenous ligands.

2.3.2.2 Parameter Estimation

The PET TAC at the voxel level has a low signal to noise ratio. Direct application of the above reference region model at the voxel level using conventional least squares fitting to estimate parameters is slow and sensitive to noise. A basis function method (Gunn et al. 2001) has been proposed to find a fast and robust solution of Equation (2.3) at the voxel level,

$$C_T(t) = \theta_1 C_R(t) + \theta_2 C_R(t) \otimes e^{-\theta_3 t} \tag{2.6}$$

where $\theta_1 = R_I$, $\theta_2 = k_2 - \frac{R_I k_2}{1+BP}$ and $\theta_3 = \frac{k_2}{1+BP} + \lambda$. Defining the following set of basis functions

$$B_i(t) = C_R(t) \otimes e^{-\theta_{3i} t}. \tag{2.7}$$

The number and range of the discrete values for θ_{3i} should be decided in advance. The target tissue time-activity curve can then be expressed as,

$$C_T(t) = \theta_1 C_R(t) + \sum_{i=1}^{q} \theta_{2i} B_i(t) \tag{2.8}$$

, where q is the number of basis functions. θ_1 and θ_{2i} can be solved for by using standard weighted linear least squares fitting (Gunn 1996). However, the overcomplete basis can leads to an under-determined set of equations. Basis pursuit denoising (Gunn et al.

2001) can be used to determine a sparse selection of kinetic basis functions. Each $B_i(t)$ corresponds to one θ_{3i} which has been calculated in advance. After that R_I, BP and k_2 are easily computed, enabling the binding potential and relative delivery associated with each voxel to be obtained. It is then possible to produce parametric images of these quantities.

2.4 Reference Region Localisation

The above reference region model is widely used in PET parametric modelling. However, the use of the reference region model requires a reference region TAC as an input to the model. Although there is a certain degree of consistency in reference region TACs from different PET scans, the reference region TAC is dependent upon factors in the PET scan such as injection dose, subject, weight, etc. An improper reference region TAC chosen in a PET experiment will lead to biased results. To use the reference region model for a specific PET scan, an important step is to localise reference regions in that scan.

Currently, there are two types of methods to localise reference regions:

- Anatomical information from a co-registered Magnetic Resonance Image (MRI);
- Unsupervised techniques followed by manually choosing the closest TAC.

MRI can be used to define an anatomical region to be the reference region, usually the cortex or thalamus. Then the reference region in PET is obtained by co-registration of the MRI and PET images. This means an extra MRI image is required for each PET image, which is expensive and the image co-registration from different modalities can be very difficult.

An alternative way is to use cluster analysis (Ashburner et al. 1996; Boudraa et al. 1996; Kimura et al. 1999), where the dynamic PET data are partitioned into a small number of clusters, each described by a multivariate Gaussian distribution. The means of these Gaussian distributions represent the underlying TAC associated with each cluster. However, final discrimination in each cluster depends on expert knowledge, which is time-consuming since it must be repeated for each new image.

2.5 [^{11}C](R)-PK11195 PET Data

The PET data used in this thesis was obtained with the ligand [^{11}C](R)-PK11195 which is a marker for activated glial cells (Banati et al. 1999)). PK 11195 is the name commonly given to 1-(2-chlorophenyl)-N-(1-methylpropyl)-1-isoquinoline carboxamide. It is one of

the most powerful peripheral benzodiazepine binding (PBB) ligands known (Langer and Arbilla 1988). It is known to bind to PBB sites on activated microglia and macrophages in regions of active pathology in brain. It exhibits minimal binding in normal brains. However there is a large increase in PBB sites in the locality of brain lesions.

PK 11195 has been used as an imaging ligand in PET studies, where it identifies multiple sclerotic plaques, malignant gliomas and areas surrounding infarcted zones in stroke patients (Vowinckel et al. 1997; Benavides et al. 1988). Quantification of specific binding of this ligand in PET is necessary if it is to realise its potential in the longitudinal monitoring of disease progression. However, compared to other neuroreceptors, in PK 11195 PET studies, the reference region is very difficult to localise anatomically, as there is no obvious link between reference region and the anatomical regions. However the TACs in the normal brain tissue have similar kinetics, implying consistent behaviour of the ligand in the non-pathological tissue.

The 18-scan PET data set used in this thesis was kindly collected and provided by MRC cyclotron Unit, Hammersmith hospital in London. The 18 subjects considered are 17 normal volunteers and one patient. In the scans for normal volunteers, they would be expected to have a reference region represented by grey matter. Each scan contains 3-D spatial sampled images over 18 different time instants. Each 3-D image contains $128 \times 128 \times 25$ $(x \times y \times z)$ voxels with the resolution of $2.09mm \times 2.09mm \times 3.42mm$. In each scan, the TAC for each voxel was extracted from the dynamic PET images of tracer concentration. To produce the target data, the cortex, thalamus and cerebellum regions were labelled from a co-registered MRI image. Experimental evidence shows that the cortex region in all the investigated normal scans had no binding. Thus the TACs from the cortex region are treated as the reference region TAC, while TACs from other regions are treated as non-reference region TACs. Fig. 2.6 summarises the TACs in the scalp, thalamus, cerebellum and cortex region. The scalp region TAC is different from others in both the amplitude and the shape.

Table 2.1 lists the details of all eighteen [^{11}C](R)-PK11195 PET scans. Among all scans, there are four set of scans used for test-retest studies (shown as subject a, b, c, d in Table 2.1). Each test-retest study contains two scans from the same subject, who was scanned on two separate occasions. Additionally, the age of each healthy subject is recorded for age-related binding studies.

Examples of binding potential images of a healthy subject and a diseased subject at the voxel level are shown in Fig. 2.7 for comparison. The diseased subject (scan n02904) exhibits a high value of binding potential while the healthy subject shows minimal [^{11}C](R)-PK11195 binding.

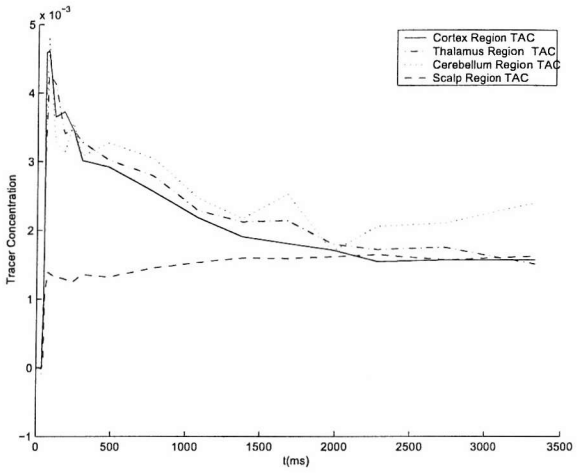


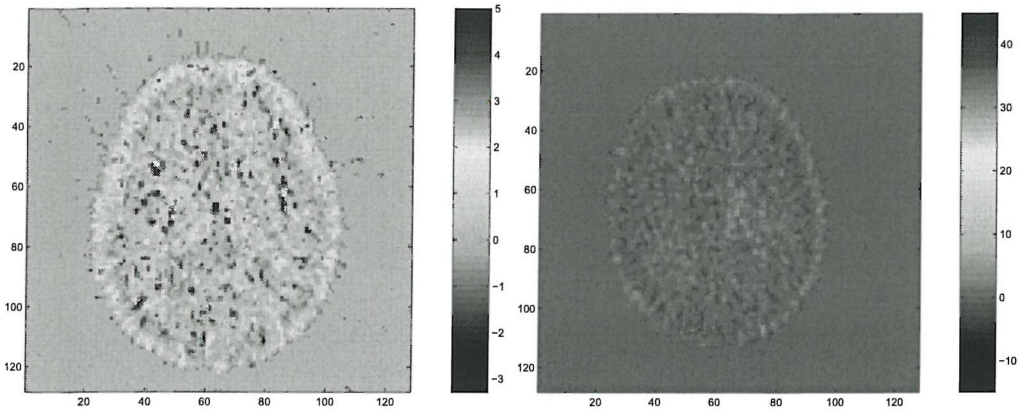
FIGURE 2.6: Example of mean TACs in different regions in a $[^{11}C](R)$-PK11195 normal scan

Scan No.	Disease	Age	Other Information
n02791	×	48	
n02805	×	54	
n02816	×	48	
n02833	×	53	
n02870	×	57	
n02904	✓		
n02907	×	80	
n02938	×	34	
n03578	×	59	subject a
n03637	×	78	
n03642	×	64	subject b
n03657	×	64	subject b
n03661	×	74	subject c
n03689	×	59	subject a
n03694	×	74	subject d
n04071	×	74	subject c
n04073	×	74	subject d
n04128	×	32	

TABLE 2.1: Data Set Description

2.6 Summary

Positron emission tomography uses radio tracers to image human biological and chemical processes *in vivo*. By using the tracer that is involved in the biological or chemical processes, the process can be quantified. In a PET experiment, the unstable nucleus emits a positron that will annihilate with a nearby electron. Annihilation produces two almost collinear gamma rays of 511keV that are detected in coincidence on either side of the active volume. The localisation of the annihilation allows the tracers spatio-temporal



(a) BP image of slice 20 of scan n03578 (b) BP image of slice 20 of scan n02904

FIGURE 2.7: Binding Potential (BP) images for a healthy subject and diseased subject

distribution to be determined. A PET experiment generates 4-D data sets where each voxel has a vector which describes how the tracer concentration changes with time, called a time-activity curve.

The data obtained from a PET scan is a realisation of a complex spatio-temporal process with many great variables and a significant noise component. The analysis of the PET data set is not a trivial problem as the raw data gives little insight of what is happening inside the human body. The parametric models (especially compartmental models) reviewed in this chapter enable the analysis of PET data using meaningful variables. These compartmental models fall into two groups, of which the reference region models are most attractive since these avoid the necessity of having an arterial blood or plasma input function.

The difficulty with reference region approaches is the determination of the reference TAC. This thesis is concerned with developing intelligent methods for this determination. The following chapters will present the machine learning methods and image segmentation methods in the view of solving the reference region localisation problem and the potential of tackling other functional image segmentation problems.

Chapter 3

Data Classification

Classification (Devroye et al. 1996) is the grouping of similar objects. To formalise a classification problem, it is assumed that objects belong to one of several classes. The task is to predict class identities based on observed features or properties of the objects. Classification is often termed pattern recognition in practical applications such as medical diagnosis, speech recognition, image understanding, handwritten recognition, and fault detection in machinery. These classification problems may be solved by humans in a seemingly effortless fashion while their solution using computers has often proved to be immensely difficult. Thus research has been carried out towards finding a theoretical way to solve this learning problem.

3.1 Learning and Classification

Learning is the process of estimating an unknown dependency or structure using a limited number of observations. The problem of a learning machine is to select a function $\widehat{y} = f(\mathbf{x}, \omega)$ that best approximates the system's response y (induction), and then use this dependency estimated between $\mathbf{x}$ and y to predict outputs for future input values (deduction), as illustrated in Fig. 3.1. The inductive step involves finding the optimum value of ω, the model parameters. The deductive step involves evaluating the function $\widehat{y} = f(\mathbf{x}, \omega)$ at $\mathbf{x}$. This two-step (induction/deduction) approach to learning is of our main interest, although a one-step (transduction) approach exists (Vapnik 1998; Devroye et al. 1996). The inductive step is an ill-posed problem: to form generalisations from finite particular examples (training data). The general inductive problem can be formulated as: given a finite set of observations $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ derived from an unknown joint probability density function (pdf) $p(\mathbf{x}, y) = p(\mathbf{x})p(y|\mathbf{x})$, how can the underlying functional relationship

$$y = f(\mathbf{x}, \omega) \tag{3.1}$$

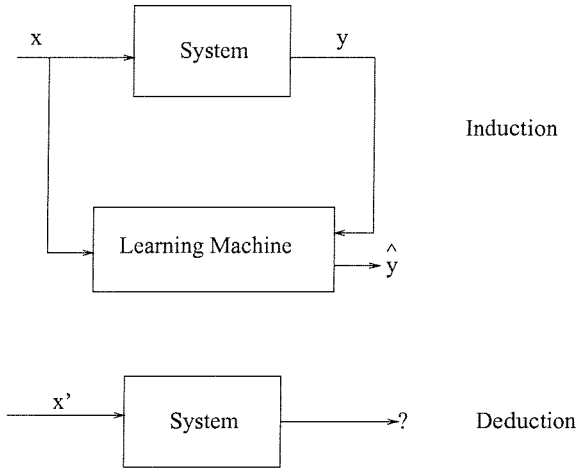


FIGURE 3.1: Induction/deduction formulation of learning

be found. A loss function $L(y, f(\mathbf{x}, \omega))$ is used to penalise the mismatch between the observations and the learning machine. The learning process could estimate the function $f(\mathbf{x}, \omega)$ by minimising the expected *risk functional*:

$$R(\omega) = \int \int L(y, f(\mathbf{x}, \omega)) \, p(\mathbf{x}, y) d\mathbf{x} dy \tag{3.2}$$

from the data set $\mathcal{D}$. It is an ill-posed problem due to the finite number of data. This general formulation of the learning problems includes three particular type of basic statistical problems (Vapnik 1998):

- Classification;
- Regression;
- Density Estimation.

In classification problems, the task is to assign an input $\mathbf{x}$ to one of a number of discrete classes y. In regression problems, the output y is a continuous variable. Both classification and regression problems can be seen as particular cases of function approximation. If the density $p(\mathbf{x}, y)$ can be estimated, then the regression and classification problem can be easily solved by minimising Equation (3.2). Hence the density estimation problem can been seen as a more general problem than classification and regression (Cherkassky and Mulier 1998). When solving a problem based on finite information, a common rule is: *Do not attempt to solve a specified problem by indirectly solving a harder general problem as an intermediate step* (Vapnik 1998). That is for a specified problem such as classification and regression, instead of estimating the joint probability $p(\mathbf{x}, y)$ directly, it is better to estimate only some features of the joint density which are critical for solving the specific problem.

Among three types of learning problem, classification is our main interest. Classification is often categorised into supervised classification and unsupervised classification (Bishop 1995). In unsupervised learning, the learning must proceed on the distribution of the patterns in the input space alone. The goal of unsupervised classification is to model the probability distribution of the input data and to use this to segment the input space into regions. The resulting model consists of a set of distributions. An expert can then decide how to assign these clusters. Each time a new data set appears, the whole learning process must repeat, and as such no information is aggregated from previous learning. In supervised classification, the desired output is known for each input pattern and the system can learn from these examples to generalise for "unseen" input patterns. Supervised learning differs from unsupervised learning in that a teacher is used to instruct the system with known examples, e.g. in a two-class classification problem, positive and negative examples of objects belonging to these classes. An inductive process is used to build up a model from the examples, producing a system which can determine the class membership of an input pattern. The merit of supervised learning is that it exploits the knowledge provided by the teacher in learning to discriminate.

Often both labelled examples and unlabelled data exist and recently research (Miller and Uyar 1996; Jaakkola and Haussler 1998) has been conducted to build hybrid unsupervised and supervised classification methods, termed semi-supervised classification methods.

3.2 Generalisation

Due to the finiteness of the data set $\mathcal{D}$, the expected risk functional in Equation (3.2) cannot be accurately evaluated.Instead it can be approximated by the empirical risk functional:

$$R_{emp}(\omega) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i, \omega)). \tag{3.3}$$

This approach causes two types of error, namely the approximation error and the estimation error (Niyogi and Girosi 1996), shown in Fig. 3.2. The approximation error arises because the exact behaviour of function $f(\mathbf{x})$ is unknown and belongs to some large space of functions, called the target space. Consequently, $f(\mathbf{x})$ has to be estimated by parameterised functions $f(\mathbf{x}, \widehat{\omega})$, and the resulting sub-space is referred to as the hypothesis space. The predictor $\widehat{f}(\mathbf{x}, \widehat{\omega})$ obtained by empirical error minimisation will approximate $f(\mathbf{x}, \widehat{\omega})$ as the number of data increases without bound. The approximation error decreases as the size of the hypothesis space increases and has more capacity to approximate $f(\mathbf{x})$. The estimation error comes from lack of knowledge about the conditional distribution $p(y|\mathbf{x}, \omega)$, which is unknown. The best that can be done is to minimise the empirical risk, as only finite number of examples are available. It is expected that as the model structure has more capacity, the estimation error can increase

(Niyogi and Girosi 1996).

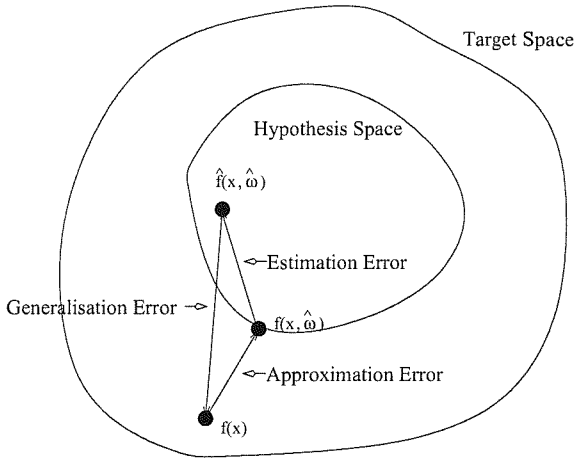


FIGURE 3.2: Generalisation error = Approximation error + Estimation error

A learning machine's performance is determined by its generalisation ability or its predictive capability. Estimation of a function from a finite data set is ill-posed, which makes precise estimation impossible. In order to make the problem well-posed, a prior, or structure, over the hypothesis space must be defined. A prior that is often consistent with the physical world is the characteristic of smoothness. Following the well known principle of *Occam's razor*[1], simpler models are preferred. A good model should have a trade-off between the empirical error and the model complexity. Thus instead of using the empirical risk, a regularised risk functional is often used to control generalisation performance,

$$R_{reg}(\omega) = R_{emp}(\omega) + \nu\varphi(f(\mathbf{x},\omega)), \tag{3.4}$$

where $\varphi(\cdot)$ is called a regularisation term. A common form for the regulariser is a measure of smoothness, assigning greater likelihood to smoother functions. The regularisation parameter ν controls the tradeoff between $R_{emp}(\omega)$ and $\varphi(f(\mathbf{x},\omega))$. Consider the two-class classification problem in Fig. 3.3, the classification with solid line causes over-fitting while the dash line is preferred.

Bayesian learning (Mackay 1992a; Neal 1994) gives a different view of selecting weights. Regularisation can be given a natural interpretation in the Bayesian framework. An advantage of the Bayesian method is that the values of regularisation coefficients can be selected using the training data. Furthermore, a Bayesian framework provides an objective and principled framework for dealing with the issues of model complexity which avoids many of the problems which arise when using maximum likelihood.

After a risk functional has been set up, an optimisation procedure is used to minimise the risk functional with respect to the adjustable parameters. Non-linear optimisation is a

[1]William of Occam (1285-1349): Causes should not be multiplied beyond necessity.

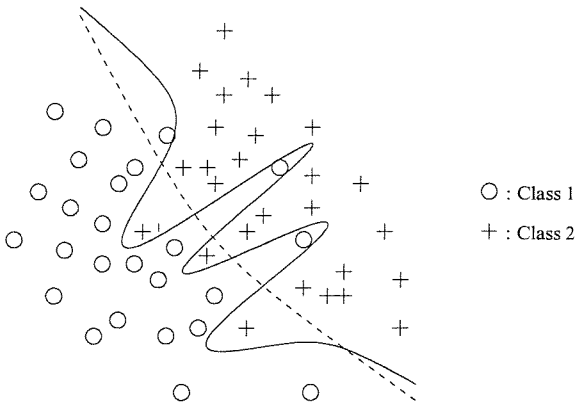


FIGURE 3.3: Over-fitting in a classification problem

very difficult problem and often only local minimum can be found. However, careful construction of the hypothesis space, and the regulariser can result in convex, even quadratic functions for which unique global solutions can be obtained, e.g. SVMs (Vapnik 1998), Gaussian processes (Barber and Williams 1996). The expectation-maximisation (EM) method (Dempster et al. 1977) and the gradient decent related method play a central role for optimisation in unsupervised learning and supervised learning respectively.

3.3 Unsupervised Classification

Unsupervised classification aims to separate a data set into a number of groups (called clusters) based on some measure of similarity. The goal is to find a set of clusters for which samples within a cluster are more similar than samples from different clusters. The task of clustering can fall outside of the framework of predictive learning, since the goal is to cluster the data at hand instead of providing an accurate characterisation of future data generated from the same probability distribution.

Unsupervised classification is a difficult problem, as the clusters may have various sizes and shapes and the number of clusters is unknown and needs to be specified by user. But it has wide applications in the real world.

As no known target y exists in unsupervised classification, learning aims to find K centres, $\boldsymbol{\mu}_k$, $k = 1, \ldots, K$, to describe the distribution of the data. In this case, the problem can be expressed as

$$R(\boldsymbol{\mu}_k) = \int \min_k \|\mathbf{x} - \boldsymbol{\mu}_k\|^2 p(\mathbf{x}) d\mathbf{x} \tag{3.5}$$

In the following sections, two efficient and widely used clustering algorithms: K-means and Gaussian Mixture Model (GMM) are introduced.

3.3.1 K-means Clustering

The K-means clustering algorithm (MacQueen 1967) is a well-known unsupervised algorithm. The algorithm involves a simple re-estimation procedure. Suppose we wish to find a set of K representative vectors $\boldsymbol{\mu}_k$ where $k = 1, \ldots, K$ for n data points x_i, $i = 1, \ldots, n$. The algorithm seeks to partition the data points $\{x_i\}_{i=1}^n$ into K disjoint subsets. S_k containing n_k data points, in such a way as to minimise the sum-of-squares clustering function given by

$$R_{emp}(\boldsymbol{\mu}_k) = \sum_{i=1}^{n} \min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2, \tag{3.6}$$

where $\boldsymbol{\mu}_k$ denotes the means of the data points in set S_k and is given by

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{i \in S_k} \mathbf{x}_i \tag{3.7}$$

Each point is re-assigned to a new set according to which is the nearest mean vector. The means of the set are then recomputed. This procedure is continued until there is no further change in the grouping of the data points. The resulting representation is one of the K vectors, which can be used to partition the input space into K regions or as the basis for further algorithms. The number of centres must be chosen in advance.

3.3.2 Mixture Models

Mixture models (MaLachlan and Basford 1988) are a flexible and powerful probabilistic modelling tool. As in the K-means algorithm, n data points $\mathbf{x}_i, i = 1, \ldots, n$ are to be classified into K clusters, where K is specified in advance. Consider the following scheme where these data are generated: There are K random sources, each characterised by a probability function $p(\mathbf{x}_i|k)$, $k = 1, 2, \cdots, K$ with mixing parameters $P(k)$, such that the overall distribution is given by $p(\mathbf{x}_i)$. $p(\mathbf{x}_i)$ is formed from the linear combination of component densities $p(\mathbf{x}_i|k)$,

$$p(\mathbf{x}_i) = \sum_{k=1}^{K} p(\mathbf{x}_i|k)P(k), \tag{3.8}$$

where $P(k)$ can be regarded as the prior probability for the data having been generated from the kth component of the mixture which are chosen to satisfy the constraints

$$\sum_{k=1}^{K} P(k) = 1 \tag{3.9}$$

$$0 \leq P(k) \leq 1. \tag{3.10}$$

The individual density component $p(\mathbf{x}_i|k)$ are normalised such that

$$\int p(\mathbf{x}_i|k)d\mathbf{x}_i = 1. \tag{3.11}$$

According to Bayes' theorem, the corresponding posterior probabilities are given by,

$$p(k|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|k)P(k)}{p(\mathbf{x}_i)}. \tag{3.12}$$

The value of $p(k|\mathbf{x}_i)$ represents the probability that a particular component k is responsible for generating the data point $\mathbf{x}_i$, which can be used to classify $\mathbf{x}_i$ to a certain class k. It is usually assumed that all the components $p(\mathbf{x}|k)$ have the same functional form, such as multivariate Gaussian. Fitting a mixture model to a set of observations $\{x_i\}_{i=1}^n$ consists of estimating the set of mixture parameters that best describes this data set.

Adopting a framework of parametric statistics, the detection of data clusters reduces mathematically to the problem of how to estimate the parameters of the probability density for a given mixture model. A powerful statistical tool for finding mixture parameters is the maximum likelihood method, i.e., one maximises the probability of the independently, identically distributed set $\{\mathbf{x}_i, i = 1, ..., n\}$ given a particular mixture model. For analytical purposes, it is more convenient to minimise the negative log-likelihood with respect $p(\mathbf{x}_i|k)$

$$E = -\sum_{i=1}^{n} \ln p(\mathbf{x}_i) = -\sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} p(\mathbf{x}_i|k)P(k) \right). \tag{3.13}$$

The parameter estimation in mixture models is often carried out by maximum likelihood learning using the Expectation-Maximisation method or the Markov chain Monte Carlo method.

3.3.2.1 EM Algorithm

The Expectation-Maximisation (EM) algorithm (Dempster et al. 1977) is the standard technique for maximum likelihood estimation of the parameters in mixture models. In general, the EM algorithm is used to solve the maximum likelihood estimation from incomplete data. The EM algorithm is a simple, practical method for estimating the mixture parameters which avoids the complexities of non-linear optimisation. An iteration of these two steps renders the following algorithm:

- E-step: Guess some values for the parameters of the mixture model ('old' parameter values). This yields the expected assignments of data to mixture components. Then with respect to this compute the expectation in Equation (3.13).

- M-step: "new" parameter values are found by minimising the expected error with respect to the "old" parameters. These parameter values then become the "old" values in the E-step.

By using Jensen's inequality, EM achieves a bound maximisation. The idea in the EM algorithm is that to find the model parameters that maximise the likelihood. Since the likelihood is unknown, its current expectation is maximised given the observed data and the current parameter fit. Provided some care is taken over the way in which the updates are performed, this algorithm is guaranteed to decrease the error function at each iteration, until a local minimum is found.

The algorithm can be modified to the generalised EM method (Neal and Hinton 1998): in each M step, the likelihood is increased but not necessarily maximised. The main difficulty in using EM for mixture models is local minimum, which makes its performance critically dependent on initialisation.

3.3.2.2 EM Algorithm for Gaussian Mixture Models

The mixture model is called a Gaussian Mixture Model (GMM) when the individual component densities are given by multivariate Gaussian distribution functions:

$$p(\mathbf{x}_i|k) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right), \tag{3.14}$$

where d is the dimensionality of input vector $\mathbf{x}_i$. $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$ are the mean and diagonal standard deviation of the kth Gaussian distribution. $|\mathbf{\Sigma}_k|$ is the determinant of $\mathbf{\Sigma}_k$.

Although mixture models can be built from different types of components, the majority of the literature focuses on Gaussian mixtures (Titterington et al. 1985). The K-means clustering algorithm can be seen as a particular limit of the Gaussian mixture model as the variance in Gaussian distribution approaches zero (Bishop 1995). Therefore, the K-means algorithm is a sensible choice to initialise the cluster centers in Gaussian mixture modelling.

Let $\mathbf{z} = \{\mathbf{z}_i\}_{i=1}^n$ denote the cluster index of data. After initialising parameters $\Phi^{(0)} = \{(\boldsymbol{\mu}_1^{(0)}, \mathbf{\Sigma}_1^{(0)}), (\boldsymbol{\mu}_2^{(0)}, \mathbf{\Sigma}_2^{(0)}), \cdots, (\boldsymbol{\mu}_K^{(0)}, \mathbf{\Sigma}_K^{(0)})\}$ and $p(z_i^{(0)=k}|\Phi^{(0)})$, the algorithm iterates the following two steps:

- E step: find the function $Q(\Phi|\widehat{\Phi}^{(t)}) = E[\log p(\mathbf{x}, \mathbf{z}|\Phi)|(\mathbf{x}, \widehat{\Phi}^{(t)})]$
- M step: find $\widehat{\Phi}^{(t+1)} = \arg\max Q(\Phi|\widehat{\Phi}^{(t)})$

In the M step, the parameter Φ is updated by (in the following $p(\cdot)$ is a simplified

notation of $p(\cdot|\Phi)$):

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)}, \tag{3.15}$$

$$\left(\boldsymbol{\Sigma}_k^2\right)^{(t+1)} = \frac{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})^T(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})}{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)}, \tag{3.16}$$

where

$$\begin{aligned} p(z_i^{(t)} = k|\mathbf{x}_i) &= \frac{p(\mathbf{x}_i|z_i^{(t)} = k)P(z_i^{(t)} = k)}{p(\mathbf{x}_i)} \\ &= \frac{p(\mathbf{x}_i|z_i^{(t)} = k)P(z_i^{(t)} = k)}{\sum_{i=1}^{n} p(\mathbf{x}_i|z_i^{(t)} = k)P(z_i^{(t)} = k)}. \end{aligned} \tag{3.17}$$

$p(\mathbf{x}_i|z_i^{(t)} = k)$ is calculated from Equation (3.14). The prior can be updated by

$$p(z_i^{(t+1)} = k) = \frac{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)}{n}. \tag{3.18}$$

The above Gaussian mixture model with EM algorithm for parameter optimisation is widely used in the unsupervised classification. The maximum likelihood estimation can also be achieved by Markov chain Monte Carlo method (Neal 1993; Richardson and Green 1997) in a fully Bayesian flavour to find a global solution instead of a local minimum in EM algorithm. But MCMC is computation expensive and thus is less used in unsupervised classification.

A problem in unsupervised classification is determining the appropriate number of clusters. The maximum likelihood criterion cannot be used to find the optimal number of clusters. There are several model selection criteria in literature: minimum description length (MDL) (Rissanen 1987; Barron et al. 1998), Bayesian inference criterion(BIC)(Schwarz 1978; Whindham and Cutler 1992), Akaike's information criterion (AIC) (Akaike 1974), Minimum message length (MML)(Sclove 1983) and reverse jumping MCMC (Green 1995). These techniques are employed with the attempt to reduce the number of parameters in the model while maintaining a reasonable performance. Among these methods, BIC and AIC are widely used. If the complexity of the true model does not increase with the size of the data set (such as the above K-means and Gaussian mixture model clustering), BIC is the preferred criterion, otherwise AIC is preferred (Burnham and Anderson 1998). Use of AIC criterion generally results in a good fit to the dataset, but it often causes an overestimation of the number of components. BIC is a likelihood criterion penalised by the model complexity, i.e. the number

of parameters in the model. Let $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{n}$ be the data set and $\mathcal{M} = \{M_i\}_{i=1}^{K}$ be the candidates for the parametric models. $\mid M_i \mid$ is the number of parameters in the model M_i, then the BIC criterion is defined as

$$BIC(M_i) = \log \Xi(\mathbf{x}, M_i) - \frac{1}{2} \mid M \mid_i \times \log(n) \tag{3.19}$$

where $\Xi(\mathbf{x}, M_i)$ is the likelihood of the data with model M_i. Reverse jump MCMC methods have been used to improve GMMs by splitting and merging clusters (Williams 2000). But as a sampling method, its computation is expensive.

3.4 Supervised Classification

Supervised classification is often referred to as "pattern recognition", which was formulated in late 1950s. A supervisor observes occurring situations and determines to which of k classes each one of them belongs. It is required to construct a machine which, after observing the supervisor's classification, carries out the classification in the same manner as the supervisor. The induction-deduction supervised learning diagram is shown in Fig. 3.4. In the induction process, the labelled data and the prior knowledge are used to build a model. Once the model has been built, it can be used in the deduction process to predict output for the testing data.

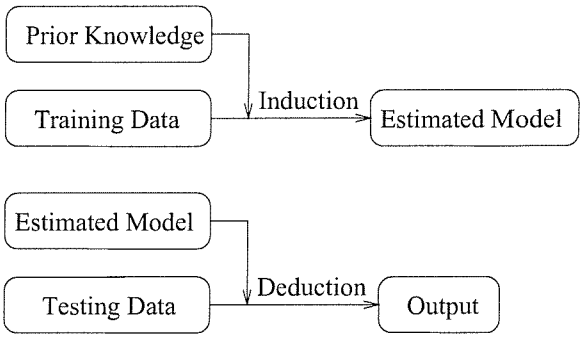


FIGURE 3.4: Induction-deduction supervised learning

In a two-class supervised classification problem, the output y can take either of two values $\{0, 1\}$, each denoting one of the two classes. The problem is to find the classifier $f(\mathbf{x}, \omega)$ that minimises the risk of misclassification:

$$R(\omega) = \int \int L(y, f(\mathbf{x}, \omega)) p(\mathbf{x}, y) d\mathbf{x} dy \tag{3.20}$$

The binary loss function $L(y, f(\mathbf{x}, \omega))$ is

$$L(y, f(\mathbf{x}, \omega)) = \begin{cases} 0 & \text{if } f(\mathbf{x}, \omega) = y \\ 1 & \text{if } f(\mathbf{x}, \omega) \neq y. \end{cases} \tag{3.21}$$

In many supervised classification problems, the cross-entropy loss function (Bishop 1995) of the form

$$L(y, f(\mathbf{x}, \omega)) = -y \ln f(\mathbf{x}, \omega) - (1 - y) \ln(1 - f(\mathbf{x}, \omega)) \tag{3.22}$$

is used. The use of this cross-entropy loss function in a two-class problem enables the classifier output $y(\mathbf{x})$ to represent the probability of belonging to one class (Hampshire and Pearlmutter 1990).

3.4.1 Classifier Type

There are a large number of various classifiers in the literature. Three widely used classifiers are introduced in this section.

3.4.1.1 Multi-Layer Perceptron

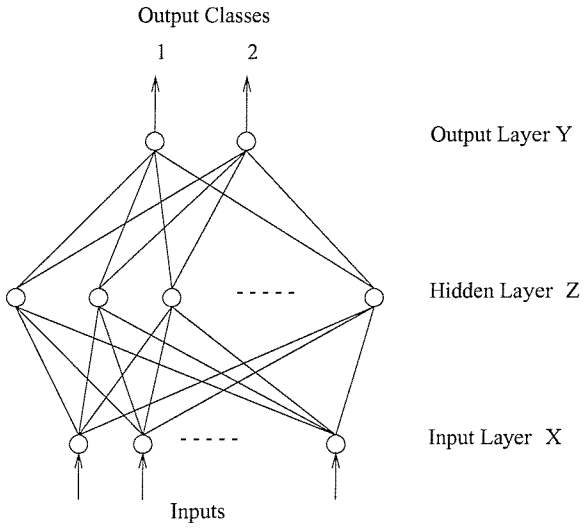


FIGURE 3.5: Architecture of a three-layer feed-forward classfier

The most widely used neural classifier is a Multi-Layer Perceptron (MLP) network which has been extensively analysed and for which many learning algorithms have been developed. The MLP classifer consists of a network of processing elements or nodes arranged in layers. Typically it requires three or more layers: an input layer which accepts the input variables used in the classification procedure, one or more hidden layers, and an output layer with one node per class. The output of hidden layer and output layer is obtained by transforming a weighted linear sum with a non-linear activation function. Fig. 3.5 shows the structure of a three-layer network. The unit in the hidden layer j has the form:

$$z_j = g\left(\sum_{i=0}^{d} w_{ji}^{(1)} x_i\right). \tag{3.23}$$

The output layer has the form:

$$y_k = g'\left(\sum_{j=0}^{M} w_{kj}^{(2)} z_j\right), \tag{3.24}$$

where $w_{ji}^{(1)}$ denotes a weight in the first layer, going from input i to hidden layer j and $w_{kj}^{(2)}$ denotes a weight in the second layer, going from hidden layer j to output k. $g(\cdot)$ and $g'(\cdot)$ are activation functions of hidden layer and output layer.

For the classification problem, the network can be trained by minimising the cross-entropy error using error back-propagation (Werbos 1994). When data from an input pattern is presented to the input layer of a trained network, the network nodes perform calculations in the successive layers until an output value is computed at each of the output nodes. This output signal should indicate which is the appropriate class for the input data.

3.4.1.2 Radial Basis Functions

The Radial Basis Functions (RBF) classifier (Broomhead and Lowe 1988) has the following form:

$$y(\mathbf{x}) = \sum_{j=1}^{K} w_j \phi_j(\mathbf{x}) + w_0 \tag{3.25}$$

where $\mathbf{x}$ is the input vector, y is the output, w_j is the weight and w_0 is bias. $\phi_j(\mathbf{x})$ denotes a local radial basis function, typically the local basis function is Gaussian:

$$\phi_j(\mathbf{x}) = \frac{1}{(2\pi\sigma_j^2)^{d/2}} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2\sigma_j^2}\right) \tag{3.26}$$

Training radial basis function networks involves a two-stage training procedure:

- Determine the parameters of basis function $\phi_j(x)$ using an unsupervised technique (Use only the input data and not the target data).
- Determine remaining weights using standard linear methods. The target values of training data are only used in this stage.

There are many potential applications where unlabelled input data are plentiful, but where labelled data are in short supply as the labelling of the data with target variables may require the time of a human expert which therefore limits the amount of the data which can be labelled within a reasonable time. With such applications, the two-stage training process for a radial basis function is particularly advantageous since the determination of the nonlinear representation can be done on a large quantity of unlabelled

data, leaving a relatively small number of parameters to be determined using the labelled data.

3.4.1.3 Support Vector Machines

A view of the regularisation method can be obtained from statistical learning theory, where an induction principle called *structural risk minimisation* (Vapnik 1998; Cristianini and Shawe-Taylor 2000) is presented.

The support vector machine (SVM) approach is motivated by results of statistical learning theory (Vapnik 1998). It is based on the Structural Risk Minimization (SRM) principle that is rooted in VC (Vapnik-Chervonenkis) dimension theory.

Given a set of examples

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), \mathbf{x} \in \mathbb{R}^d, y \in \{-1, +1\}, \tag{3.27}$$

where $\mathbb{R}^d$ denotes d-dimensional Euclidean space. The goal of learning is to find the decision function $f : \mathbb{R}^d \rightarrow \pm 1$ which provides the smallest *risk*

$$R(f) = \int \frac{1}{2}|f(\mathbf{x}) - y| dP(\mathbf{x}, y) \tag{3.28}$$

where P is the unknown distribution that data are drawn from. To realise this, the straightforward approach is to minimise the *empirical risk*

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}|f(\mathbf{x}_i) - y_i|. \tag{3.29}$$

The main idea of a support vector machine is to construct a hyperplane as the decision surface in such a way that the margin of separation between positive and negative examples is maximised (Gunn 1998). The set of vectors is said to be optimally separated by the hyperplane if it is separated without error and the distance between the closest vector to the hyperplane is maximal, as shown in Fig. 3.6.

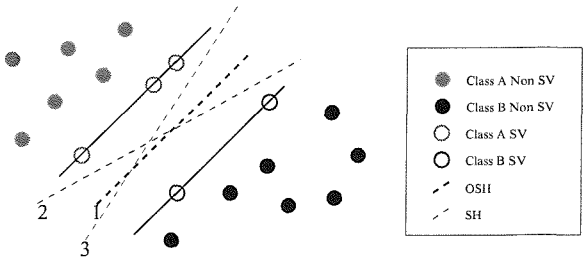


FIGURE 3.6: SVM optimal hyperplane

For examples given in Equation (3.27), we want to find a decision function

$$f_{w,b} = sgn((\mathbf{w}^T\mathbf{x}) + b) \tag{3.30}$$

with the property

$$sgn((\mathbf{w}^T\mathbf{x}_i) + b) = y_i, \tag{3.31}$$

and

$$\min_{\mathbf{x}_i} |\mathbf{x}_i^T\mathbf{w} + b| = 1. \tag{3.32}$$

If this function exists, then

$$y_i \cdot ((\mathbf{w}^T\mathbf{x}_i) + b) \geq 1. \tag{3.33}$$

In the linearly separable case, the optimal separating hyperplane which generalises well can be found by minimising the regularized risk functional (This is equivalent to minimising the VC dimension.)

$$\Psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \tag{3.34}$$

subject to Equation (3.33). To solve this convex optimisation problem, one introduces a Lagrangian

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i(y_i((\mathbf{x}_i^T\mathbf{w}) + b) - 1) \tag{3.35}$$

with multipliers $\alpha_i \geq 0$. The Lagrangian has to be minimised with respect to $\mathbf{w}, b$ and simultaneously maximised with respect to α_i. This means at the saddle point,

$$\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0, \frac{\partial}{\partial b} L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0. \tag{3.36}$$

This leads to

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \tag{3.37}$$

and

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.38}$$

By solving its dual problem, the solution is given by:

$$\boldsymbol{\alpha}^* = \arg\min \left(\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j) - \sum_{i=1}^{n} \alpha_i \right) \tag{3.39}$$

with constraints

$$\alpha_i \geq 0, \tag{3.40}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{3.41}$$

In many applications, most of the α_i which are found by solving a quadratic program turn out to be 0. Those with $\alpha_i \neq 0$ are called *Support Vectors*. On substitution of Equation (3.37) into the decision function (3.30), gives the expression for the decision function in terms of dot products:

$$f_{\mathbf{w},b} = sgn\left(\sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}^T \mathbf{x}_i) + b\right) \tag{3.42}$$

The data only appears in the training problem in the form of dot products, $\mathbf{x}_i \cdot \mathbf{x}_j$. In the nonlinear case, one can first nonlinearly transform input vectors into a high-dimensional feature space by a mapping $\Phi : R^n \rightarrow H$ and then do a linear separation there. The training algorithm only depends on $\Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$. We have

$$f_{\mathbf{w},b} = sgn\left(\sum_{i=1}^{n} \alpha_i y_i (\Phi(\mathbf{x})^T \Phi(\mathbf{x}_i)) + b\right) \tag{3.43}$$

If there is a kernel function, K, such that $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$, it is only necessary to evaluate K in the input space and it is not even necessary to know Φ. The decision function becomes

$$f_{\mathbf{w},b} = sgn\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b\right) \tag{3.44}$$

According to Mercer's condition (Courant and Hilbert 1953): There exists a mapping Φ and an expansion $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$ if and only if, for any $g(\mathbf{x})$ such that

$$\int g(\mathbf{x})^2 d\mathbf{x} < \infty, \tag{3.45}$$

the following inequality holds

$$\int K(\mathbf{x}_i, \mathbf{x}_j) g(\mathbf{x}_i) g(\mathbf{x}_j) d\mathbf{x}_i d\mathbf{x}_j \geq 0. \tag{3.46}$$

Depending on how the inner-product kernel K is generated, it is possible to construct different learning machines characterized by their nonlinear decision surface. Gunn (1998) gives the details of several kernel functions which satisfy Mercer's conditions, such as:

- Polynomial

$$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + 1)^d, d = 1, \ldots \tag{3.47}$$

- Gaussian Radial Basis Function (RBF)

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{(\mathbf{x} - \mathbf{x}')^2}{2\sigma^2}\right) \tag{3.48}$$

In the nonseparable case, slack variables are introduced to tolerate misclassifications,

$$\xi_i \geq 0. \tag{3.49}$$

The constraint of (3.33) is modified to

$$y_i f_{\mathbf{w},b}(\mathbf{x}_i) \geq 1 - \xi_i. \tag{3.50}$$

The regularised risk functional is given by

$$\Psi(\mathbf{w}, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i, \tag{3.51}$$

where C is a given value determining the trade-off between minimizing training errors and minimising the model complexity term $\|\mathbf{w}\|^2$. The whole optimization process is similar to separable case except additional constraints

$$0 \leq \alpha_i \leq C \tag{3.52}$$

are required.

3.4.2 Optimisation

After setting up a model and specifying a loss function (and regulariser), the supervised pattern recognition problem is reduced to an error minimisation problem. Thus choosing suitable optimisation methods is of critical importance to pattern recognition. When gradient information is available, this offers a natural way to perform the minimisation. Many standard iterative optimisation methods are based on gradients, such as gradient descent, conjugate gradients and Newton's method.

Expectation-Maximisation (EM) based optimisation method (Dempster et al. 1977) is a standard technique for maximum likelihood optimisation for incomplete data problem. The EM technique has been used widely to solve supervised learning problems (Ghahramani and Jordan 1994) (Jordan 1998).

The main idea of the EM algorithm is to decouple the problem by estimating the distribution of the hidden parameters given the data and the estimated model parameters in the E-step and in the M-step, estimate the model parameters by maximising this distribution, which is a lower bound problem.

3.4.3 Transductive Classification

The supervised classification method reviewed above belongs to induction-deduction learning category. Unlike the induction-deduction learning, transduction learning does not need an explicit classifier function estimated everywhere (Fig. 3.7). As most classification problems only require one to estimate the outputs of the unknown function for a given test set, the global function estimation (in the induction step) may be overkill (Cherkassky and Mulier 1998). Notice that transductive learning does not combine the unlabelled data in the learning process. Only labelled data are effective in setting up the transductive classifier. Thus transductive learning belongs to supervised learning.

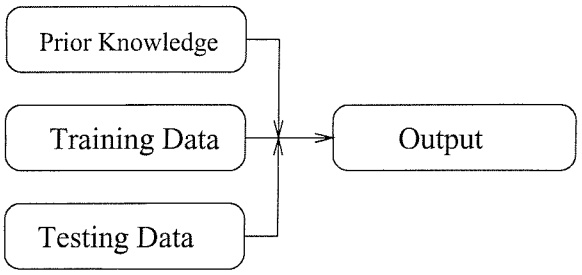


FIGURE 3.7: Transduction

K-nearest-neighbour (KNN) (Bishop 1995) ia a popular non-parametric memory-based model. No assumptions are made about the distribution of the data. Simple nearest neighbour models do not require any training. The algorithm for making predictions involves finding the smallest hypersphere centred around the point X (Fig. 3.8) which contains K points (independent of their class membership), and then assigning it to the class having the largest number of representatives inside the hypersphere. This is done by a majority voting which states it should be assigned the label which occurs the most amongst its K neighbours. KNN performs transduction as the local misclassification error is minimised directly.

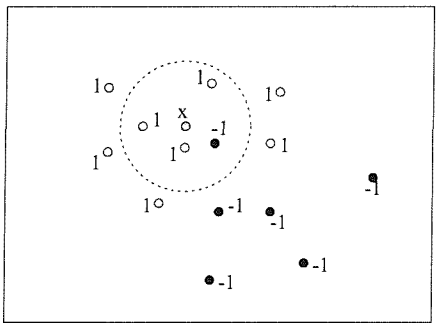


FIGURE 3.8: An example of K-nearest-neighbour classifier

Two important problems need to be considered in defining the classifier. One is how to measure 'closeness'. The simple and common choice is Euclidean distance. The other question is how to choose a suitable value for K. K acts as a smoothing parameter. Too

large a value of K may lead to a relatively poor estimator whereas too small a value of K makes the model very sensitive to the individual data points, which can result in poor generalisation. The value of K is problem dependent and an optimal value for K can be estimated by using cross-validation.

3.5 Semi-supervised Classification

Semi-supervised classification refers to learning a classifier using both labelled data and unlabelled data. It seeks to combine the two well-developed fields: supervised classification and unsupervised classification. The semi-supervised learning diagram is shown in Fig. 3.9. Note that it is a induction-deduction type of learning process. It is different from transductive learning as no unlabelled data are involved in transductive learning process, while labelled data, prior knowledge and unlabelled data are three factors needed for the inductive step of semi-supervised classification.

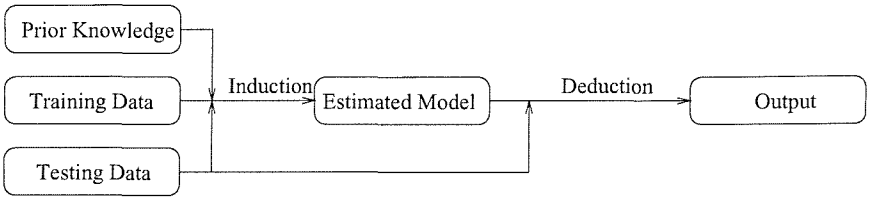
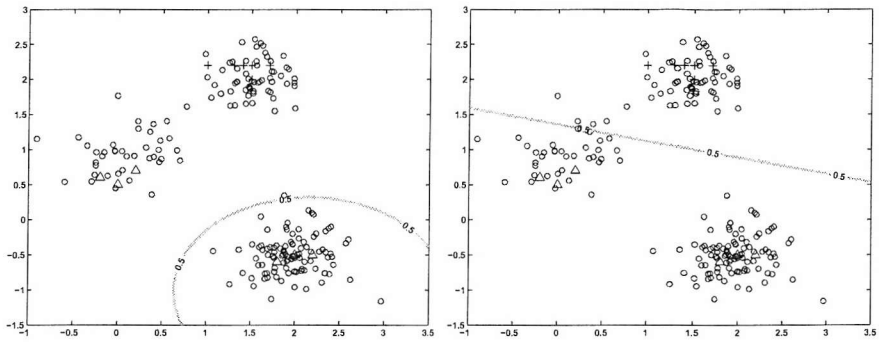


FIGURE 3.9: Semi-supervised induction-deduction learning

3.5.1 A Toy Example

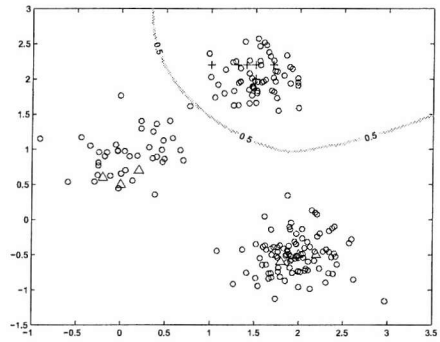
A toy example is used to demonstrate the necessity of using both labelled and unlabelled examples in classification. A set of 2-D data is generated independently from three Gaussians as two classes, where the two clusters close to each other belongs to different classes. Each class has 6 labelled data, as shown in Fig. 3.10. Unsupervised learning with two Gaussian mixtures fails as it tends to classify the data based on the distance (Fig. 3.10(a)); supervised classification using Bayesian MLP network with two hidden-layers gives a relatively large error on test data as a result of limited unrepresentative training data (Fig. 3.10(b)) while a much better classification is shown in Fig. 3.10(c).

This example shows the importance of using all the available information in the data. Although the labelled data contains important class information, it fails to represent the distribution of the whole data due to its small size. A theoretical framework that can exploit the unsupervised data to enhance classification, particularly when obtaining supervised examples is expensive, is a fruitful goal for research.



(a) Classification with no labelled samples

(b) Classification using 6 labelled samples



(c) An ideal classification

FIGURE 3.10: A toy example (+, △: labelled samples in two different classes; ○: unlabelled data)

3.5.2 A Review of the Value of Labelled and Unlabelled Data in Learning

The problem of combining labelled data and unlabelled data in learning has been an active research area in recent years, since traditional supervised techniques provide no means to incorporate extra knowledge contained in unlabelled data. Unsupervised classification techniques only uses unlabelled data and group them into clusters and provides no class information. Both supervised and unsupervised classification techniques learn information from data with the same form, i.e. either with class label or without class label.

Castelli and Cover (1996) shows that labelled samples are necessary to construct a classifier and are more valuable than unlabelled samples. However, supervised neural networks have not played an important role in image segmentation, since labelled data are often difficult to obtain or is unable to represent all possible variations of the operational environment to which the classifier will be applied (e.g. medical and remote sensing images). When labelled data are very difficult to obtain and input dimensionality is high, the unlabelled data can enhance learning. However, there is no framework

to address the optimisation problem in combining both labelled examples and image models.

Amongst early work, Gutfinger and Sklansky (1991) discuss mixed adaption to make the classifier robust. Shahshahani and Landgrebe (1994) analyse the value of unlabelled data in classification from the point of Fisher information matrix. They propose a method of using unlabelled data in classification using mixture models where each cluster is manually assigned to a class in the learning, which is termed "hard-partition". A cluster-class "soft-partition" method is proposed by Miller and Uyar (1996). Miller also analyses the connection of his method to the radial basis function networks. Nigam et al. (2000) use a method similar to Shahshahani and Landgrebe (1994) in text classification using both labelled and unlabelled documents. Jaakkola and Haussler (1998) consider a two-step learning for supervised techniques such as support vector machines: they build a generative model from unlabelled data and then exploit this generative model in supervised learning. Joachims (1999) first proposes a sparse and regularised method for inference for text classification using support vector machines using both labelled and unlabelled examples. Blum and Mitchell (1998) propose a learning paradigm called *co-training* to address the problem where strong structural prior knowledge is available. In (Nigam and Ghani 2000), Nigam analyses the effectiveness and applicability of co-training. Cohn et al. (1996) discuss active learning with statistical models. These results indicate that it is advantageous to incorporate unlabelled data with labelled data in the learning process.

More recently, Jebara and Pentland (1998) propose a maximum conditional likelihood method as an extension of the EM algorithm to conditional density estimation under missing data, where a bounding and maximisation process is given to specifically optimise conditional likelihood instead of the usual joint likelihood. Schuurmans and Southey (2001) discuss metric-based methods for adaptive model selection and regularisation that exploits unlabelled data to adaptively control hypothesis complexity in supervised learning tasks. The idea is to impose a metric structure on hypothesis by determining the discrepancy between their predictions across the distribution of unlabelled data. Blum and Chawla (2001) discuss learning from labelled data and unlabelled data using graph minicuts which is seen to be robust to noise on the labelled data. This method uses a similarity measurement between data to construct a graph and then outputs a classification corresponding to partitioning the graph in a way that minimises the number of similar pairs of examples that are given different labels. Ivanov et al. (2001) discuss EM method for weakly labelled data.

As a problem involving both labelled and unlabelled data, it is believed that semi-supervised learning is capable of learning more information from the data and thus achieve improved performance. However, as both labelled data and unlabelled data are involved, the model assumption and model optimisation become very difficult. Recently, kernel machines such as support vector machines are an very active research

field by mapping input spaces into kernel spaces. The capacity and flexibility provided by kernel machines might enable them to be used in the complex modelling problem with labelled and unlabelled data. In terms of optimisation, the Expectation Maximisation (EM) algorithm can be used by treating the unlabelled data's labels as the missing data. Thus kernel machines and EM algorithm may provide useful insights and powerful methodologies in the direction of semi-supervised learning research.

The following section will give an probabilistic method for semi-supervised classification. The detailed model assumption and parameter optimisation will be discussed.

3.5.3 A Probabilistic Method

In the following, the data to be classified will be denoted $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ with mixture (cluster) labels $\mathbf{z} = \{z_1, z_2, \cdots, z_n\}$, $z_i \in \{1, 2, \cdots, K\}$ and class labels $\mathbf{y} = \{y_1, y_2, \cdots, y_n\}$, $y_i \in \{1, 2, \cdots, J\}$ to be decided, as shown in Fig. 5.4. $\mathbf{x}_i$ is an m-dimensional feature vector for pixel i. Note that the class labels c are distinguished from mixture labels z, as z are not necessarily the meaningful physical class labels c. In addition, $\mathcal{D} = \{(\mathbf{x}_1^l, y_1^l), (\mathbf{x}_2^l, y_2^l), \cdots, (\mathbf{x}_{n^l}^l, y_{n^l}^l)\}$ are the labelled samples and their associated class labels, where n^l is the number of labelled samples.

A probabilistic method for using both labelled and unlabelled data are proposed in (Miller and Uyar 1996). Given the model, the joint data log-likelihood is written in the form

$$\begin{aligned} \log \Xi &= \log p(\mathbf{x}) + \log p(\mathbf{x}^l, y^l) \\ &= \log p(\mathbf{x}) + \sum_{i=1}^{n^l} \log p(\mathbf{x}_i^l, y_i^l). \end{aligned} \tag{3.53}$$

This objective function consists of a "supervised" term $\sum_{i=1}^{n^l} \log p(\mathbf{x}_i^l, y_i^l)$ based on the labelled data $\mathbf{x}_l$ and an "unsupervised" term $\log p(\mathbf{x})$ based on unlabelled data $\mathbf{x}$. The joint likelihood allows the inclusion of unlabelled data in the learning. The supervised term can be expressed as

$$p(\mathbf{x}_i^l, y_i^l) = \sum_{k=1}^{K} p(y_i^l | \mathbf{x}_i^l, z_i^l = k) p(\mathbf{x}_i^l | z_i^l = k) p(z_i^l = k). \tag{3.54}$$

The labelled examples make possible the establishment of a probabilistic distribution $p(y_i = j | z_i = k)$ to describe the connections between the "mixture label" z_i and the "class label" y_i.

A model is needed to classify each mixture to each class. A hard "partitioned" mixture model is used in (Shahshahani and Landgrebe 1994), where each mixture component is

hard-partitioned to classes, that is the mixture component k belonging to which class j is predetermined. A probabilistic "generalised mixture" (GM) model is introduced in (Miller and Uyar 1996) and shows improved performance:

$$\gamma_{j|k} \equiv p(y_i = j|z_i = k) = \frac{\sum\limits_{\mathbf{x}_i^l, c_i^l = j} p(z_i^l = k|\mathbf{x}_i^l, c_i^l)}{\sum\limits_{\mathbf{x}_i^l} p(z_i^l = k|\mathbf{x}_i^l, c_i^l)}, \tag{3.55}$$

where $\gamma_{j|k}$ is used as $p(y_i = j|z_i = k)$ is independent of the pixel i. Finally, the class membership of each data is decided by

$$p(y_i = j|\mathbf{x}_i) = \sum_{k=1}^{K} \gamma_{j|k} p(z_i = k|\mathbf{x}_i). \tag{3.56}$$

An algorithm is needed to solve the integrated optimisation in Equation (3.53). This optimisation is very difficult as it involves both labelled and unlabelled data. It is noted that the EM algorithm is an efficient method to solve an optimisation problem where there is hidden data. For supervised learning, a gradient-based method can be used for optimisation. In recent years some EM-based optimisation algorithms have been proposed for parameter estimation in unsupervised learning. Jordan and Jacobs (1994) use an EM algorithm to maximise conditional likelihood in a mixture of experts framework. The mathematical connections between the EM algorithm and the gradient-based approaches for maximum likelihood learning of finite Gaussian mixtures were developed in (Xu and Jordan 1996). These make the EM algorithm a natural method for solving optimisation problems involving both unlabelled and labelled data.

Starting with an initial estimate of model parameters, $\widehat{\Phi}^{(0)}$, the algorithm iterates:

- E-step: Estimate $Q(\Phi|\widehat{\Phi}^{(t)}) = E[\log p(\mathbf{x}, z) + \log p(\mathbf{x}^l, c^l)|\mathbf{x}, \widehat{\Phi}^{(t)}]$
- M-step: Find $\widehat{\Phi}^{(t+1)} = \arg\max_{\Phi} Q(\Phi|\widehat{\Phi}^{(t)})$.

Here t represents the tth iteration. As the unlabelled data are independent of each other, the mixture label's prior is updated by,

$$p(z_i^{(t+1)} = k) = \frac{\sum\limits_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i) + \sum\limits_{i=1}^{n^l} p(z_i^{l(t)} = k|\mathbf{x}_i^l, c_i^l)}{n + n^l}. \tag{3.57}$$

When using a Gaussian distribution to model the conditional distribution $p(\mathbf{x}_i|z_i = k, \Phi)$,

$$p(\mathbf{x}_i|z_i = k, \Phi) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right\}}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}}, \tag{3.58}$$

the mean vector and the covariance matrix of component k, $\phi_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, are re-estimated using:

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)\mathbf{x}_i + \sum_{i=1}^{n^l} p(z_i^{(t)} = k|\mathbf{x}_i^l, c_i^l)\mathbf{x}_i^l}{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i) + \sum_{i=1}^{n^l} p(z_i^{(t)} = k|\mathbf{x}_i^l, c_i^l)}, \tag{3.59}$$

$$\left(\boldsymbol{\Sigma}_k^2\right)^{(t+1)} = \frac{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)S_{ik} + \sum_{i=1}^{n^l} p(z_i^{(t)} = k|\mathbf{x}_i^l, c_i^l)S_{ik}}{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i) + \sum_{i=1}^{n^l} p(z_i^{(t)} = k|\mathbf{x}_i^l, c_i^l)}, \tag{3.60}$$

where $S_{ik} \equiv (\mathbf{x}_i - \boldsymbol{\mu}_k^{(t)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t)})^T$,

$$p(z_i^{(t)} = k|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|z_i^{(t)} = k)P(z_i^{(t)} = k)}{\sum_{k=1}^{K} p(\mathbf{x}_i|z_i^{(t)} = k)P(z_i^{(t)} = k)}, \tag{3.61}$$

$$p(z_i^{(t)} = k|\mathbf{x}_i^l, c_i^l) = \frac{p(\mathbf{x}_i|z_i^{(t)} = k)\gamma_{j|k}^{(t)}P(z_i^{(t)} = k)}{\sum_{k=1}^{K} p(\mathbf{x}_i|z_i^{(t)} = k)\gamma_{j|k}^{(t)}P(z_i^{(t)} = k)} \tag{3.62}$$

with $\gamma_{j|k}^{(t+1)}$ updated by Equation (3.55).

This EM re-estimation continues until the updates fall below a tolerance or the maximum number of iteration reaches. Finally, the class label is determined using Equation (3.56).

3.6 Summary

This chapter has reviewed classical approaches to pattern recognition, with the emphasis on learning and generalisation. There may be several factors that influence the performance of the pattern recognition classifier. First the correctness and efficiency of the labelling process to generate data for training the classifier. This part may need a large amount of expert knowledge. The correct labelling process is often critical to the following training process. Secondly what kind of environment the classifier will be used. Most classifier performs better on the data that is similar to the labelled data used for setting up the classifier while fails on the data that is different from the labelled data. Thirdly the way the classifier will be trained. This is a very important step and the field that quite a lot of research have been done. Many classifiers have been set up using various model assumptions and optimisation algorithms. The problem of setup of a classifier which has good generalisation with respect to the testing data is still an open question. The semi-supervised techniques are a good way to treat this problem.

The next chapter will use the three different types of classification techniques introduced in this chapter in simulated and real PET reference region extraction.

Chapter 4

Temporal PET Reference Region Extraction

The purpose of the work described in this chapter is to investigate whether the characterisation of dynamic PET images will benefit from the application of expert knowledge contained in the labelled data, using supervised and semi-supervised classification.

Pattern recognition techniques are applied to solve the PET reference region localisation and modelling problem for [^{11}C](R)-PK11195 PET images. Results on simulated PET data are presented, followed by the results on real PET data. For both types of data, results of reference region exaction via unsupervised, supervised and semi-supervised pattern recognition are compared.

4.1 Parametric and Non-parametric Modelling

There are two general approaches to the analysis of PET images: parametric approaches and non-parametric approaches. Parametric methods construct an explicit TAC model, using a set of biochemically or physiologically meaningful parameters. Often the parameters are chosen such that they represent the desired characteristic, and so estimation of the parameters is the main task of interest. After the modelling, a "statistical image" can be displayed to show the value of one feature (non-parametric modelling) or one parameter such as binding potential in PET (parametric modelling) associated with each voxel (Gunn et al. 2001).

The parametric approach (model-based approach) has traditionally been applied to the design of signal processing algorithms. A mathematical model is derived that describes the physical signal-generating system, such as compartmental models in PET. This model is then used to derive a mathematical procedure that should constitute an optimal solution to the processing problem faced. An optimisation procedure is then devised

which minimises some error between the model and the data to recover the optimal model parameters. However, the approach is often hampered by the lack of knowledge about the signal generating system. Thus simplified assumptions are often made about the system, producing suboptimal solutions.

Non-parametric approaches apply some manner of filtering or transformation to the TAC such that features which represent a desired characteristic can be obtained (principal component analysis; factor analysis; clustering technique etc.). Data-driven approaches are an alternative to the model-based approach, which avoid the definition of a mathematical model based on the prior assumptions about the system. Examples of the signals are provided as the inputs to the system. Then, by learning from these examples, the characteristics of the real system (non-linear, non-stationary, non-Gaussian etc.) are taken into the learning system implicitly. No explicit model is necessary. However, the parameters in the generated learning system are not directly related to meaningful parameters.

PET modelling in this thesis first uses non-parametric methods to extract the reference region, then uses a parametric model, a simplified reference region model, to generate statistical images of parameters of interest such as binding potential. The next three sections describe three non-parametric methods - unsupervised classification, supervised classification and semi-supervised classification - in PET modelling.

4.2 Unsupervised Reference Region Extraction

Most classification methods used in PET are unsupervised. A basic method is Principal Component Analysis (PCA) which has been used in PET for dimensionality reduction (Yap et al. 1996). However, the orthogonal factors produced by PCA are not necessarily related to physiologically meaningful time-activity distributions. For PET reference region segmentation, K-means and Gaussian Mixture Model (GMM) are two efficient techniques that have been used (Ashburner et al. 1996). No *apriori* knowledge is involved in the segmentation process, and the number of underlying patterns and the final discrimination are determined manually.

In (Ashburner et al. 1996), cluster analysis is used for the characterisation of dynamic $[^{11}C]$flumazenil PET data. The data are partitioned according to its probability of belonging to each of k clusters, as described in the Gaussian mixture unsupervised modelling techniques introduced in Section 3.3.2 (Hartigan 1975). The superiority of Gaussian mixture models for clustering over other methods like K-means is that it provides a powerful probabilistic representation. In the experiments in this chapter, the unsupervised classification method used is the Gaussian mixture models. It is optimised using EM algorithm described in Section 3.3.2.2.

Fig. 4.1 shows the procedure in the unsupervised classification for PET image modelling to generate binding potential images. After setting up an unsupervised clustering algorithm using Gaussian mixture models, an expert manually selects one or two clusters as the data from the reference region with the other clusters representing the non-reference region. The mean TAC from all the manually chosen reference clusters is calculated. This is fed into the reference region model, allowing the generation of parametric maps of binding potential and relative delivery of each voxel.

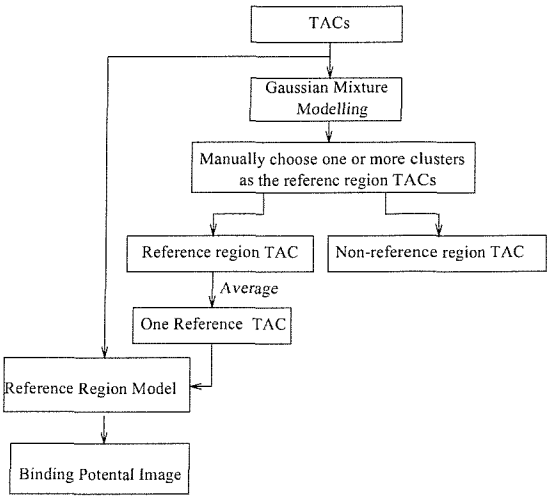


FIGURE 4.1: The unsupervised Gaussian mixture models and the reference region model to generate the binding potential image

4.3 Supervised Reference Region Extraction

No supervised classification techniques have been used before in PET modelling. In this section a supervised classification approach is proposed to extract of the reference region from PET dynamic images. When enough representative training data are available, supervised classification is more suitable for extracting the reference region from PET dynamic images than unsupervised classification. In unsupervised learning, the learning must proceed on the distribution of the patterns in the input space alone. The goal of unsupervised classification is to model the probability distribution of the input data and to use this to segment the input space into regions. As no classifier is produced, an expert must decide how many patterns there are, as well as how to map from the clusters to target patterns after unsupervised learning. Each time a new data set appears, the whole learning process must repeat, and as such no information is aggregated over different scans. In supervised classification, the desired output is known for each input pattern and the system can learn from these examples to generalise to "unseen" values for a new input pattern. The basic notion of supervised learning is to construct a model from examples. For a system to use supervised learning, a teacher must help the system in its model construction by providing positive and negative examples of objects belonging

to these classes. An inductive process is used to build up a model from the examples, producing a system which can determine the class membership of an input pattern. The merit of supervised learning is that it sets up a classifier that can remember the expert knowledge and can be reused.

As Cherkassky and Mulier (1998) observe, no single universally accepted theoretical framework for predictive learning currently exists. Neural networks have received an immense research interest and have been applied in many areas. The Bayesian approach to neural networks can be considered to be an example of the parameter space approach to learning, viz. the learning machine is parameterised by a weight vector and the task of learning is to find optimal values of these weights. Kernel methods such as support vector machine can find the unique and optimal solution. In this application, there are a large number of data sets and also a relative large amount of noise. If a support vector machine classifier were used, a large proportion of training data would be kept as support vectors. As a result, the classification process will be relatively slow. Additionally, the inaccurate data kept as support vectors will affect the accuracy of the classification. Thus, instead of support vector machine or other kernel methods, a Bayesian multi-layer perceptron classifier is used.

4.3.1 Bayesian Network Structure

The Multi-Layer Perceptron (MLP) network is one of the most widely used supervised neural networks, but it can suffer from "overfitting", especially when the data are noisy. A Bayesian framework (Mackay 1992b) avoids "overfitting" by incorporating capacity control to the model.

In this section, a multi-layer perceptron (MLP) network with a logistic activation function is trained with a regularised cost function according to the Bayesian framework. The network is trained using the cost function

$$R_{reg}(\omega) = R_{emp}(\omega) + \frac{1}{2}\sum_{c} \alpha_c \|\omega\|_c^2, \tag{4.1}$$

where α_c is the regularizer for the subset c of weights W_c, $\|\omega\|_c^2 = \sum_{w \in W_c} w^2$ is the regulariser for the weights W_c, ω is the weight vector and $R_{emp}(\omega)$ is the sum of cross-entropy loss term

$$R_{emp}(\omega) = -\sum_{i=1}^{n} \{y_i \ln f(x_i, \omega) + (1 - y_i) \ln(1 - f(x_i, \omega))\}. \tag{4.2}$$

The initial values for α_c are set to small arbitrary values and the network is then trained using gradient descent to minimise $R_{reg}(\omega)$. The regularisers α_c are then re-estimated

using the evidence framework

$$\alpha_c = \frac{\gamma_c}{\|\omega\|_c^2}, \tag{4.3}$$

where

$$\gamma_c = \sum_{j=1}^{J} \frac{\lambda_j}{\lambda_j + \alpha_c} (\mathbf{V}^T I_c \mathbf{V})_{jj}, \tag{4.4}$$

J is the total number of network weights, λ_j and the jth column of $\mathbf{V}$ are the jth eigenvalue and eigenvector of the Hessian matrix $\nabla^2 G$.

Once the regularisation parameters have been re-estimated, further minimisation is performed. This re-estimation and further training continues until the updates fall below a tolerance. This re-estimation scheme enables the network to adjust its regulariser controlling the capacity of the network and limiting over-fitting.

4.3.2 Decision Making

Let C_1 and C_2 denote reference region TAC class and non-reference region TAC class respectively. The logistic activation function allows the network output y_i to be interpreted as the probability $p(C_1|\mathbf{x}_i)$. In this two-class problem, $p(C_1|\mathbf{x}_i) + p(C_2|\mathbf{x}_i) = 1$. The discrimination rule is:

$$\begin{aligned} \mathbf{x}_i &\sim C_1 \quad \text{if} \quad p(C_1|\mathbf{x}_i) \geq 0.5; \\ \mathbf{x}_i &\sim C_2 \quad \text{if} \quad p(C_1|\mathbf{x}_i) < 0.5. \end{aligned}$$

In a PET scan, once all the voxels belonging to the reference region are extracted, the mean reference region TAC can be obtained.

4.3.3 Supervised Neural Network Reference Region Extraction

After setting up a supervised neural network, each voxel in the PET image is segmented as the reference region and the non-reference region. The mean TAC from the extracted reference region is calculated. This is fed into the reference region model, allowing the generation of parametric maps of binding potential and relative delivery of each voxel. Fig. 4.2 shows the procedure used in the supervised reference region extraction and binding potential image generating.

4.4 Semi-supervised Reference Region Extraction

Semi-supervised classification enables learning information from both labelled and unlabelled data. This section gives the detailed procedure of extracting reference region using the semi-supervised learning method and generating binding potential images.

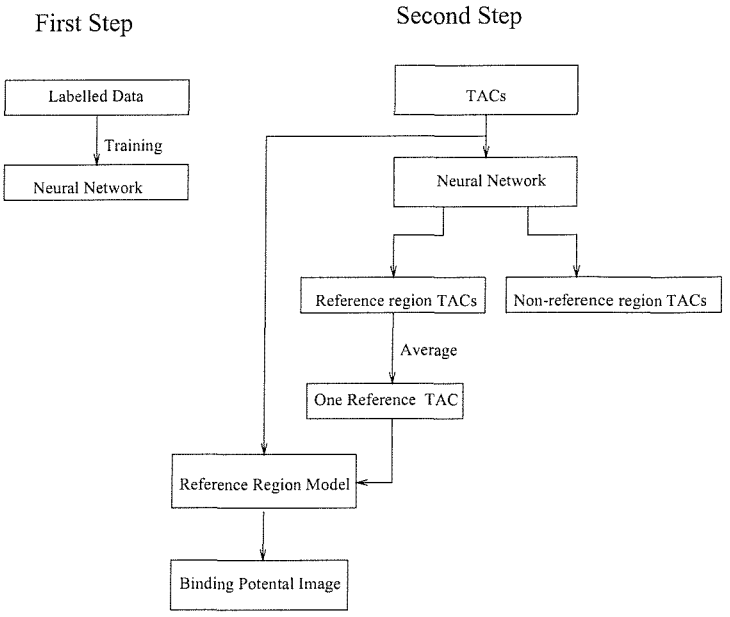


FIGURE 4.2: The supervised neural network and the reference region model to generate the binding potential image

The labelled reference and non-reference TAC data and all the TACs in the PET images are used to train the probabilistic semi-supervised classifier.

Similarly, let C_1 and C_2 denote reference region TAC class and non-reference region TAC class respectively. The semi-supervised classification output y_i can be interpreted as the probability $p(C_1|\mathbf{x}_i)$. Each voxel in the PET image can be segmented as the reference region and the non-reference region according to the same decision rule as described in the supervised decision making (Section 4.3.2).

In a PET scan, once all the voxels belonging to the reference region are extracted, the mean reference region TAC can be obtained. This is fed into the reference region model, allowing generating parametric maps of binding potential and relative delivery of each voxel.

Fig. 4.3 shows the procedure of using semi-supervised classification in PET reference region extraction and parametric image generation.

4.5 Simulated PET Experiments

In this section, synthetic positron emission tomography (PET) images are used to examine the performance of the algorithms.

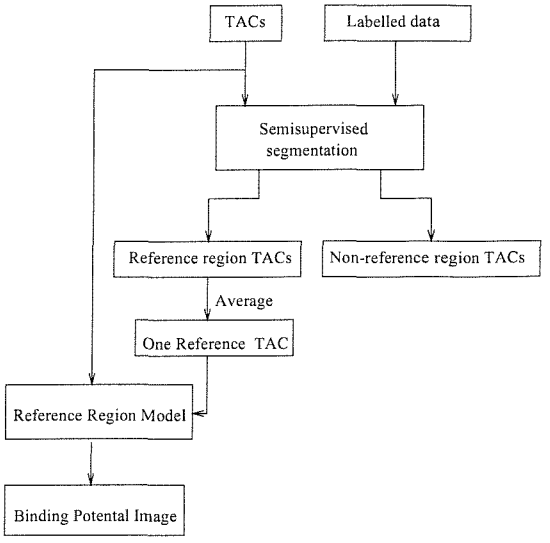


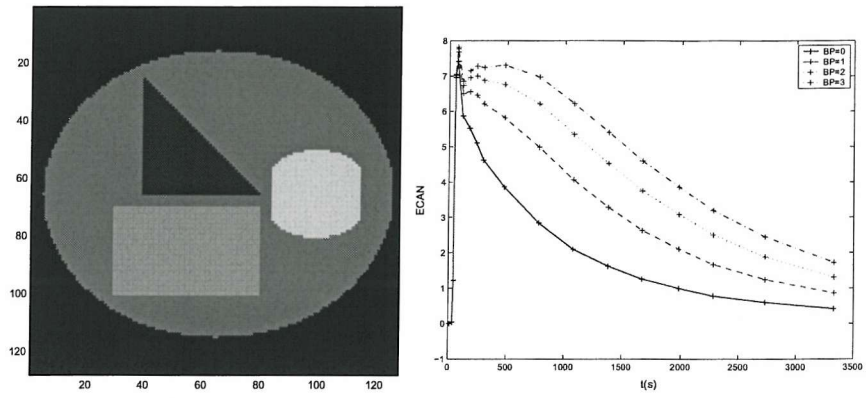
FIGURE 4.3: The probabilistic semi-supervised method and the reference region model to generate the binding potential image

4.5.1 Data Description

Simulated data are generated as follows: an 18-D vector from the cortex in a real PET scan is extracted as a reference region time-activity curve with binding potential value $BP = 0$. Three other vectors are generated by using this reference region TAC and the simplified reference region model (Mintun et al. 1984) with $BP = 1, 2, 3$ respectively. Thus four vectors with different binding potential are generated. These vectors are assigned to each of four regions in the simulated image as shown in Fig. 4.4, with added noise for TACs in voxels. The reference region, with $BP = 0$ is the oval background. Additionally, labelled examples are obtained by generating 18-D vectors with $BP = 0$ in one class and $BP = 1, 2, 3$ as the other class. In all experiments, Gaussian noise with zero mean and a certain standard deviation is added to the TACs. The standard deviation varies from 0.3 to 2.6 with the increment of 0.1 in the trials. In each trial, the noise level in the labelled samples and the image to be segmented are the same. The standard deviation is directly used to represent the noise level instead of the traditional Signal-to-Noise Ratio (SNR) representation, as the underlying signal is constant in the trials with different noise level.

Three different methods are compared to extract the region with $BP = 0$ from the regions with $BP = 1, 2, 3$:

- Unsupervised classification using Gaussian mixture model described in Section 4.2;
- Supervised segmentation using Bayesian neural networks as described in Section 4.3;
- Semi-supervised method described in Section 4.4.



(a) Four regions in the image

(b) Four time-activity curves associated with four regions

FIGURE 4.4: The simulation data: Each voxel in the image has an 18-D feature vectors

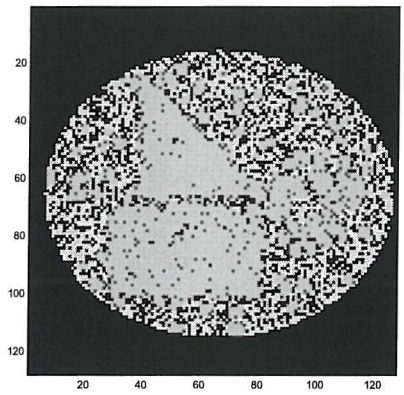
The detailed description of the procedures are given in the previous three sections and illustrated in Fig. 4.1, Fig. 4.2 and Fig. 4.3. In supervised and semi-supervised algorithms, the final classification is carried out automatically by assigning every voxel to the class where it has the highest posterior probability. In the unsupervised segmentation, the number of clusters is manually specified. In this simulated experiment, the cluster assignment can be done automatically by choosing the cluster whose centre has lowest BP as the reference class, while the rest are treated as the other class.

4.5.2 Results

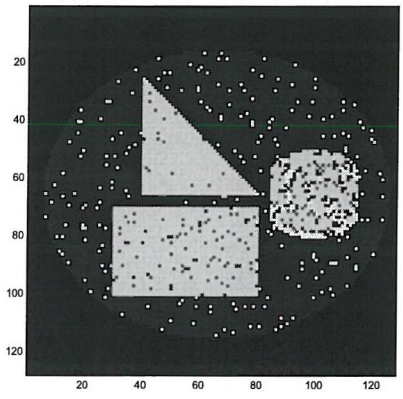
The image segmentation results using these three methods, with 500 labelled data and noise standard deviation $\sigma = 0.8$, are displayed in Fig. 4.5. The unsupervised segmentation result (Fig. 4.5(a)) fails to localise the reference region, when 4 clusters were used. The supervised segmentation result in Fig. 4.5(b) gives an improved result, while the semi-supervised segmentation result in Fig. 4.5(c) gives the best classification accuracy.

A 3×3 median filter is applied to the segmentation result in Fig. 4.5, as shown in Fig. 4.6. It can be seen that after the median filtering, a large amount of scattered points inside the images are removed. However, the scattered points remain around the image edge. The semi-supervised method gives the best accuracy. In the supervised segmentation image, the edge of the right side object is still not ideal, compared to the original image.

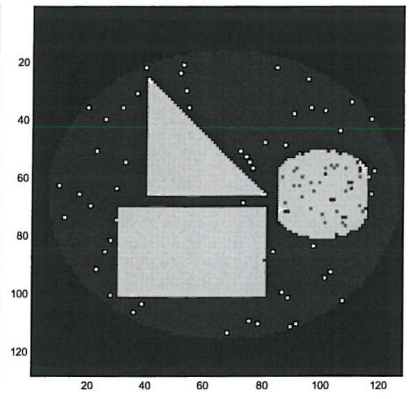
To further compare the algorithms' performance, the influence of the number of labelled examples and the noise level in the data is examined. Although the unsupervised segmentation results are not affected with the change in the number of labelled examples, for the sake of comparison, they are displayed in the same way as the other two meth-



(a) Unsupervised segmentation



(b) Supervised segmentation

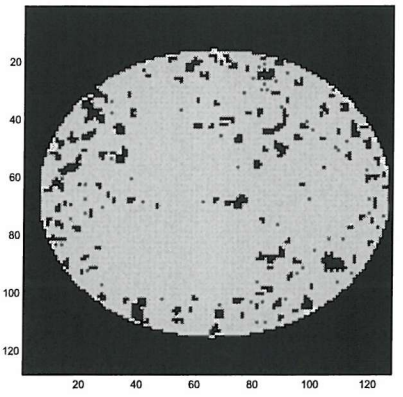


(c) Semi-supervised segmentation

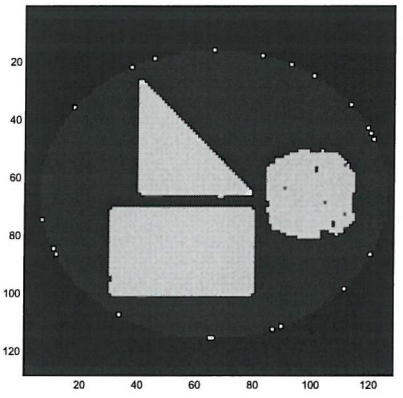
FIGURE 4.5: Segmentation results with 500 labelled examples, noise standard deviation $\sigma = 0.8$

ods. Fig. 4.7, 4.8, 4.9 show the mean and standard deviation for the voxel classification accuracy, the non-reference region classification error and the BP value of the extracted TAC respectively. The four rows in each figure correspond to the labelled examples's number 100, 200, 500 and 1000 respectively. The error bars are generated by running every method for ten times with different initial values for model parameters. The number of clusters is set to 4 in both the unsupervised and semi-supervised classification. The three types of error are defined as follows:

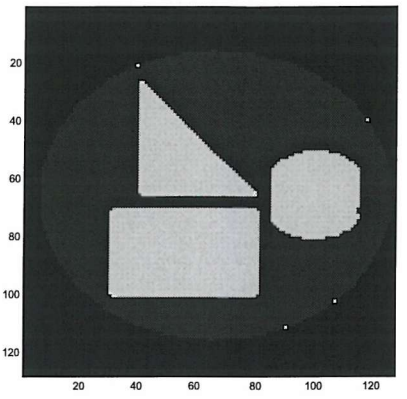
- The total classification error is calculated by $\frac{Number\ of\ misclassified\ voxels}{Total\ number\ of\ voxels} * 100\%$;
- The non-reference region classification error is $\frac{Number\ of\ misclassified\ non-reference\ region\ voxels}{Total\ number\ of\ non-reference\ region\ voxels} * 100\%$;
- The BP value of the extracted reference TAC is calculated using the simplified reference region model (Equation (2.3)) and the basis function method (Equation (2.6),(2.7),(2.8)).



(a) Unsupervised segmentation



(b) Supervised segmentation



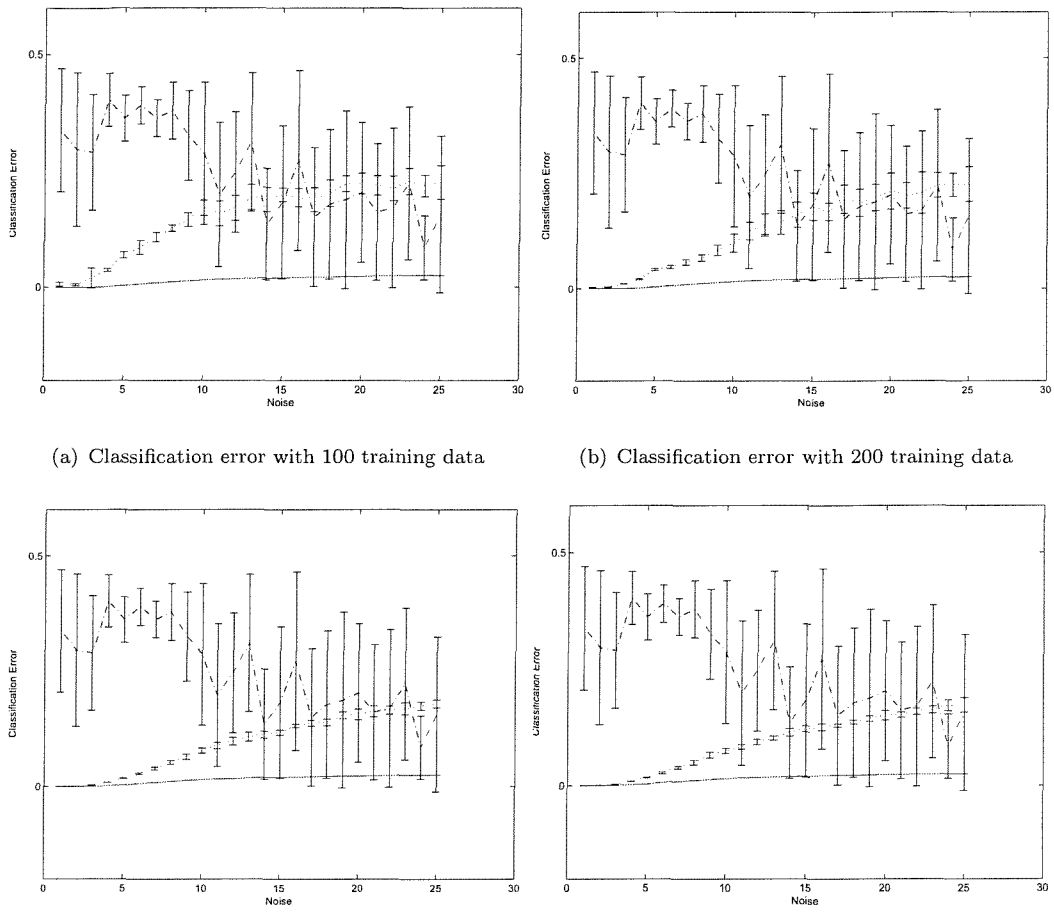
(c) Semi-supervised segmentation

FIGURE 4.6: Results after median filtering

Each figure shows the change of one of these three types of error with a different noise level. The noise level runs from 0.2 to 2.6.

Fig. 4.10, 4.11, 4.12 show the results with the number of clusters in the unsupervised and semi-supervised classification changed to 10.

Fig.4.7 and Fig.4.10 shows the total classification error with 100, 200, 500, 1000 training data. In both figures, the semi-supervised segmentation achieves best classification performance while the unsupervised segmentation result gives the worst classification error. As each optimisation process involved in this simulation can only find a local minimum, the classification results vary with different initial values for model parameters. With smallest standard deviation, the semi-supervised method shows its robustness with different starting points. Unsupervised classification performance is also unstable as it gives highest standard deviation. As the noise level increases, the classification error for supervised and semi-supervised classification increases, as expected. However, the unsupervised classification error decreases in the four-cluster case. One possible explanation is that when the noise level is low, the algorithm tends to split the true reference



(a) Classification error with 100 training data

(b) Classification error with 200 training data

(c) Classification error with 500 training data

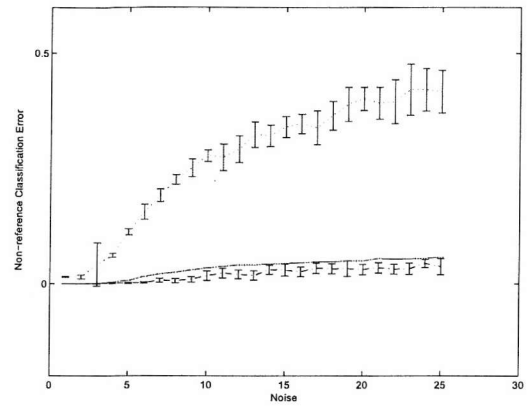
(d) Classification error with 1000 training data

FIGURE 4.7: Simulation results on classification accuracy, with 4 clusters in unsupervised classification (Red solid : Error bar for semi-supervised classification; Blue dotted: Error bar for supervised classification; Black dashdot: Error bar for unsupervised classification.)

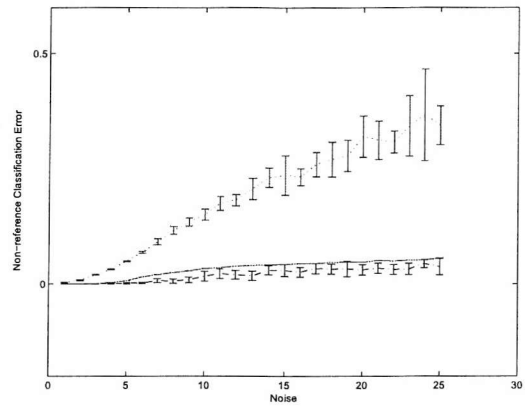
region data into more than one clusters, as only one cluster will be picked, the total classification error is large. When the noise level increases, the chosen cluster contains more true reference region data. This downtrend is less severe when the cluster number increases to 10, as shown in Fig.4.10.

As our main interest is the accuracy of reference regions, the percentage of non-reference regions classified as reference regions are emphasized. Fig.4.8 and Fig.4.11 show the non-reference region mis-classification error. The unsupervised classification achieves best accuracy. The result for the 10-cluster case achieves more accuracy than the 4-cluster case. However, semi-supervised classification has the lowest standard deviation, showing that its performance is very stable.

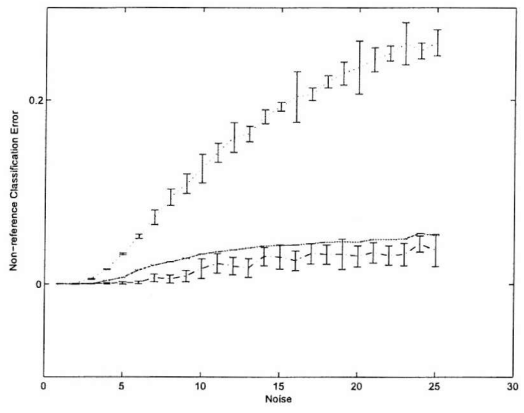
Fig.4.9 and Fig.4.12 show the binding potential error of the classified reference region voxels. The binding potential is the meaning parameter that will be extracted to de-



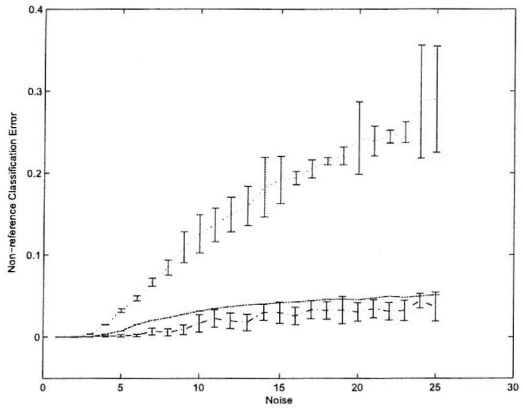
(a) Non-reference region classification error with 100 training data



(b) Non-reference region classification error with 200 training data



(c) Non-reference region classification error with 500 training data

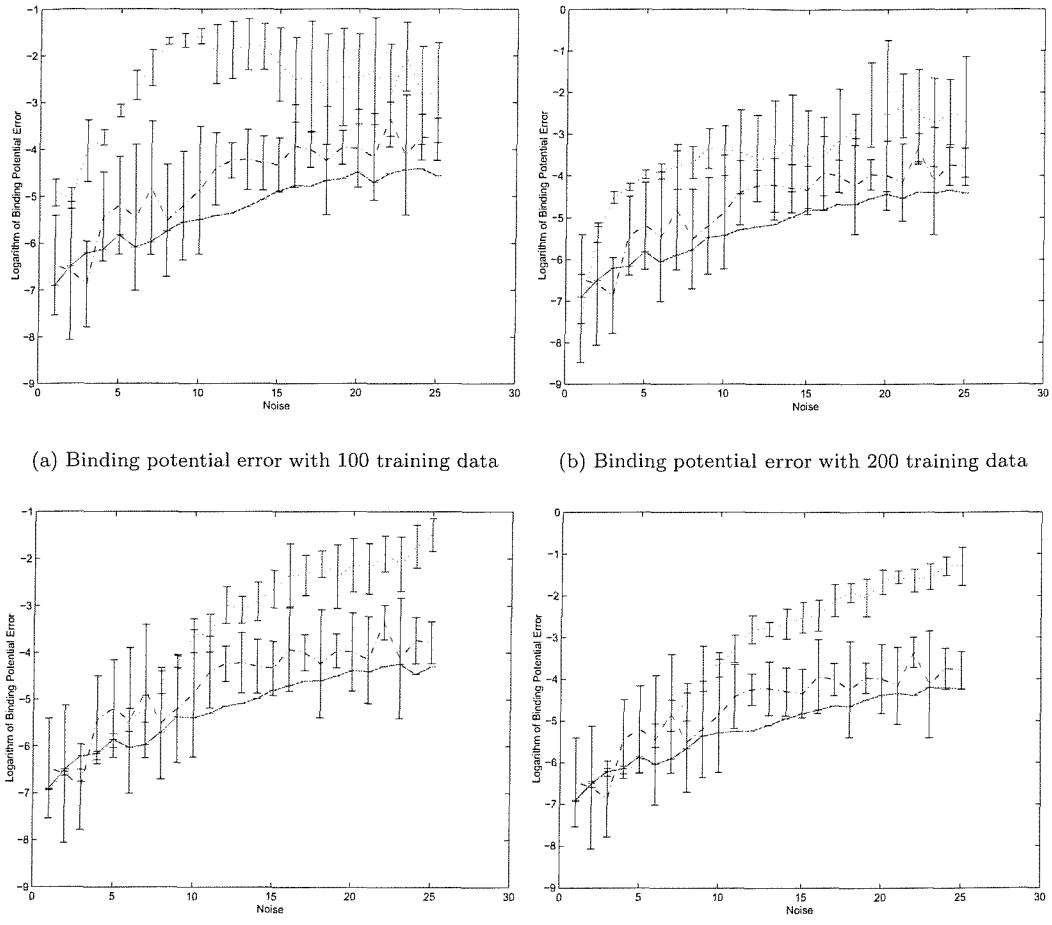


(d) Non-reference region classification error with 1000 training data

FIGURE 4.8: Simulation results on non-reference region classification accuracy, with 4 clusters in unsupervised classification (Red solid : Error bar for semi-supervised classification; Blue dotted: Error bar for supervised classification; Black dashdot: Error bar for unsupervised classification.)

scribe the characteristic of PET images. Thus minimising the binding potential error is very important in PET segmentation. The semi-supervised method achieves the best performance with excellent stability reflected by the low standard deviation in Fig.4.7 and Fig.4.10. The unsupervised method also achieves competitive performance but it is less stable.

Overall, the semi-supervised method obtains the best accuracy in this simulation. The semi-supervised performance is also very stable in that the error bar for the semi-supervised classification in all the above figures are small compared to the other two methods. One possible reason is that the simulated data is generated from a Gaussian distribution with additional noise. As there are Gaussian mixtures in the semi-supervised model (Section 3.5.3), it is easier to find similar optimal solutions for the



(a) Binding potential error with 100 training data

(b) Binding potential error with 200 training data

(c) Binding potential error with 500 training data

(d) Binding potential error with 1000 training data

FIGURE 4.9: Simulation results on binding potential accuracy, with 4 clusters in unsupervised classification (Red solid : Error bar for semi-supervised classification; Blue dotted: Error bar for supervised classification; Black dashdot: Error bar for unsupervised classification.)

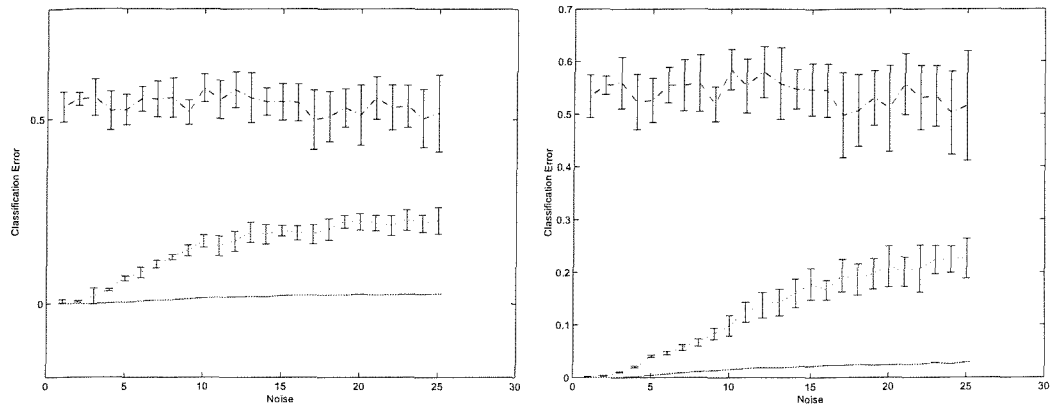
model parameters from different starting points in the optimisation process. Although Gaussian mixtures are also used in the unsupervised model, the unsupervised method often fails to find the right class membership for Gaussian components, making the final result less stable.

4.5.3 Discussions

4.5.3.1 Different Ways to Estimate Unsupervised Classification Error

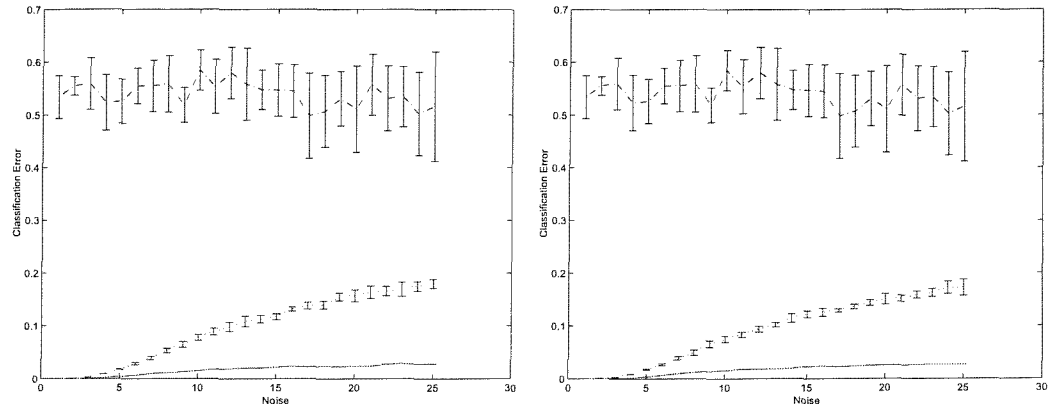
In the above section, the total classification error and the non-reference region classification error (i.e. false positive) is used. Another aspect of classification error is

- the reference region classification error (i.e. false negative), defined as



(a) Classification error with 100 training data

(b) Classification error with 200 training data



(c) Classification error with 500 training data

(d) Classification error with 1000 training data

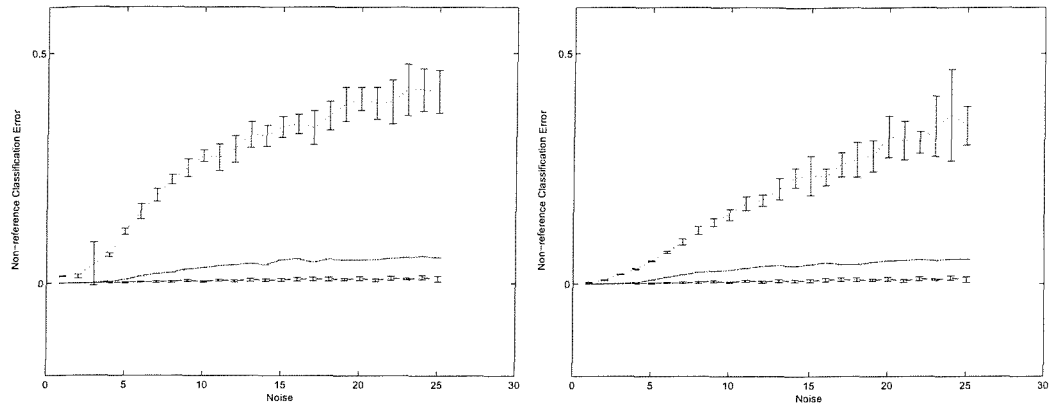
FIGURE 4.10: Simulation results on classification accuracy, with 10 clusters in unsupervised classification (Red solid : Error bar for semi-supervised classification; Blue dotted: Error bar for supervised classification; Black dashdot: Error bar for unsupervised classification.)

$$\frac{Number\ of\ misclassified\ reference\ region\ voxels}{Total\ number\ of\ reference\ region\ voxels} * 100\%.$$

These three different aspects of the same unsupervised classification result is shown in Fig. 4.13. The non-reference region classification error is very small. As we only choose one cluster that has the smallest binding, there may be a large amount of reference region data missed, leading to a large reference region classification error.

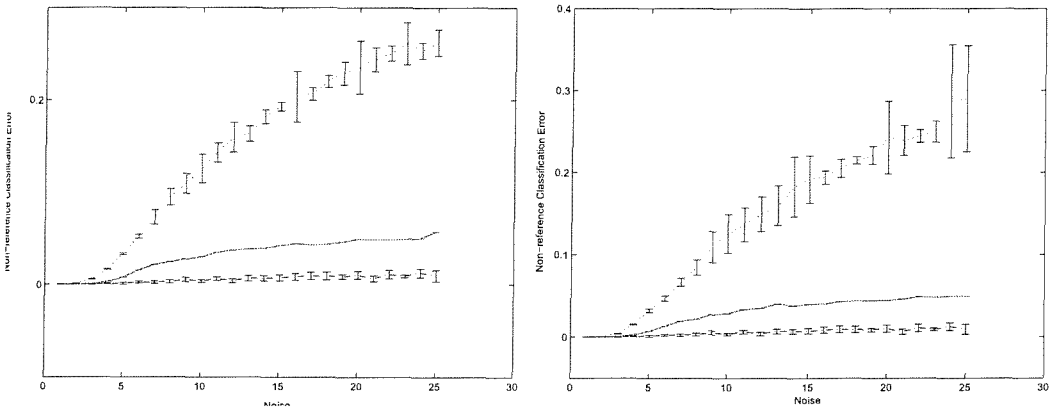
4.5.3.2 The Influence of Cluster Number

In this simulation experiment, we notice that the specification of cluster number is subjective and could significantly influence the unsupervised classification result. To compare the influence of the number of clusters, the noise level in data is fixed at 0.8 and the cluster centre vectors for 4-cluster and 10-cluster classification after K-means



(a) Non-reference region classification error with 100 training data

(b) Non-reference region classification error with 200 training data

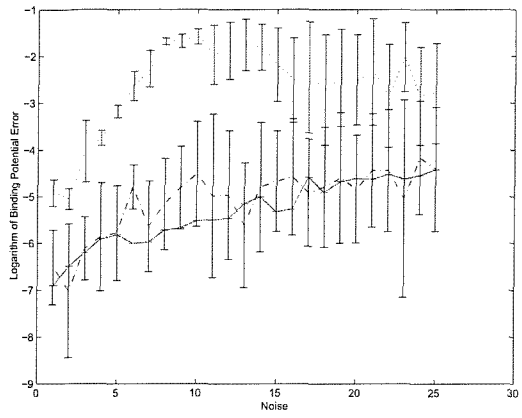


(c) Non-reference region classification error with 500 training data

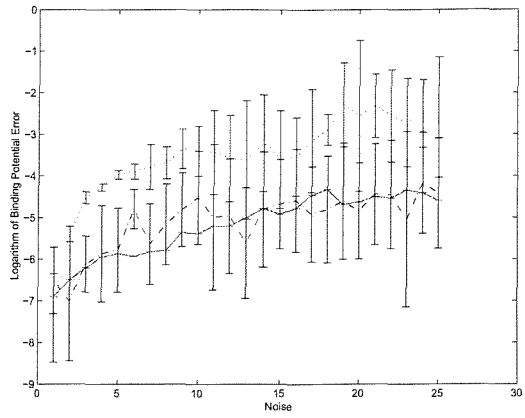
(d) Non-reference region classification error with 1000 training data

FIGURE 4.11: Simulation results on non-reference region classification accuracy, with 10 clusters in unsupervised classification (Red solid : Error bar for semi-supervised classification; Blue dotted: Error bar for supervised classification; Black dashdot: Error bar for unsupervised classification.)

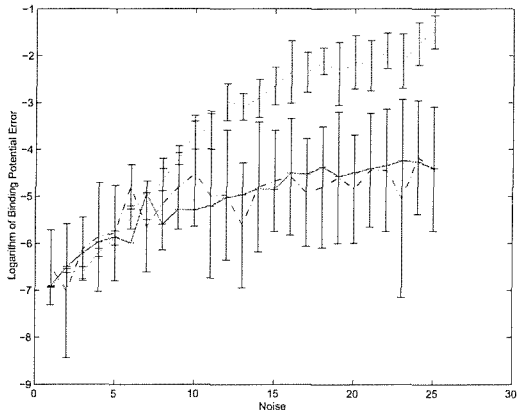
initialisation and Gaussian mixture modeling are plotted separately. As shown in Fig. 4.14 and Fig. 4.15, the cluster centre chosen from the 10 cluster result appears to have less binding but more noise. The higher this cluster number we specify, the more chance of choosing the cluster that has very small binding (i.e. very close to reference region). However, the cluster may have less data inside and is thus subject to noise. A trade off needs to be made in choosing a suitable cluster number.



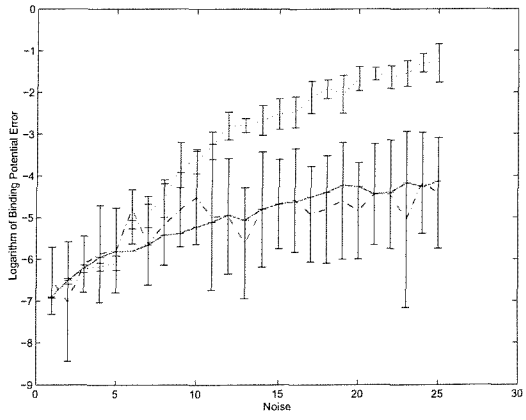
(a) Binding potential error with 100 training data



(b) Binding potential error with 200 training data



(c) Binding potential error with 500 training data



(d) Binding potential error with 1000 training data

FIGURE 4.12: Simulation results on binding potential accuracy, with 10 clusters in unsupervised classification (Red solid : Error bar for semi-supervised classification; Blue dotted: Error bar for supervised classification; Black dashdot: Error bar for unsupervised classification.)

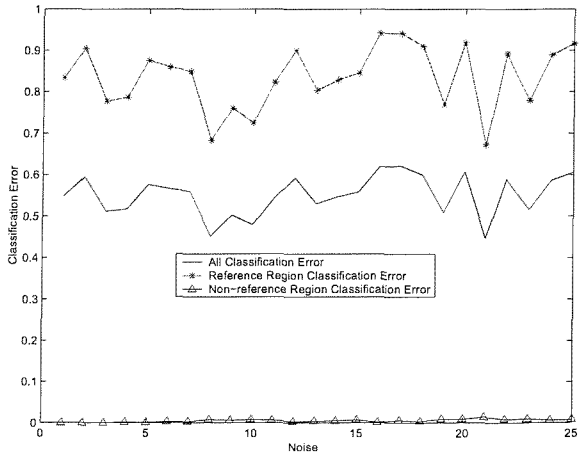
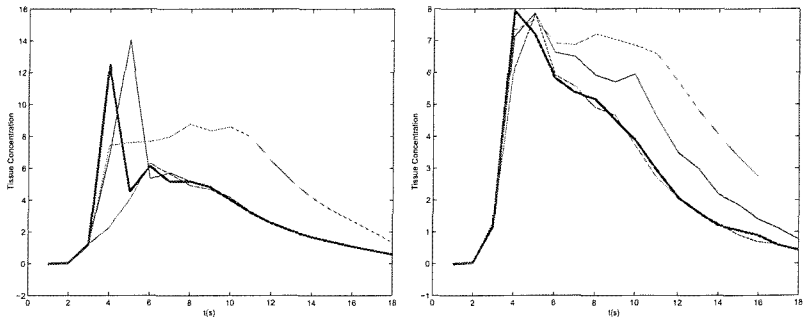


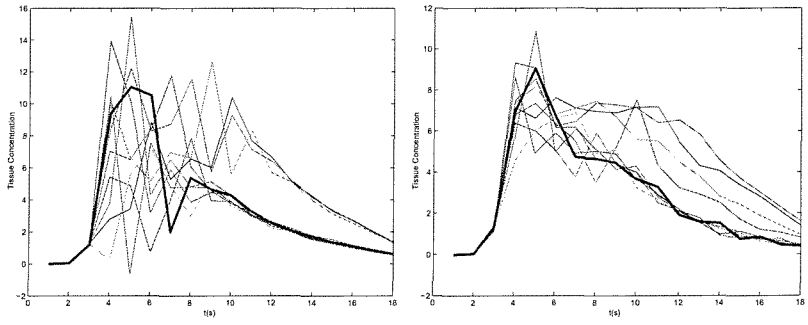
FIGURE 4.13: Three different ways to measure the unsupervised classification error



(a) Cluster centres after k-means initialisation

(b) Cluster centres after Gaussian mixture modeling

FIGURE 4.14: Cluster Centres for four cluster classification



(a) Cluster centres after k-means initialisation

(b) Cluster centres after Gaussian mixture modeling

FIGURE 4.15: Cluster centres for 10 cluster classification

4.6 Real PET Experiments

As in the simulation, the three different methods are compared to extract the reference region from the regions with binding:

- Unsupervised image segmentation using EM with a Gaussian mixture model, as introduced in Section 4.2;
- Supervised segmentation using Bayesian neural networks as described in Section 4.3;
- Semi-supervised method proposed in Section 4.4.

Eighteen [^{11}C](R)-PK11195 PET scans from 17 normal volunteers and one patient were used in the experiment. The data details have been described in Section 2.5. Our goal is to extract the prototype cortex region. This is carried out as follows. 2800 TACs from the cortex region and 2100 TACs from the scalp, thalamus and cerebellum regions

were randomly sampled from seven scans as the labelled data in supervised and semi-supervised segmentation. The data from the cortex region is labelled as the reference region examples, while the 2100 data from other regions are labelled as the other class, the non-reference region examples. The choice of the cortex as the ground truth is based on the evidence that the cortex part of the brain in all healthy scans investigated has no significant binding. The seven scans involved in training are listed in Table 4.1. The segmentation results for all 18 scans will be generated. The segmentation results on the 11 scans not used in the training are independent results for performance evaluation.

Scan used in training
n02791
n02805
n02816
n02833
n02870
n03637
n03657

TABLE 4.1: Scan used in training

In the unsupervised and semi-supervised segmentation, after several trials with various cluster number, the cluster number is set as 10. 10 is also found as a suitable cluster number in the previous simulations and PET image segmentation experiments in literature (Ashburner et al. 1996). In supervised classification, after initial investigations varying the number of hidden nodes, four hidden nodes are found to be suitable for this segmentation problem. In supervised and semi-supervised algorithms, the final classification is carried out automatically by assigning every voxel to the class that has the highest posterior probability. As there is no information about cluster identification available in unsupervised segmentation, the final classification is carried out by choosing the cluster whose centre has lowest BP as the reference class, while the rest are treated as the other class.

4.6.1 Data Pre-processing and Input Normalisation

The characterisation of each voxel to the reference region and the non-reference region is decided by its TAC. An 18 dimensional feature vector for each voxel is extracted from the tracer concentration at 18 different time instants, see Fig. 4.16. This is based on *a priori* knowledge that the information contained in this 18 dimensional feature vector is sufficient to represent the time activity curve and distinguish the reference region and the non-reference region.

Input normalisation ensures that all of the input dimensions and the network weights are of order unity. Data are normalised before input to the network. By treating each of the 18 input dimensions independently, each dimension's mean and variance with respect to

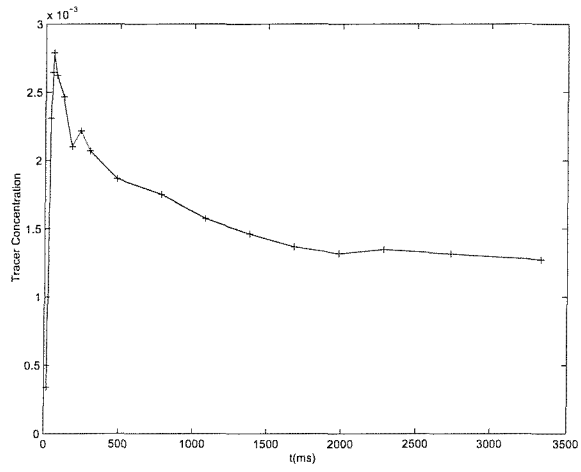


FIGURE 4.16: An example of feature vector extracted from 18 different time instants in a TAC

the training set are calculated. Thus each dimension of the transformed training data has zero mean and unit standard deviation.

4.6.2 Segmentation Result

Fig. 4.17 shows the extracted reference TACs in the 7 scans, where part of data are used in training supervised and semi-supervised classifiers. Three curves are shown in each figure: the curve extracted from the semi-supervised classification; the curve extracted from the supervised neural network and the curve extracted from the unsupervised clustering. In unsupervised classification, the mean curve is obtained by choosing the largest cluster, as the way used in Hammersmith hospital.

To assess the performance across different PET scans, the performance on 11 independent scans are tested. Fig. 4.18 shows the independent testing results. The curves extracted from supervised and semi-supervised classification catch the shape of the reference region curves very well. However, the unsupervised classification performs inconsistently. It fails to capture the features of reference region curve in 4 scans n02833, n02870, n03637, n03657 in Fig. 4.18. In the test scans, unsupervised classification also fails on scans n02907, n02938, n3642, n03661, n04071, n04073 and n02904.

By using the labelled data information, the supervised and semi-supervised classification have successfully learnt the information to discriminate the reference region from the other regions based on the shape of their TACs. In each sub-figure of Fig. 4.17 and Fig. 4.18, the curve extracted from the semi-supervised method has lower tracer concentration value in last several time instants. Comparing these to the curves extracted from the supervised method, it indicates that the extracted reference region TACs extracted from the semi-supervised method have lower binding potential value than the reference region

TACs extracted from the supervised method. Their performances are consistent across subjects.

The output TACs from the supervised neural network capture the shape of the curve well as a result of successfully learning the knowledge in the network. However, the unsupervised clustering technique suffers from the mismatch between the number of unknown underlying clusters and the number of clusters specified in advance. Additionally, manual selection of one cluster makes the process time-consuming and user dependent.

In the case where the cortex has significant binding that makes it unsuitable to be the reference region in PET reference region, the supervised and semi-supervised methods have the advantage of finding the reference region, unrestricted to a specific region. As no ground truth is available, a quantitative performance comparison based on these figures are difficult. However, this is done by test-retest experiments.

4.6.3 Parametric Images

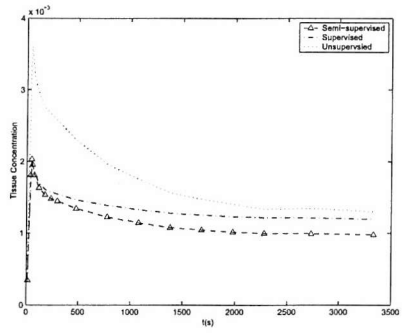
Applying the learnt TAC to the reference region model enables parametric images to be obtained from PET scans. Parametric images such as binding potential images give functional information instead of anatomical information of the scan subject. Binding occurs when the ligand and receptor collide due to diffusion, and when the collision has the correct orientation and enough energy. Measuring the rate and extent of binding provides information on the number of binding sites, and their affinity and accessibility for various drugs. The scan subject's anatomical information can be obtained by registering PET images with an anatomical image (such as CT, MRI) of the subject.

Fig. 4.19 shows an example of binding potential image generated for plane no.20 of scan n03578, a healthy subject. These three figures correspond to using the reference region TAC generated from unsupervised, supervised and semi-supervised method respectively. The binding potential images are generated as follows: Firstly most voxels outside the scalp with low TACs need to be filtered out to avoid unnecessary computation. Let x represent the TAC to be filtered and x_r represent the reference region TAC extracted by the semi-supervised method. The filtering (i.e. thresholding) rule is :

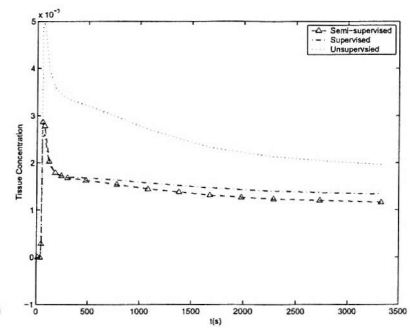
$$\text{if } \sum_{j=1}^{18} x^j > 0.5 \sum_{j=1}^{18} x_r^j, \text{ calculate } BP \text{ value;}$$
$$\text{if } \sum_{j=1}^{18} x^j \leq 0.5 \sum_{j=1}^{18} x_r^j, \; BP = 0.$$

The calculation of BP value is realised by applying the simplified reference region model (Equation (2.3)) and the basis function method.

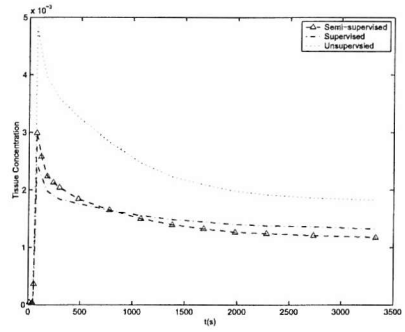
The binding potential image for plane no. 20 of the patient scan n02904 using three reference region extraction techniques are shown in Fig. 4.20. The binding potential



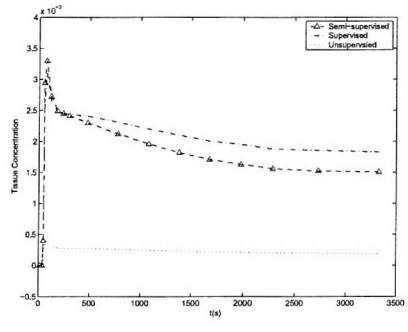
(a) n02791



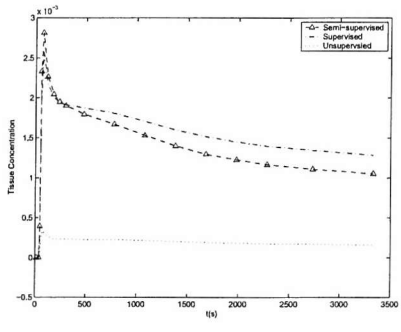
(b) n02805



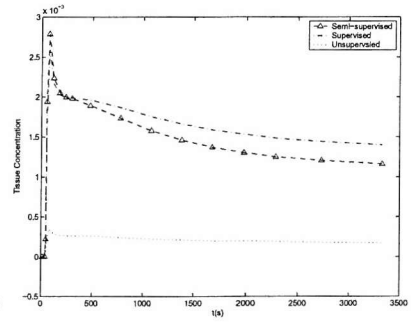
(c) n02816



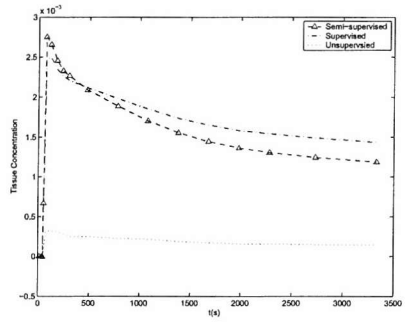
(d) n02833



(e) n02870

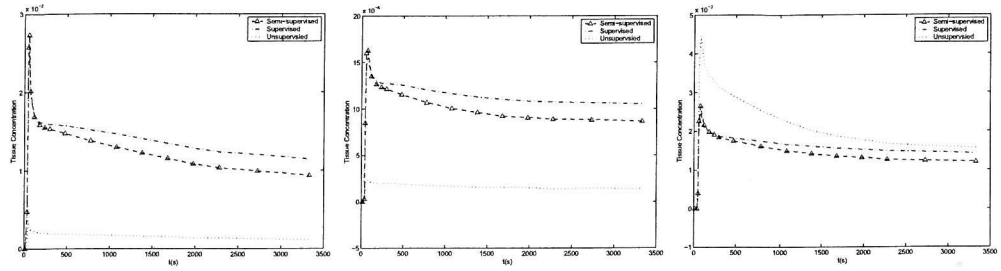


(f) n03637

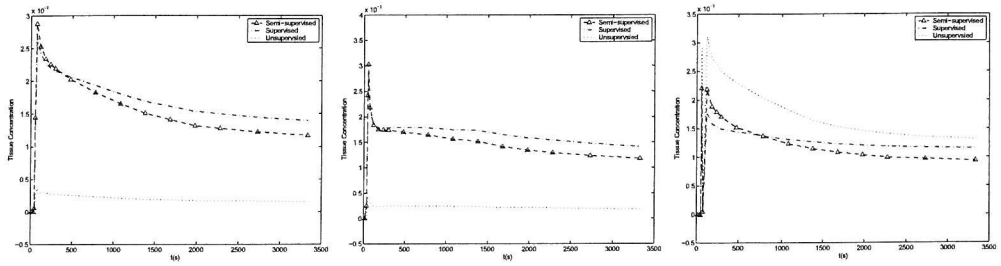


(g) n03657

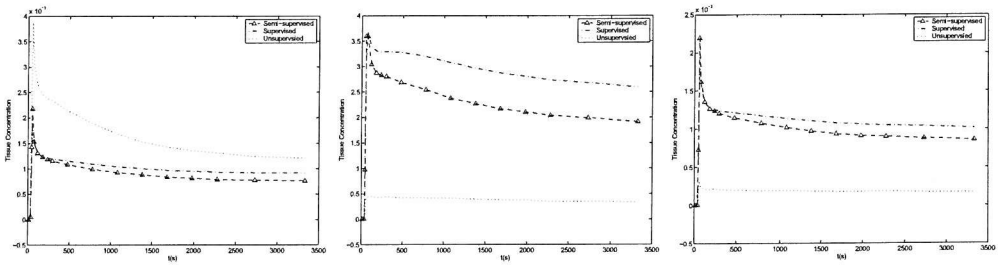
FIGURE 4.17: Results in seven scans (Each figure shows the mean TACs from the reference region extracted from unsupervised, supervised and semi-supervised classification.)



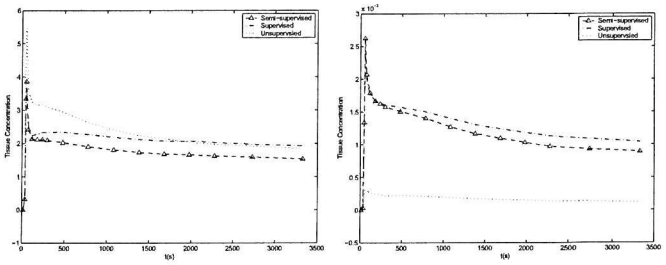
(a) n02907 (b) n02938 (c) n03578



(d) n03642 (e) n03661 (f) n03689

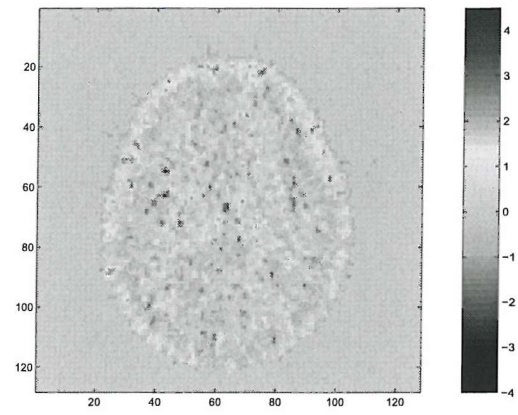


(g) n03694 (h) n04071 (i) n04073

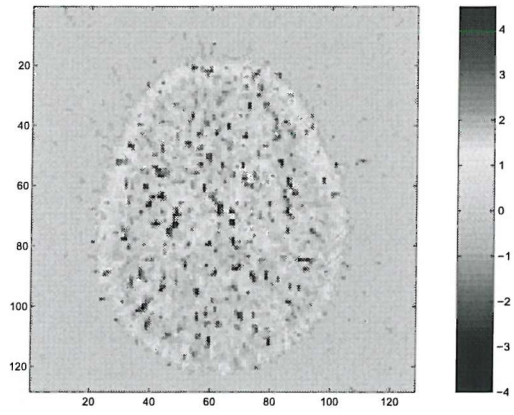


(j) n04128 (k) n02904

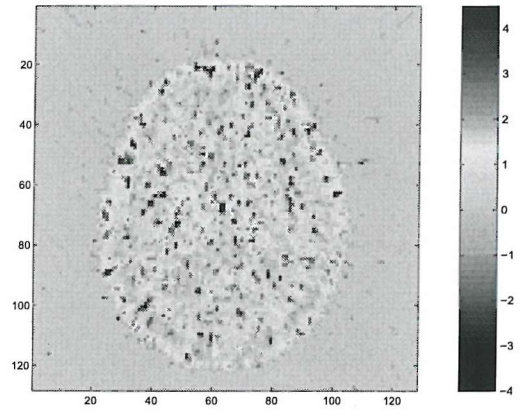
FIGURE 4.18: Results in 11 independent scans (Each figure shows the mean TACs from the reference region extracted from unsupervised, supervised and semi-supervised classification.)



(a) BP image using unsupervised extracted reference region



(b) BP image using supervised neural network extracted reference region



(c) BP image using semi-supervised extracted reference region

FIGURE 4.19: Parametric image of binding potential (BP) for a healthy subject's PET scan n03578 plane 20

images are generated using the same procedure as for the scan n03578. The reference region TAC extracted by the unsupervised segmentation has a large error and fails to extract the reference region TAC, as shown in Fig. 4.18(k). The binding potential image using the unsupervised extracted reference TAC is dramatically different from the others, with the binding potential value in the range $[-15, 30]$. From the binding potential image, it can be seen that the scan contains a strong binding area around the middle-right region.

4.6.4 Improved Tests

As there is no obvious ground truth for performance evaluation in real PET data experiment, improved tests for performance evaluation are introduced.

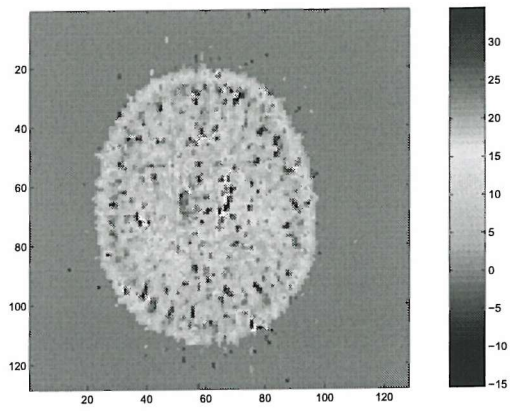
The performance of the different classifiers is to be compared by 4 test-retest experiments. Each test-retest experiment contains a scan pair, where two PET scans are carried out for the same subject over an internal of several days, under the same scanning condition. Thus these two scans should be very close to each other. Any robust and good modelling methodology should be able to generate similar BP values for the same region in these two scans. The difference between the segmentation results for these two scans can be measured and used as a criterion for evaluating the performance of the segmentation method. The scan pairs from the same subjects for test-retest experiments are listed in Table 4.2. The details of the test-retest experiment are illustrated in Fig. 4.21.

Test-Retest Scan Pair
n03578, n03689
n03642, n03657
n03661, n04071
n03694, n04073

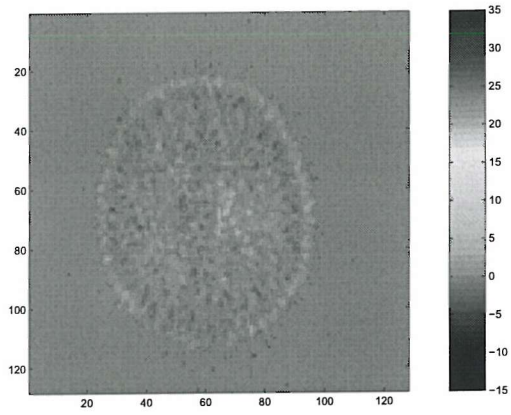
TABLE 4.2: Test-Retest Data Set

Fig. 4.22 shows the test-retest results. Each sub-figure shows the test-retest difference between a scan pair, where the binding potential differences for thalamus, cerebellum and cortex are displayed. The binding potential of each region is calculated by using the extracted reference region TAC and the simplified reference region model. Three bars for each region correspond to three different classification methods.

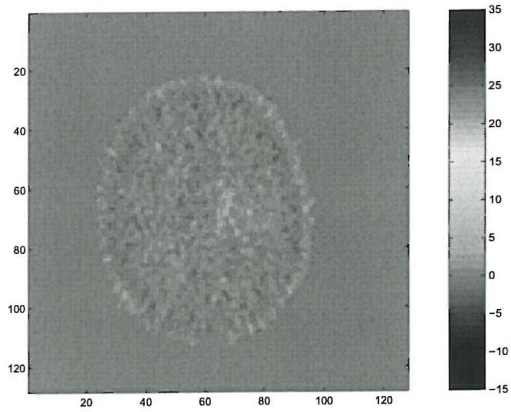
Fig. 4.22 shows that unsupervised classification methods achieve best performance in the first test-retest subject while it gives a large variance in the other three test-retest case. The supervised and semi-supervised classification performs quite stable in all these four test-retest cases. However, there may be more than one factor that influences the test-retest results and making final conclusions needs more experiments and further investigation.



(a) BP image using unsupervised extracted reference region



(b) BP image using supervised neural network extracted reference region



(c) BP image using semi-supervised extracted reference region

FIGURE 4.20: Parametric image of Binding Potential (BP) for patient PET scan n02904 plane 20

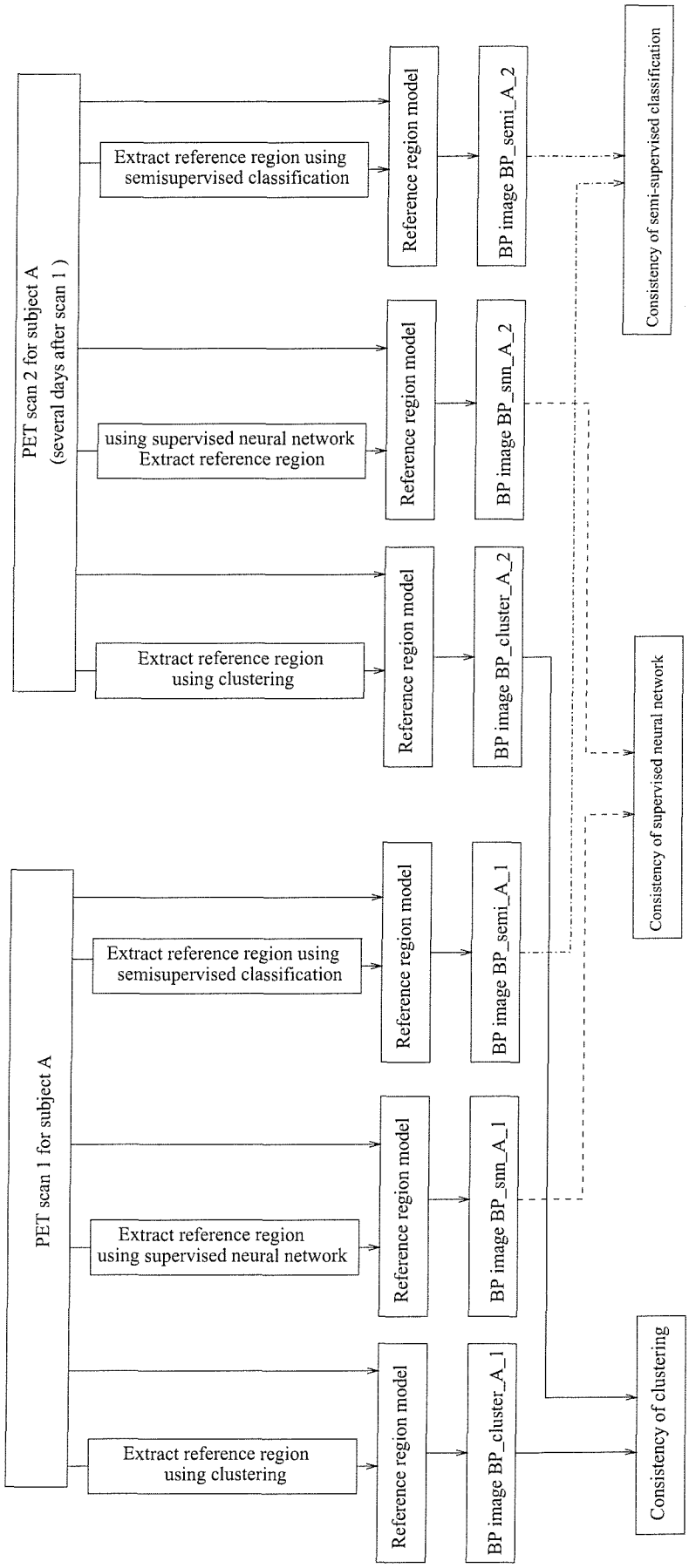


FIGURE 4.21: Improved test scheme

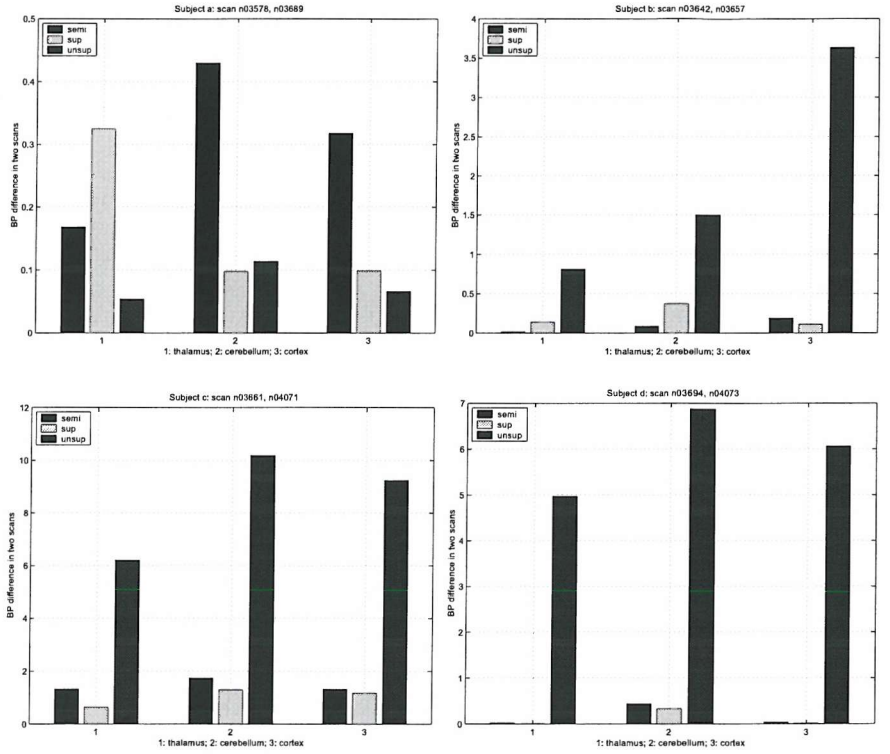


FIGURE 4.22: Test-retest result

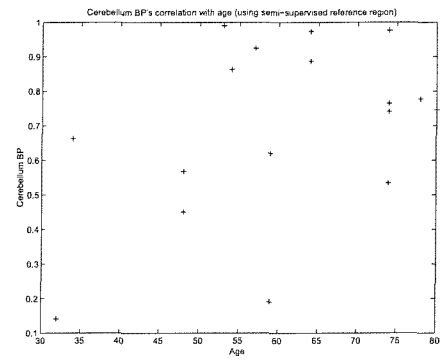
4.6.5 Cerebellum Binding's Correlation with Age

The subject's age is increasingly being recognised as an important factor influencing the brain's function. However, there is a relatively small literature on actual neurochemical differences on their interaction with age (Zubieta et al. 1999; Iidaka et al. 1999). The age-associated variations in the receptor binding is examined using PET data from a group of heathy human subjects at different ages.

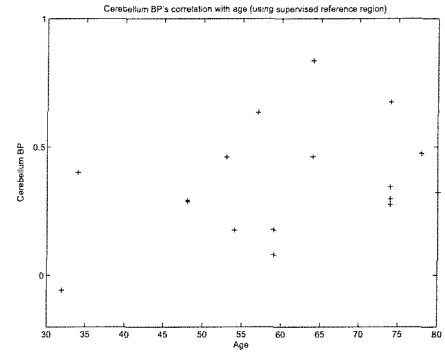
The binding of the cerebellum in 17 healthy scans are examined. Fig. 4.23 shows the estimation of cerebellum in these scans with extracted reference region TAC and the simplified reference region model. Three sub-figures correspond to the results with three different reference region extraction methods. R^2 statistics is a descriptive measure of how the variation in binding potential can by explained by the variability of age. Table 4.3 shows the R^2 statistics analysis of binding potential's variation explained by age for three methods. No obvious correlation of age with cerebellum binding potential value can be found in these 17 scans.

Method	R^2 statistics
Unsupervised	3.28%
Supervised	14.4%
SemiSupervised	2.06%

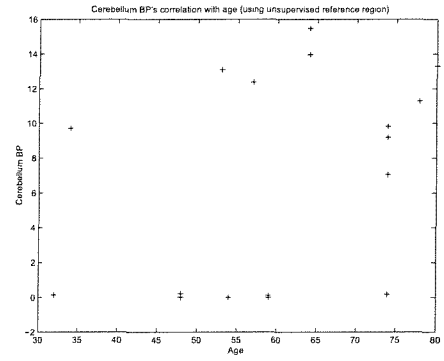
TABLE 4.3: R^2 statistics



(a) Using semi-supervised extracted reference region



(b) Using supervised extracted reference region



(c) Using unsupervised extracted reference region

FIGURE 4.23: Cerebellum binding's correlation with age

4.6.6 Summary

From this series of segmentation and data analysis for 18 [^{11}C](R)-PK11195 PET data, the importance of using labelled data (i.e. expert knowledge) in the segmentation process has been justified . The extracted TAC using methods using labelled data, supervised and semi-supervised classification, leads to more stable and reliable results than using unsupervised clustering. Although the performance evaluation is very difficult for real PET data, the test-retest scheme and cerebellum's binding experiment show that the supervised and semi-supervised extracted reference region TACs are relatively consistent.

4.7 Conclusions

Different data modelling techniques are used in this chapter for PET reference region extraction based on the time-activity curves. Both simulated and real PET data are used. In simulations, performance evaluation can be examined as there is ground truth. The unsupervised segmentation technique with Gaussian mixture modelling is very unstable as there are great uncertainties in the choice of cluster number and the physical meaning of each cluster. The semi-supervised segmentation technique outperforms both unsupervised and supervised Bayesian neural networks in that it uses all the information provided in the data at the same time. The performance of these three segmentation methods is also tested and compared on real PET data. As there is no ground truth available in the real data set, a test-retest scheme is used to examine the consistency of the segmentation methods. Finally the correlation of thalamus's binding with age is also analysed.

By building expert knowledge into the segmentation process, and integrating this knowledge across scans, we may enhance the robustness of the segmentation process and hence, confidence in the extraction of the reference regions. Additionally, the development of these methods to segment reference regions enables automatic generation of parametric maps of microscopic parameters such as binding potential.

As a result of the data independence assumption for all these learning methods, the connections between image voxels are not modelled. As a result, the segmented image appears to be noisy. The next two chapters will deal with incorporating spatial information within the image segmentation problem.

Chapter 5

Image Segmentation

This chapter presents the theory of Markov random fields and show how they can be used in PET reference region segmentation. It is widely recognised that the image segmentation and processing algorithms perform more satisfactory if they have a solid foundation in mathematical models. The Markov random field (MRF) is a class of statistical models for random variables that is well suited for image models. It provides a rich framework for modelling images and other spatial systems (Besag 1986; Ripley and Sutherland 1990); and its flexibility making it applicable in various image segmentation problem. In this chapter, Markov random field models are used to incorporate local dependency into the segmentation process. The theory of Markov random fields, including the details of the model, how to incorporate them into various learning process and the related optimisation methods, will be described.

5.1 Markov Random Field Models

Contextual constraints are ultimately necessary in vision system. The use of context is indispensable to analysic images. The idea of using contextual information in image modelling and segmentation has given rise to algorithms based on Markov random fields (MRFs) (Geman and Geman 1984; Derin et al. 1984; Li 1995). Markov random field theory provides a convenient and consistent way for modelling context dependent entities such as image voxels and correlated features. This is achieved through characterising mutual influences among such entities using conditional MRF distributions.

A Markov random field consists of a collection or a lattice of random variables (voxel values) with local interactions. MRF models define a probability distribution over a set of interacting variables. A number of concepts are fundamental to the use of MRFs, namely neighbourhoods, cliques, the Hammersley-Clifford theorem, Gibbs distributions and potential functions.

5.1.1 Neighbourhood System and Cliques

If the observed image $\mathbf{x} = \{\mathbf{x}_i, i \in \mathcal{S}\}$ is defined on a rectangular lattice which contains n voxels,

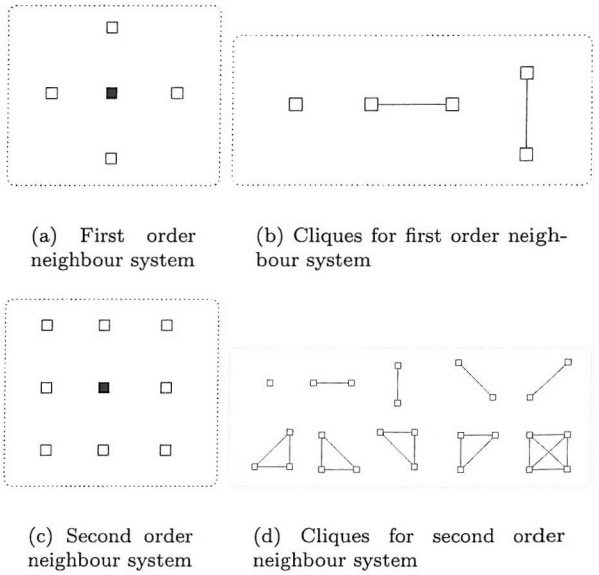
$$\mathcal{S} = \{i | 1 \leq i \leq n\}. \tag{5.1}$$

Each voxel can be called a site. For each site i, there is a label z_i. The labelling problem is to assign a label from the label set to each of the sites in $\mathcal{S}$. The sites in $\mathcal{S}$ are related to one another via a neighbourhood system, defined as

$$\mathcal{N} = \{\mathcal{N}_i | i \in \mathcal{S}\}, \tag{5.2}$$

where $\mathcal{N}_i$ is the set of sites neighbouring i.

Neighbourhoods on the voxel lattice are defined in the following way. The set of neighbours of lattice point z_i are those points j such that the functional form of the local conditional probability, $p(\mathbf{z}_i | \mathbf{z}_1, \cdots, z_{i-1}, z_{i+1}, \cdots, z_n)$, depends only on z_j; if site i is a neighbour of site j, then site j is also a neighbour of site i. A clique is a set of points which are all neighbours of each other. Fig. 5.1 shows the first order and second order neighbourhood system and corresponding cliques for images. Clearly, the number and the types of cliques grows very rapidly with the increase of the neighbourhood size, and allows almost unlimited forms of interaction between voxels in a neighbourhood to be specified. In applications, first and second-order clique systems are widely used.



(a) First order neighbour system

(b) Cliques for first order neighbour system

(c) Second order neighbour system

(d) Cliques for second order neighbour system

FIGURE 5.1: First order and second order neighbourhood systems and cliques

5.1.2 Markov Random Fields

As stated above, MRF models are built by specifying the local probability structure - how voxels in a neighbourhood interact - and this is a natural way of modelling. The difficulty lies in the lack of a formulation specifying these conditional probabilities which leads to a consistent form for the joint probability. The distribution of image labels $\mathbf{z} = \{z_i, i \in \mathcal{S}\}$ is a Markov random field over $\mathcal{S}$ if

$$p(\mathbf{z}) > 0 \tag{5.3}$$

$$p(z_i|z_{\mathcal{S}-\{i\}}) = p(z_i|z_{\mathcal{N}_i}). \tag{5.4}$$

This depicts the local characteristics of $\mathbf{z}$. That is in MRFs, only neighbouring labels have direct interaction with each other. The conditional distribution of any variable in the random field given the remaining field is identical to the distribution conditioned only on values in a finite size local clique of neighbouring voxel values.

5.1.3 Gibbs Random Fields

A set of random variables $\mathbf{z}$ is a Gibbs Random Field (GRF) on $\mathcal{S}$ with respect to $\mathcal{N}$ if and only if its configurations obey a Gibbs distribution of the form

$$p(\mathbf{z}) = Q^{-1}e^{-\frac{1}{T}U(\mathbf{z})}, \tag{5.5}$$

where Q is a normalising constant called the *partition function* and T is a constant. $U(\mathbf{z})$ is the *energy function.* The energy

$$U(\mathbf{z}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{z}). \tag{5.6}$$

is the sum over all possible cliques $\mathcal{C}$ on the lattice and potentials $V_c(\mathbf{z})$ is the *potential function* depending on the variables z_i within the clique $\mathcal{C}$.

5.1.4 Markov-Gibbs Equivalence

The Hammersley-clifford theorem (Hammersley and Clifford 1971) establishes the equivalence of the Markov random field and the Gibbs distribution, The theorem states that $\mathbf{z}$ *is a Markov random field on* $\mathcal{S}$ *with respect to* $\mathcal{N}$ *if and only if* $\mathbf{z}$ *is a Gibbs random field on* $\mathcal{S}$ *with respect to* $\mathcal{N}$.

The proof that a Gibbs random field is a Markov random field is relatively straightforward and shown in the following:

Let $p(\mathbf{z})$ be a Gibbs distribution on $\mathcal{S}$ with respect to neighbourhood system $\mathcal{N}$. The conditional probability

$$p(z_i|z_{\mathcal{S}-i}) = \frac{p(z_i, z_{\mathcal{S}-i})}{p(z_{\mathcal{S}-i})} = \frac{p(\mathbf{z})}{\sum_{z_i'} p(\mathbf{z}')} \tag{5.7}$$

where $\mathbf{z}' = \{z_1, \cdots, z_{i-1}, z_{i+1}, \cdots, z_n\}$. Expanding $p(\mathbf{z}) = Q^{-1} e^{-\sum_{c \in \mathcal{C}} V_c(\mathbf{z})}$ gives

$$p(z_i|z_{\mathcal{S}-i}) = \frac{e^{-\sum_{c \in \mathcal{C}} V_c(\mathbf{z})}}{\sum_{z_i'} e^{-\sum_{c \in \mathcal{C}} V_c(\mathbf{z})}} \tag{5.8}$$

The potential functions that do not include voxel i cancel each other in the dominator and numerator, thus

$$p(z_i|z_{\mathcal{S}-i}) = \frac{e^{-\sum_{c \in \mathcal{A}} V_c(\mathbf{z})}}{\sum_{z_i'} e^{-\sum_{c \in \mathcal{A}} V_c(\mathbf{z})}} \tag{5.9}$$

where $\mathcal{A}$ consists of cliques that contains voxel i. That is it only depends on i's neighbours. So it is a Markov random field. This proves that a Gibbs random field is a Markov random field. The proof that a Markov random field is a Gibbs random field is more involved. For details see (Clifford 1990).

This theorem provides a simple way of specifying the joint probability $p(\mathbf{z})$ in Markov random field using Equation (5.5).

In summary, Markov random field model builds up hierarchically in the following manner:

- Specify a neighbourhood system.
- Determine the cliques associated with that neighbourhood system.
- Specify the potential functions associated with each clique.
- Form $p(\mathbf{z}) = Q^{-1} e^{-U(\mathbf{z})}$.

A MRF is said to be homogenous if potential function $V_c(\mathbf{z})$ in the Gibbs distribution (Equation (5.5)) is independent of the relative position of the clique $\mathcal{C}$. It is said to be isotropic if $V_c(\mathbf{z})$ is independent of the orientation of $\mathcal{C}$. For mathematical and computational convenience, the homogeneity is assumed in most MRF models.

5.2 Segmentation Technique

Markov random field models allow the spatial continuity of image voxel labels to be incorporated into the modelling process. This section details a heuristic and statistical

physics based optimisation method to find the maximum a posteriori (MAP) solution to image segmentation based on MRFs. An Example of PET image segmentation is given to illustrate the process.

5.2.1 MAP Estimation

Let $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ denote the observation of an image which contains n voxels while $\mathbf{z} = \{z_1, z_2, \cdots, z_n\}$ is the true label for the image.

A maximum a posteriori (MAP) estimation is achieved at $\mathbf{z} = \mathbf{z}^*$, where

$$\mathbf{z}^* = \arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}). \tag{5.10}$$

As data model $p(\mathbf{x})$ is known,

$$p(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{z})p(\mathbf{z}). \tag{5.11}$$

5.2.2 MRF-MAP Estimation

Assuming the data are conditionally independent,

$$p(\mathbf{x}|\mathbf{z}) = \prod_{i \in \mathcal{S}} p(\mathbf{x}_i|z_i). \tag{5.12}$$

and

$$p(\mathbf{x}_i|z_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\mathbf{x}_i - z_i)^2}{2\sigma^2}}. \tag{5.13}$$

Then

$$p(\mathbf{x}|\mathbf{z}) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\sum_{i=1}^{n} \frac{(\mathbf{x}_i - z_i)^2}{2\sigma^2}}. \tag{5.14}$$

The prior model $p(\mathbf{z})$ is a Gibbs distribution of the form in Equation (5.5), where

$$U(\mathbf{z}) = \sum_{i=1}^{n} \sum_{i' \in \mathcal{N}_i} (z_i - z_{i'})^2. \tag{5.15}$$

Then from Equation (5.11), the posterior probability

$$p(\mathbf{z}|\mathbf{x}) \propto e^{-U(\mathbf{z}|\mathbf{x})} \tag{5.16}$$

where

$$U(\mathbf{z}|\mathbf{x}) = \sum_{i=1}^{n} \frac{(z_i - \mathbf{x}_i)^2}{2\sigma^2} + \sum_{i=1}^{n} \sum_{i' \in \mathcal{N}_i} (z_i - z_{i'})^2. \tag{5.17}$$

Thus the MAP estimation is equivalently found by

$$\mathbf{z}^* = \arg\min_{\mathbf{z}} U(\mathbf{z}|\mathbf{x}). \tag{5.18}$$

However, the optimisation of Equation (5.18) is a difficult problem. A computationally expensive method - simulated annealing (Geman and Geman 1984) - is often used to find a global minimum for combinational optimisation problems. This is a statistical mechanics formulation of the optimisation problem. Simulated annealing is capable of - at least at the limit of infinite computing time - localising the mode of the energy function even in the presence of local minima.

The function in Equation (5.17) satisfies the convexity property. Convexity guarantee simplifies the optimisation process. The local minimisation algorithms such as iterated conditional modes (ICM) (Besag 1986) could be used. This method finds an approximate solution to Equation (5.18), by iteratively minimising the function with respect to each voxel z_i,

$$z_i^* = \arg\min_{z_i} U(z_i|\mathbf{x}_i) \tag{5.19}$$

where

$$U(z_i|\mathbf{x}_i) = \frac{(z_i - \mathbf{x}_i)^2}{2\sigma^2} + \sum_{i' \in \mathcal{N}_i} (z_i - z_{i'})^2. \tag{5.20}$$

The initial value for $\mathbf{z}$ can be set to the maximum likelihood solution for each voxel, which ignores the local dependence and merely choose z_i to maximise $p(z_i|\mathbf{x}_i)$ at each i separately. The solution is obtained by iterating Equation (5.19) for a certain number of cycles or until convergence. This approach will converge to a local minimum. Often in the iteration process, the raster is varied from cycle to cycle to reduce the small directional effects.

5.3 Unsupervised Image Segmentation

5.3.1 Joint Segmentation and Parameter Estimation

When the model parameters are known, image segmentation can be achieved by *ad hoc* ICM iterations, as described in the previous section. When the model parameters are unknown, an algorithm that enables joint parameter estimation and image segmentation is needed. EM, mean field annealing and Markov chain Monte Carlo methods can all be used to find (or approximate) the solution to the problem with different computational demands. These segmentation techniques are referred to as unsupervised image segmentation here as no labels are involved.

5.3.1.1 EM based Optimisation

As before, let $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ denote the observation of an image which contains n voxels while $\mathbf{z} = \{z_1, z_2, \cdots, z_n\}$ is the true label for the image. After setting up the model for $\mathbf{z}$ and $\mathbf{x}$, we need to estimate the distribution $\mathbf{z} = z_1, z_2, \cdots, z_n$ and the parameter Φ. A general and effective algorithm for solving this problem is the Expectation Maximisation (EM) (Dempster et al. 1977) algorithm. In chapter 3, EM algorithm is introduced to solve the incomplete data problem in unsupervised and supervised classification.

Starting with an initial estimate of $\widehat{\Phi}^{(0)}$, the algorithm iterates the following two steps:

- E step: find the function $Q(\Phi|\widehat{\Phi}^{(t)}) = E[\log p(\mathbf{x}, \mathbf{z}|\Phi)|(\mathbf{x}, \widehat{\Phi}^{(t)})]$
- M step: find $\widehat{\Phi}^{(t+1)} = \arg\max Q(\Phi|\widehat{\Phi}^{(t)})$

For the observed image model $\mathbf{x}$ in Equations (5.12) and (5.13), after initialising parameters $\Phi^{(0)} = \{(\boldsymbol{\mu}_1^{(0)}, \boldsymbol{\Sigma}_1^{(0)}), (\boldsymbol{\mu}_2^{(0)}, \boldsymbol{\Sigma}_2^{(0)}), \cdots, (\boldsymbol{\mu}_K^{(0)}, \boldsymbol{\Sigma}_K^{(0)})\}$ and $p(z_{ik}^{(0)}|\Phi^{(0)})$, the parameter Φ is updated by (in the following $p(\cdot)$ is a simplified notation of $p(\cdot|\Phi)$):

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)}, \tag{5.21}$$

$$\left(\boldsymbol{\Sigma}_k^2\right)^{(t+1)} = \frac{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})^T(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})}{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)}, \tag{5.22}$$

where

$$\begin{aligned} p(z_i^{(t)} = k|\mathbf{x}_i) &= \frac{p(\mathbf{x}_i|z_i^{(t)} = k)P(z_i^{(t)} = k)}{p(\mathbf{x}_i)} \\ &= \frac{p(\mathbf{x}_i|z_i^{(t)} = k)P(z_i^{(t)} = k)}{\sum_{i=1}^{n} p(\mathbf{x}_i|z_i^{(t)} = k)P(z_i^{(t)} = k)}. \end{aligned} \tag{5.23}$$

$p(\mathbf{x}_i|z_i^{(t)} = k)$ is calculated from Equation (5.13). In the case that z_i's are independent, the prior can be updated by

$$p(z_i^{(t+1)} = k) = \frac{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)}{N}. \tag{5.24}$$

When the assumption of independence of image voxels does not hold, the estimation of the prior model $p(z_i^{(t+1)} = k)$ $(k = 1, 2, \cdots, K)$ is very difficult. An approximate

technique is considered here, using a simple state prior model (Besag 1986; Zhang et al. 1994), approximate Equation (5.24) by

$$p(z_i = k) \approx p(z_i = k|\widehat{z}_l, l \in \mathcal{N}_i). \tag{5.25}$$

By substituting an independent voxel prior with a Markov random field model, an approximate implementation of the EM procedure for the MRF model can be obtained. Here a simple Markov random field model

$$p(z_i = k|\widehat{z}_l, l \in \mathcal{N}_i) = \frac{e^{\beta\delta_i(k)}}{\sum_{k=1}^{K} e^{\beta\delta_i(k)}}, \tag{5.26}$$

is used, where $\delta_i(k)$ is the number of neighbours of i in state k and $\beta > 0$ is a parameter controlling the influence of neighbouring voxels $\mathcal{N}_i$ on voxel i. The neighbour of voxel i is selected to be 3×3 voxel grid.

Currently, Markov random field models have been used for modelling static images where $\mathbf{x}_i$ for each voxel i is a scalar. As shown in Equation (5.5), it is the hidden labelled image $\mathbf{z}$ instead of $\mathbf{x}$ that is modelled as a Markov random field, so the MRF model can be directly extended to dynamic image modelling.

The algorithms used to segment images in the independent voxel case and dependent voxel case (MRF model) are listed in Table 5.1 and Table 5.2 respectively.

β in Equation (5.26) is a regularisation parameter. The larger the parameter, the more influence the pixel's neighbour on the pixel. Optimal parameter estimation of the MRF model parameter is a difficult problem as the intractable nature of the partition function (See Equation (5.5)).

TABLE 5.1: Segmentation Algorithm for Independent Voxel Case

Given data set $\mathbf{x}_i, i = 1, 2, \cdots, n$;
1. Data preprocessing;
2. Set the number of Gaussian distributions, K;
3. Initialise the parameter Φ;
4. Apply K-means until a maximum iteration number is reached;
5. Apply EM iteration until convergence or a maximum iteration number is reached:
   5a. Calculate each probability $p(\mathbf{x}_i)$ using Equation (5.13), (5.23);
   5b. Update the prior using Equation (5.24); Update parameter Φ using Equation (5.21), (5.22);
6. Segment the image: $z_i = \arg\max_k p(z_i = k|\mathbf{x}_i)$.

TABLE 5.2: Segmentation Algorithm for Dependent Voxel Case

Given data set $\mathbf{x}_i, i = 1, 2, \cdots, n$;
1. Data preprocessing;
2. Set the number of Gaussian distribution, K;
3. Initialise the parameter Φ;
4. Apply K-means until a maximum iteration number is reached;
5. Set parameter β;
6. Iterate until convergence or the maximum iteration number is reached:
   6a. Calculate each probability $p(\mathbf{x}_i)$ using Equation (5.13), (5.23);
   6b. Update prior using Equation (5.41); Update parameter Φ using Equation (5.21), (5.22);
7. Segment the image: $z_i = \arg\max_k p(z_i = k|\mathbf{x}_i)$.

5.3.1.2 Mean Field Theory and Mean Field Annealing

If the random variable set $\mathbf{z} = \{z_i, i \in S\}$ is a Markov random field, then it has a Gibbs distribution: $p(\mathbf{z}) = Q^{-1}e^{-U(\mathbf{z})}$ where Q is the *partition function*, $U(\mathbf{z})$ is the *energy function*. The energy $U(\mathbf{z}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{z})$ is the sum of clique potentials $V_c(\mathbf{z})$ over all possible cliques $\mathcal{C}$. Mean field theory (Chandler 1987) concerns the following problem: How to find the mean of the above field?

The mean value at site i is given by

$$\begin{aligned} \langle z_i \rangle &= \sum_z z_i p(z) \\ &= Q^{-1} \sum_z z_i e^{-U(z)}. \end{aligned} \tag{5.27}$$

However, due to the interaction between the z_i's, the sum in Equation (5.27) involves all the possible realisations of the MRF and the precise calculation of $\langle z_i \rangle$ is not computational feasible (Chandler 1987).

The mean field assumption maintains that the influence of $z_{i'}, i' \neq i$ in the calculation of $\langle z_i \rangle$ can be approximated by the influence of $\langle z_{i'} \rangle$. For the sake of simplicity, here only the second-order cliques are considered. That is the energy function can be written as

$$U(\mathbf{z}) = \sum_i \left[V_c(u_i) + \frac{1}{2} \sum_{i' \in N_i} V_c(z_i, z_{i'}) \right]. \tag{5.28}$$

where $V_c(\cdot)$ and $V_c(\cdot, \cdot)$ represent clique potentials for a single site and a pair of neighbouring sites respectively. The mean field local energy at site i is

$$U_i^{mf'} = V_c(z_i) + \sum_{i' \in N_i} V_c(z_i, \langle z_{i'} \rangle). \tag{5.29}$$

Define:

$$\begin{aligned} U_i^{mf}(z_i) &= U(z)|_{z_{i'}=\langle z_{i'}\rangle, i'\neq i} \\ &= U_i^{mf'} + R_i^{mf'}(\langle z_{S-i}\rangle). \end{aligned} \tag{5.30}$$

Similarly, define

$$\begin{aligned} Q_i^{mf} &= \sum_{z_i} \exp\left[-\beta U_i^{mf}(z_i)\right] \\ &= Q_i^{mf'} \exp\left[-\beta R_i^{mf'}(\langle z_{S-i}\rangle)\right], \end{aligned} \tag{5.31}$$

where $Q_i^{mf'}$ is called the mean field local partition function, given by

$$Q_i^{mf'} = \sum_{z_i} \exp\left[-\beta U_i^{mf'}(z_i)\right] \tag{5.32}$$

Then, by the mean field approximation

$$\begin{aligned} \langle z_i\rangle &\simeq \frac{1}{Q_i^{mf}} \sum_{z_i} \exp\left[-\beta U_i^{mf}(\langle z_i\rangle)\right] \\ &= \frac{1}{Q_i^{mf'}} \sum_{z_i} \exp\left[-\beta U_i^{mf}(\langle z_i\rangle)\right]. \end{aligned} \tag{5.33}$$

Mean field theory suggests that when estimating the mean field at i, the influence of the field at other sites can be approximated by that of their mean. Therefore, the fluctuations on these sites are neglected. The ability to calculate the mean field component at site i in terms of its neighbours allows the complete mean field to be found using an iterative update procedure. If the inverse temperature parameter β is increased so that $\beta \to \infty$, then the system undergoes an annealing process, converging to the MAP estimate, and is known as mean field annealing.

Zhang (1992) and Greiger and Firosi (1991) propose a further approximation, neglecting the fluctuations at neighbouring sites in the partition function Q.

$$\begin{aligned} U(\mathbf{z}) &= \sum_c V_c(\mathbf{z}) \\ &= \sum_i \left[V_c(z_i) + \frac{1}{2} \sum_{i'\in N_i} V_c(z_i, z_{i'}) \right] \\ &\simeq \sum_i \left[V_c(z_i) + \frac{1}{2} \sum_{i'\in N_i} V_c(z_i, \langle z_{i'}\rangle) \right] \\ &= U^{mf}(\mathbf{z}) \end{aligned} \tag{5.34}$$

and

$$\begin{aligned} Q &\simeq U^{mf}(\mathbf{z}) = \sum_{\mathbf{z}} \exp\left[-\beta U^{mf}(\mathbf{z})\right] \\ &= \prod_{i \in S} \sum_{z_i} \exp\left[-\beta\left(V_c(z_i) + \frac{1}{2}\sum_{i' \in N_i} V_c(z_i, \langle z_{i'}\rangle)\right)\right] \end{aligned} \tag{5.35}$$

This is very similar to Besag's pseudo-likelihood approximation. The only difference is it is conditioned on the mean values instead of current state. While pseudo-likelihood approximation is an *ad-hoc* method, the mean field approximation has a plausible physical meaning.

5.3.1.3 Simulated Annealing

Simulated annealing is a numerical optimisation technique based on the principles of thermodynamics. Annealing refers to the process in which a solid material is first melted and then allowed to cool by slowly reducing the temperature. The primary advantage of simulated annealing is the ability to move from local optima. Thus the ability to find the global optimum is not dependent upon the starting point.

The temperature T in Equation (5.5) is set to a constant. If we allow the value of T to be varied, then it will alter the kurtosis of the distribution and is therefore known as the "temperature" by analogy with physical systems governed by Gibbs distributions. Simulated annealing uses this as a control parameter to enable the MAP configuration of the field to be found. For large values of T the distribution in Equation (5.5) is relatively flat. The temperature is then steadily lowered, making the local and global minima become further pronounced. Then as the temperature tends to zero, the surface moves towards a set of inverted peaks, the deepest at the global minimum.

Geman and Geman (1984) provide a Gibbs sampler to perform the simulated annealing, called Markov chain Monte Carlo. First a cooling schedule is designated for the Gibbs distribution temperature parameter. The temperature begins in a hot state, causing the energy surface to be relatively flat, since the system is relatively excited. The Gibbs sampler can be used to draw samples from the distribution of Equation (5.5) at any temperature. However as the temperature of the distribution is altered the transition probabilities will change, causing the Gibbs sampler to sample from an inhomogeneous Markov chain. Geman and Geman (1984) proved that if the temperature is reduced according to

$$T = \frac{C}{\log(1+k)} \tag{5.36}$$

where k denotes the number of full scans of the lattice and C is a fixed constant, then the Gibbs sampler will converge to the uniform distribution over all the configurations

with maximum probability.

Two important factors governing the cooling schedule is the temperature and the step size for perturbation. Geman also showed that the constant $C = N \triangle U$, where N is the number of sites in the lattice and $\triangle U$ is the maximum difference in energy function for two configurations which differ at only one site. This is a huge number, making simulated annealing computationally expensive. Many authors have carried out research in an attempt to find an optimal annealing schedule or an adaptive annealing schedule. Much of these are reviewed in Ruanaidh and Fitzgerald (1996) and Neal (1993).

Barker and Rayner (2000) proposed a reversible jump Markov chain Monte Carlo method for image segmentation, enabling the sampling to include the cluster number. The reversible jump was developed by Green (1995) to allow a Metropolis-Hastings based algorithm to sample the model order - the number of clusters from the posterior distribution.

5.3.2 Multi-resolution Segmentation

In recent years, substantial interest has been devoted to developing multi-resolution algorithms to the image and signal processing problems. One reason for this is that the multi-resolution methods mimic the human vision system. Multi-resolution enables combination of local and global information in images. In Bouman and Shapiro (1994), a multi-resolution random field model replaced the Markov random field model in image segmentation and a sequential maximum a posterior optimisation was proposed.

Founded in re-normalisation group theory (Chandler 1987), the multi-resolution approach to image segmentation usually begins by establishing the model parameters *a priori* at the highest resolution of interest in the original image. The image is then repeatedly down-sampled, while fresh model parameters are estimated at each of these ensuing resolutions. Segmentation is first carried out at the coarsest resolution through the use of a standard segmentation algorithm. The resulting segmentation is then used to initialise or constrain the segmentation process at the next coarsest level. The scheme is repeated at each resolution until the segmentation is carried out at the highest one.

5.3.3 Summary

In terms of learning, the above image segmentation methods (with specified or unspecified model parameters) are unsupervised methods, as no image data are labelled. The Markov random field model is one of the most efficient statistical models for image segmentation. However, most works related to its application belong to the unsupervised learning category and cannot include labelled data. Consequently, unsupervised learning suffers from two problems: choosing and validating the correct number of clusters and

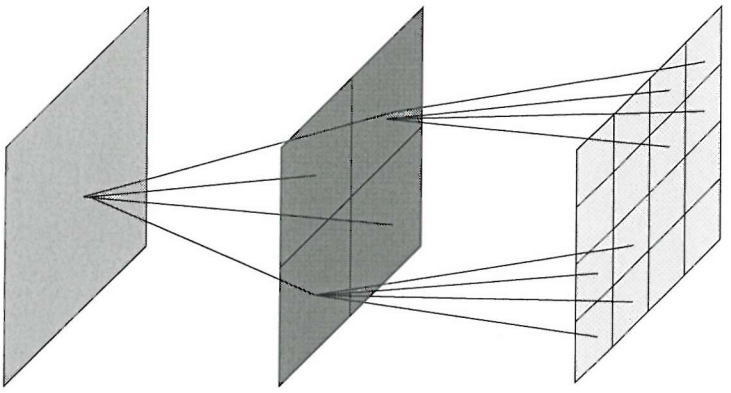


FIGURE 5.2: Multi-resolution image segmentation

ensuring that algorithmic cluster labels correspond to meaningful physical labels. This is why supervised classifiers such as Bayesian networks, support vector machines etc. can be preferred for data classification. However, these traditional supervised learning methods provide no way to incorporate either image models or unlabelled data.

In the following, two new methods for incorporating labelled samples in image segmentation will be presented, termed supervised and semi-supervised image segmentation respectively. The first one is a baseline method, hierarchically applying a neural network to the image models in the learning process, details of which will be shown in Section 5.4. In Section 5.5, a new semi-supervised image segmentation scheme is proposed by using both labelled and unlabelled data as well as imposing local constraints on image voxels in the learning process.

5.4 Supervised Image Segmentation - A Hierarchical Method

In the following, the image to be segmented will be denoted $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ with mixture (cluster) labels $z = \{z_1, z_2, \cdots, z_n\}$, $z_i \in \{1, 2, \cdots, K\}$ and class labels $c = \{c_1, c_2, \cdots, c_n\}$, $c_i \in \{1, 2, \cdots, J\}$ to be decided. $\mathbf{x}_i$ $(i = 1, 2, \cdots, n)$ is an m-dimensional feature vector for voxel i. Note that the class label c is distinguished from mixture label z, as z is not necessarily the meaningful physical class labels c. The image is often either under-segmented or over-segmented. In addition, $(\mathbf{x}^l, c^l) = \{(\mathbf{x}_1^l, c_1^l), (\mathbf{x}_2^l, c_2^l), \cdots, (\mathbf{x}_{n^l}^l, c_{n^l}^l)\}$ are the labelled samples and their associated class labels.

Supervised classification learns the mapping from the input feature to an output class label from labelled examples. Thus the classifier can be used to determine the class label c of the feature for image voxels $\mathbf{x}$ (Fig. 5.3). Because of the success of supervised learning methods (such as neural networks) in pattern recognition, they have been applied to image segmentation such as texture and within MPEG4 (Jain and Karu 1996; Doulamis

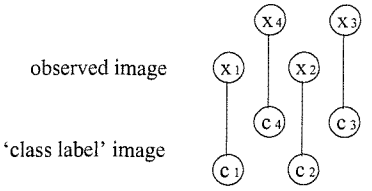


FIGURE 5.3: Image segmentation by supervised learning

et al. 2000). However, supervised techniques has not been widely used in medical image segmentation. One likely reason is that it is sometimes very expensive to obtain labelled examples. This may also be because the standard supervised learning techniques such as error back-propagation often assume the data are generated independently and provide no way to incorporate either unlabelled data or image models. Here a hierarchical method is proposed. Given labelled data $(\mathbf{x}^l, c^l)$, first a Bayesian multi-layer perceptron (MLP) network is trained with a regularized cost function according to the evidence framework (Mackay 1992b). As this neural network considers no spatial continuity, the output of the neural network was further modelled as a Markov random field, using Besag's *ad hoc* iterated conditional modes (ICM) (Besag 1986) method.

The image segmentation can be carried out by hierarchically training a supervised neural network and modelling the test image's spatial continuity as a Markov random field. This hierarchical fusion method provides a way to use labelled data in the image segmentation process. However, the two different steps of this fusion process are based on two contradictory assumptions: the setup of the neural network in the first stage assumes that the data are independent upon each other while the second step using a Markov random field to model data's inter-connections. This problem comes from the neural networks being based on the data-independence assumptions and provide no way to incorporate spatial dependence within the data.

5.5 Semi-supervised Image Segmentation - A New Combined Learning Framework

5.5.1 The Advantage of Using Labelled and Unlabelled Data in Image Segmentation

The importance of combining labelled data and unlabelled data in learning has been illustrated by an example in Section 3.5.1. In image segmentation, the combination of labelled and unlabelled samples is necessary to achieve ideal performance, as the images obtained at different times may be subject to various changes, and expert knowledge embedded in the labelled samples is critical in understanding the images. This section presents a combined framework for image segmentation.

5.5.2 Image at Three Levels: Observed Image; "Mixture Label" Image; "Class Label" Image

In the new framework, the image to be segmented is modelled at three different levels: the observed image; the "mixture label" image; the "class label" image, as shown in Fig. 5.4. Probability distributions are used to model the connections between different levels, enabling more accurate estimation to be made.

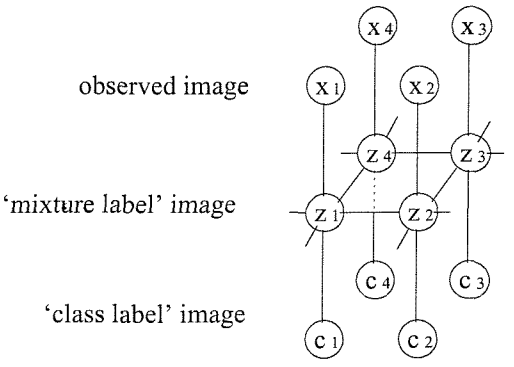


FIGURE 5.4: Image modelled at three levels

5.5.3 Incorporate Labelled Data In Image Segmentation

An image segmentation scheme is proposed here for the case that the image to be segmented $\mathbf{x}$ and labelled examples $(\mathbf{x}^l, c^l)$. We call this learning scheme "semisupervised" as it involves both labelled and unlabelled examples. As shown in Fig. 5.4, the image to be segmented is modelled at three different levels: the observed image $\mathbf{x}$, the mixture label image z and the class label image c. The model assumptions for $\mathbf{x}$ and z are the same as the unsupervised case introduced in Section 5.2: an MRF model (see Equation (5.5)) for z and the conditional independence assumption (Equation (5.12)) for $\mathbf{x}$. Additionally, the connections between different levels are modelled by a probability distribution determined from the labelled examples.

Labelled samples are incorporated into image segmentation by maximising

$$\begin{aligned} \log L &= \log p(\mathbf{x}, z) + \log p(\mathbf{x}^l, c^l) \\ &= \log p(\mathbf{x}, z) + \sum_{i=1}^{n^l} \log p(\mathbf{x}_i^l, c_i^l). \end{aligned} \tag{5.37}$$

This expression differs from both the unsupervised segmentation function (Equation (5.10)) and the supervised segmentation function. The framework integrates three parts of information from different sources: the image data and their spatial continuity embedded in $p(\mathbf{x}, z)$; further knowledge in the labelled samples given in $p(\mathbf{x}^l, c^l)$. The last

term in Equation (5.37) can be expanded as

$$p(\mathbf{x}_i^l, c_i^l) = \sum_{k=1}^{K} p(c_i^l|\mathbf{x}_i^l, z_i^l = k)p(\mathbf{x}_i^l|z_i^l = k)p(z_i^l = k). \tag{5.38}$$

The labelled examples enable the establishment of a probabilistic distribution $p(c_i = j|z_i = k)$ to describe the connection between the "mixture label" z_i and the "class label" c_i. Here the "generalized mixture" (GM) model proposed in (Miller and Uyar 1996) is used:

$$\gamma_{j|k} \equiv p(c_i = j|z_i = k) = \frac{\sum\limits_{\mathbf{x}_i^l, c_i^l = j} p(z_i^l = k|\mathbf{x}_i^l, c_i^l)}{\sum\limits_{\mathbf{x}_i^l} p(z_i^l = k|\mathbf{x}_i^l, c_i^l)}. \tag{5.39}$$

($\gamma_{j|k}$ is used as $p(c_i = j|z_i = k)$ is independent of the voxel i). Finally, the class membership of each data is decided by

$$p(c_i = j|\mathbf{x}_i) = \sum_{k=1}^{K} \gamma_{j|k} p(z_i = k|\mathbf{x}_i). \tag{5.40}$$

5.5.4 Model the Image's Spatial Distribution in "Mixture Label" Image

The modelling of spatial continuity within images is important as the contextual information is vital in understanding images. A Markov random field model is used in the second level of the three level framework, i.e. "mixture label" image z is modelled as a Markov random field, i.e., $p(\mathbf{z}) > 0$ and $p(z_i|z_{S-\{i\}}) = p(z_i|z_{\mathcal{N}_i})$.

Thus $p(z)$ obeys Gibbs distribution. To incorporate neighbourhood interactions into the model, a pseudo-likelihood approximate technique proposed in (Zhang et al. 1994) will be used to approximate the mixture label's prior

$$p(z_i = k) \approx p(z_i = k|\widehat{z}_m, m \in N_i) = \frac{e^{\beta\delta_i(k)}}{\sum\limits_{k=1}^{K} e^{\beta\delta_i(k)}}, \tag{5.41}$$

where $\delta_i(k)$ is the number of neighbours of i whose mixture label is k and $\beta > 0$ is a parameter controlling the influence of neighbouring voxels. This approximation technique is chosen for computational convenience.

5.5.5 Optimisation

An algorithm is needed to solve the optimisation problem in Equation (5.37). It is noted that the EM algorithm is an efficient method to solve an optimisation problem where

there is hidden data. For supervised learning, a gradient-based method can be used for optimisation. In recent years some EM-based optimisation algorithms have been proposed. (Jordan and Jacobs 1994) uses an EM algorithm to maximise conditional likelihood in a mixture of experts framework. The mathematical connections between the EM algorithm and the gradient-based approaches for maximum likelihood learning of finite Gaussian mixtures were developed in (Xu and Jordan 1996). These make the EM algorithm a natural method for solving optimisation problems that involve both unlabelled and labelled data.

Starting with an initial estimate of model parameters $\widehat{\Phi}^{(0)}$, the algorithm iterates:

- E-step: Estimate
  $Q(\Phi|\widehat{\Phi}^{(t)}) = E[\log p(\mathbf{x}, z) + \log p(\mathbf{x}^l, c^l)|\mathbf{x}, \widehat{\Phi}^{(t)}]$
- M-step: Find $\widehat{\Phi}^{(t+1)} = \arg\max_{\Phi} Q(\Phi|\widehat{\Phi}^{(t)})$.

Here t represents the tth iteration. When the unlabelled data are independent of each other, the mixture label's prior is updated by,

$$p(z_i^{(t+1)} = k) = \frac{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i) + \sum_{i=1}^{n^l} p(z_i^{l(t)} = k|\mathbf{x}_i^l, c_i^l)}{n + n^l}. \tag{5.42}$$

When using a Gaussian distribution to model the conditional distribution $p(\mathbf{x}_i|z_i = k, \Phi)$,

$$p(\mathbf{x}_i|z_i = k, \Phi) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right\}}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_k|^{1/2}}, \tag{5.43}$$

the mean vector and the covariance matrix of component k, $\phi_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_\mathbf{k})$, are re-estimated using:

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)\mathbf{x}_i + \sum_{i=1}^{n^l} p(z_i^{(t)} = k|\mathbf{x}_i^l, c_i^l)\mathbf{x}_i^l}{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i) + \sum_{i=1}^{n^l} p(z_i^{(t)} = k|\mathbf{x}_i^l, c_i^l)}, \tag{5.44}$$

$$\left(\boldsymbol{\Sigma}_k^2\right)^{(t+1)} = \frac{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i)S_{ik} + \sum_{i=1}^{n^l} p(z_i^{(t)} = k|\mathbf{x}_i^l, c_i^l)S_{ik}}{\sum_{i=1}^{n} p(z_i^{(t)} = k|\mathbf{x}_i) + \sum_{i=1}^{n^l} p(z_i^{(t)} = k|\mathbf{x}_i^l, c_i^l)}, \tag{5.45}$$

where $S_{ik} \equiv (\mathbf{x}_i - \boldsymbol{\mu}_k^{(t)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t)})^T$,

$$p(z_i^{(t)} = k|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|z_i^{(t)} = k)P(z_i^{(t)} = k)}{\sum_{k=1}^{K} p(\mathbf{x}_i|z_i^{(t)} = k)P(z_i^{(t)} = k)} \tag{5.46}$$

and

$$p(z_i^{(t)} = k|\mathbf{x}_i^l, c_i^l) = \frac{p(\mathbf{x}_i|z_i^{(t)} = k)\gamma_{j|k}^{(t)} P(z_i^{(t)} = k)}{\sum_{k=1}^{K} p(\mathbf{x}_i|z_i^{(t)} = k)\gamma_{j|k}^{(t)} P(z_i^{(t)} = k)} \tag{5.47}$$

with $\gamma_{j|k}^{(t+1)}$ updated by Equation (5.39). This EM re-estimation continues until the updates fall below a specified threshold or the maximum number of iteration reaches. Finally, the class label is determined using Equation (5.40).

5.5.6 Summary

A new semi-supervised image segmentation framework is proposed with the view to enabling the labelled examples and the image to be considered in the learning process. Thus the generalisation performance can be improved. The learning scheme proposed here is a three-level image model with probability distributions describing the connection. The image voxel's connection is modelled as a Markov random field model in the second level. An EM based optimisation algorithm is also proposed to solve this difficult combined learning problem.

The proposed approach is superior in that:

- Both the labelled data and unlabelled data are used in the learning phase, so that the classifier uses knowledge in the labelled samples as well as additional knowledge of the data distribution from the unlabelled data, which is very important when labelled data are sparse.
- Each image is modelled at three different levels: the observed image, the "mixture label" image and the "class label" image, where connections between different levels being described by probability distributions, enabling a posterior probability distribution of the data to be recovered.
- A Markov random field model is used to incorporate neighbourhood interaction into the learning process.
- All information is considered within an integrated framework instead of in a hierarchical way.

5.6 Conclusions

Low-level image segmentation is a fundamental and yet difficult task in machine vision. Markov random field (MRF) models are one of the widely used model-based methods for image segmentation. The main theory of Markov random field model-based image

segmentation has been presented in this chapter. The local interactions between image voxels, Markov random fields models are introduced. The Markov random field model enables consideration of contextual dependence which is indispensable in image analysis. By modelling interactions between image voxels, regularisation in image segmentation is realised, thus maximum a posterior (MAP) image segmentation can be achieved. However, the inclusion of MRF models makes the optimisation process difficult. The computationally expensive MCMC method can be used with the potential to find the global minimum. Further more if the cluster number is unknown, reversible jump MCMC can be used. If the model parameters are known, ICM is an efficient approximation optimisation method to find a local solution. Mean field annealing is a method to find the minimum variance approximation. Both ICM and mean field annealing can to used in Expectation-Maximisation algorithm to do joint model parameter and image segmentation.

As Markov random field models express global relationships in terms of local statistics, with the expense of heavy computation burden. Multi-resolution techniques attempt to circumvent this problem by providing a mechanism by which longer range interactions can quickly propagate the image model.

Following the introduction of general image segmentation issues and the unsupervised image segmentation techniques, a new hierarchical supervised method for image segmentation was proposed in Section 5.4 and a new combined image segmentation framework and the optimisation algorithm in Section 5.5. It can be seen that the new combined image segmentation framework has an integrated learning process without any contradictory assumptions as exists in the hierarchical supervised image segmentation process.

As the combination of labelled and unlabelled data are a developing research field in machine learning, there may be other ways to combine labelled samples in the image segmentation as well. However, the combined framework proposed in this chapter provides a natural way to incorporate the Markov random field data model into the combined learning process. The use of this combined image segmentation scheme and other unsupervised and supervised segmentation schemes will be compared in the next chapter.

Chapter 6

Spatio-temporal PET Reference Region Extraction

The methods presented in chapter 5 enable labelled data, image models and unlabelled data to be used in the image segmentation process. In this chapter, these methods will be used in simulated and real PET images for reference region extraction. This is referred as spatio-temporal PET reference region extraction as both spatial information (voxel location) and the temporal information (time activity curves) are used.

6.1 Simulation Studies

6.1.1 Description of Experiments

The synthetic data used here is the same as described in Section 4.5.1. Three different methods are compared to extract the region with $BP = 0$ from the regions with $BP = 1, 2, 3$:

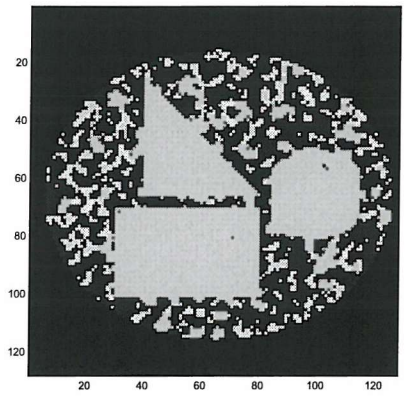
- Unsupervised image segmentation using EM with a pseudo-likelihood prior approximation in Section 5.3.1.1;
- Supervised segmentation by hierarchically using trained neural networks and ICM as described in Section 5.4;
- Semi-supervised method proposed in Section 5.5.

A 3×3 pixel neighbourhood is used in the Gaussian Markov random field model, β in Equation (5.41) is set to 1. In supervised and semi-supervised algorithms, the final classification is carried out automatically by assigning every pixel to the class that it has the highest posterior probability. In the unsupervised segmentation, the cluster

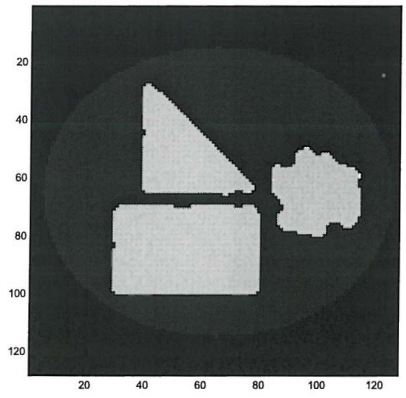
number is set to 4. As there is no information about cluster identification available, the classification is carried out by choosing the cluster whose centre has the lowest BP value as the reference class, while the rest are treated as the other class.

6.1.2 Segmentation Results

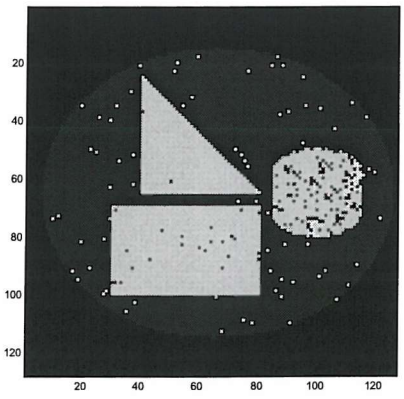
The image segmentation results using these three methods, with 500 labelled data and noise standard deviation $\sigma = 2.1$, are displayed in Fig. 6.1. The unsupervised segmentation result (Fig. 6.1(a)) fails to localise the reference region, although the mixture number is set as 4, which is the real cluster number. The semi-supervised segmentation result in Fig. 6.1(c) gives the best classification accuracy as it captures the edges of different classes better than the supervised segmentation result in Fig. 6.1(b).



(a) Unsupervised segmentation



(b) Supervised segmentation



(c) Semi-supervised segmentation

FIGURE 6.1: Image segmentation results with 500 labelled examples, noise standard deviation $\sigma = 0.8$

To further compare the algorithms' performance, the influence of the number of labelled examples and the noise level in the data is examined. Although the unsupervised segmentation results are not affected by the change in the number of labelled examples, for

the sake of comparison, they are displayed in the same way as the other two methods. Fig. 6.2 shows the mean and standard deviation for the voxel classification accuracy, the non-reference region classification error and the BP value of the extracted TAC. The four rows correspond to the labelled examples's number 100, 200, 500 and 1000 respectively. The error bars are generated by running every method for ten times with different initial values for model parameters. The number of clusters is set to 4 in both the unsupervised and semi-supervised classification. The three types of error are defined as follows:

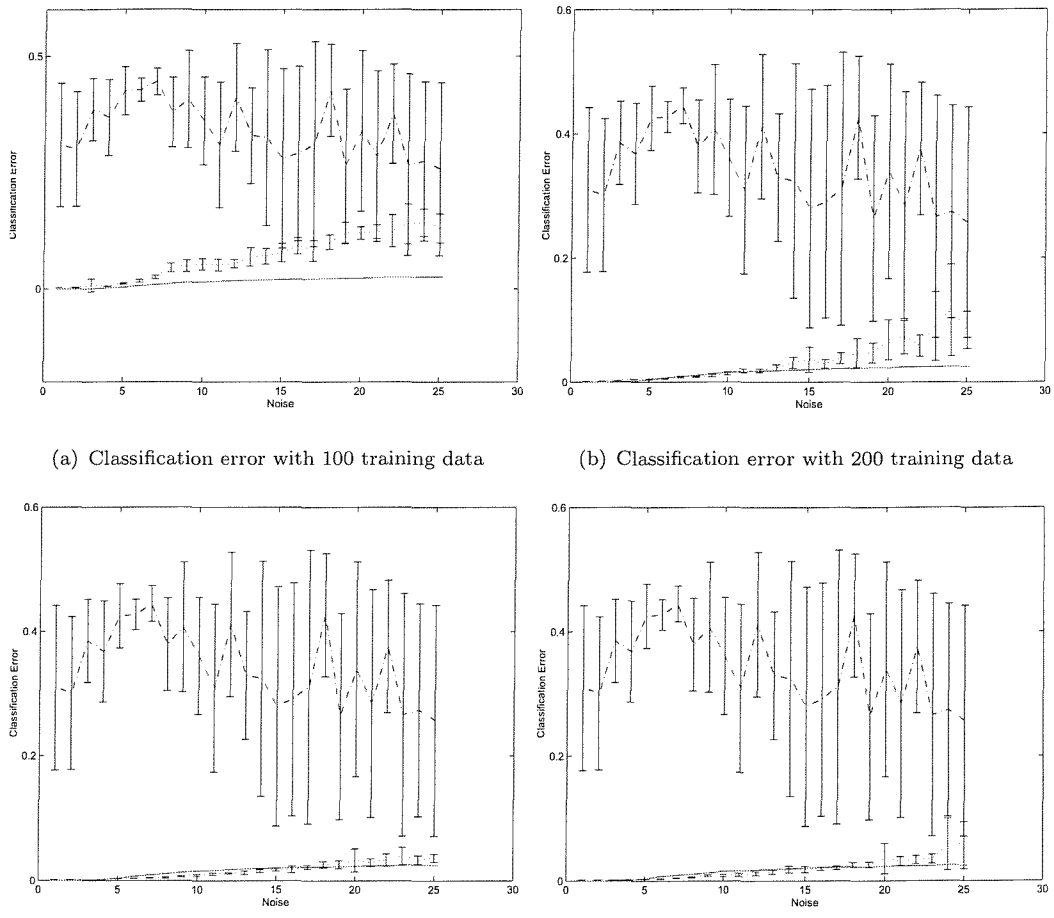
- The total classification error is calculated by $\frac{Number\ of\ misclassified\ voxels}{Total\ number\ of\ voxels} * 100\%$;
- The non-reference region classification error is $\frac{Number\ of\ misclassified\ non-reference\ region\ voxels}{Total\ number\ of\ non-reference\ region\ voxels} * 100\%$;
- The BP value of the extracted reference TAC is calculated using the simplified reference region model (Equation (2.3)) and the basis function method (Equation (2.6),(2.7),(2.8)).

Each sub-figure shows the change of one of these three types of error with a different noise level. The noise level runs from 0.2 to 2.6.

Fig. 6.5 shows the results with the number of clusters in the unsupervised and semi-supervised classification changed to 10.

Fig. 6.2 and Fig. 6.5 shows the total classification error. In both figures, the semi-supervised segmentation achieves best classification performance while the unsupervised segmentation result gives the worst classification error. As each optimisation process involved in this simulation can only find a local minimum, the classification results vary with different initial values for model parameters. With smallest standard deviation, the semi-supervised method shows its robustness with different starting points. Unsupervised classification performance is also unstable as it gives a highest standard deviation. Similar to the independent data case in Chapter 3, when the noise level increases, the classification error for supervised and semi-supervised classification increases as expected while the unsupervised classification error decreases slightly in the four-cluster case. One possible explanation is that when the noise level is low, the algorithm tends to split the true reference region data into more than one clusters, as only one cluster will be picked, the total classification error is large. When the noise level increases, the chosen cluster contains more true reference region data. This downtrend is less severe when the cluster number increases to 10, as shown in Fig. 6.5.

Fig. 6.3 and Fig. 6.6 give the non-reference region mis-classification error. The unsupervised classification achieves best accuracy. The result for the 10-cluster case achieves more accuracy than the 4-cluster case. However, semi-supervised classification has the lowest standard deviation, showing that its performance is very stable.



(a) Classification error with 100 training data

(b) Classification error with 200 training data

(c) Classification error with 500 training data

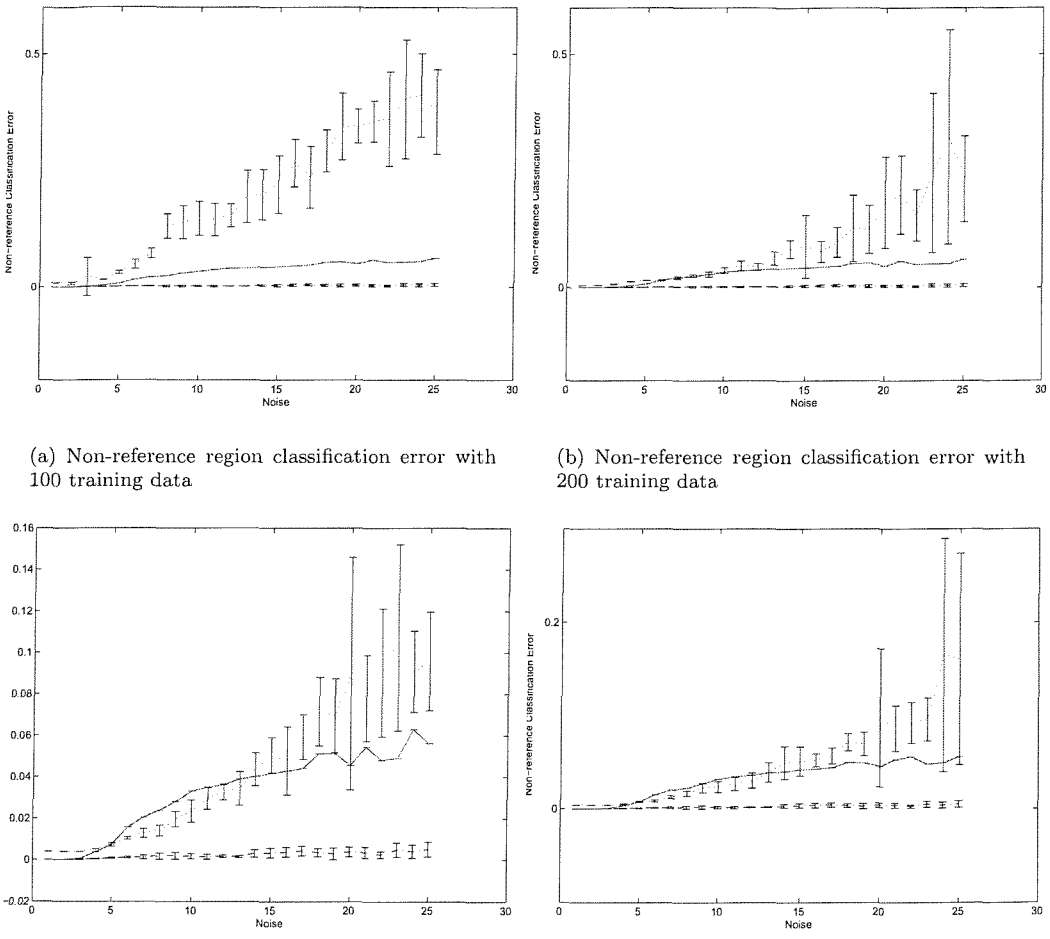
(d) Classification error with 1000 training data

FIGURE 6.2: Simulation results on classification accuracy, with 4 clusters in unsupervised classification (Red solid : Error bar for semi-supervised classification; Blue dotted: Error bar for supervised classification; Black dashdot: Error bar for unsupervised classification.)

Fig. 6.4 and Fig. 6.7 show the binding potential error of the classified reference region voxels. Minimising the binding potential error is very important in PET segmentation as binding potential describes the characteristics of PET images. The semi-supervised method achieves the best performance with excellent stability reflected by the low standard deviation. The unsupervised method also achieves competitive performance but it is less stable.

6.1.3 Comparison of the Results Based on Markov Random Field Models and the Independence Assumptions

Fig. 6.8, 6.9, 6.10 give the differences between independent assumption results and MRF-model assumption results, for all three learning methods with 4 clusters in unsupervised



(a) Non-reference region classification error with 100 training data

(b) Non-reference region classification error with 200 training data

(c) Non-reference region classification error with 500 training data

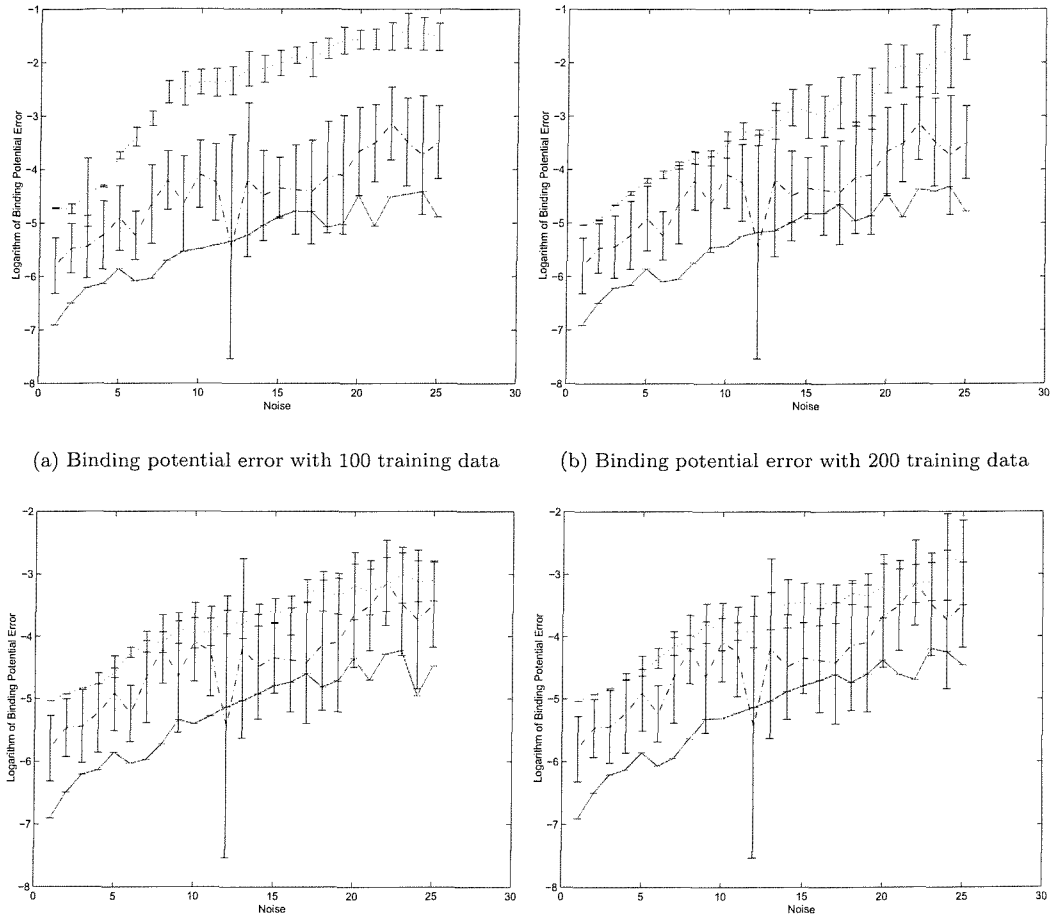
(d) Non-reference region classification error with 1000 training data

FIGURE 6.3: Simulation results on non-reference region classification accuracy, with 4 clusters in unsupervised classification (Red solid : Error bar for semi-supervised classification; Blue dotted: Error bar for supervised classification; Black dashdot: Error bar for unsupervised classification.)

and semi-supervised segmentation.

Fig. 6.8 shows the difference of the total classification error, generated by the mean value of each variable from error bars in Fig. 4.7 minus the corresponding mean value from error bars in Fig. 6.2. Thus positive value in the figures indicates the MRF model-based segmentation improves the image segmentation accuracy compared to the independent segmentation. For supervised segmentation, the MRF model-based segmentation improves the classification accuracy significantly. No obvious improvement or deterioration for the semi-supervised segmentation while the unsupervised segmentation with a MRF model has more classification error.

Fig. 6.9 shows the difference of the non-reference region classification error, generated by the mean value of each variable from error bars in Fig. 4.8 minus the corresponding



(a) Binding potential error with 100 training data

(b) Binding potential error with 200 training data

(c) Binding potential error with 500 training data

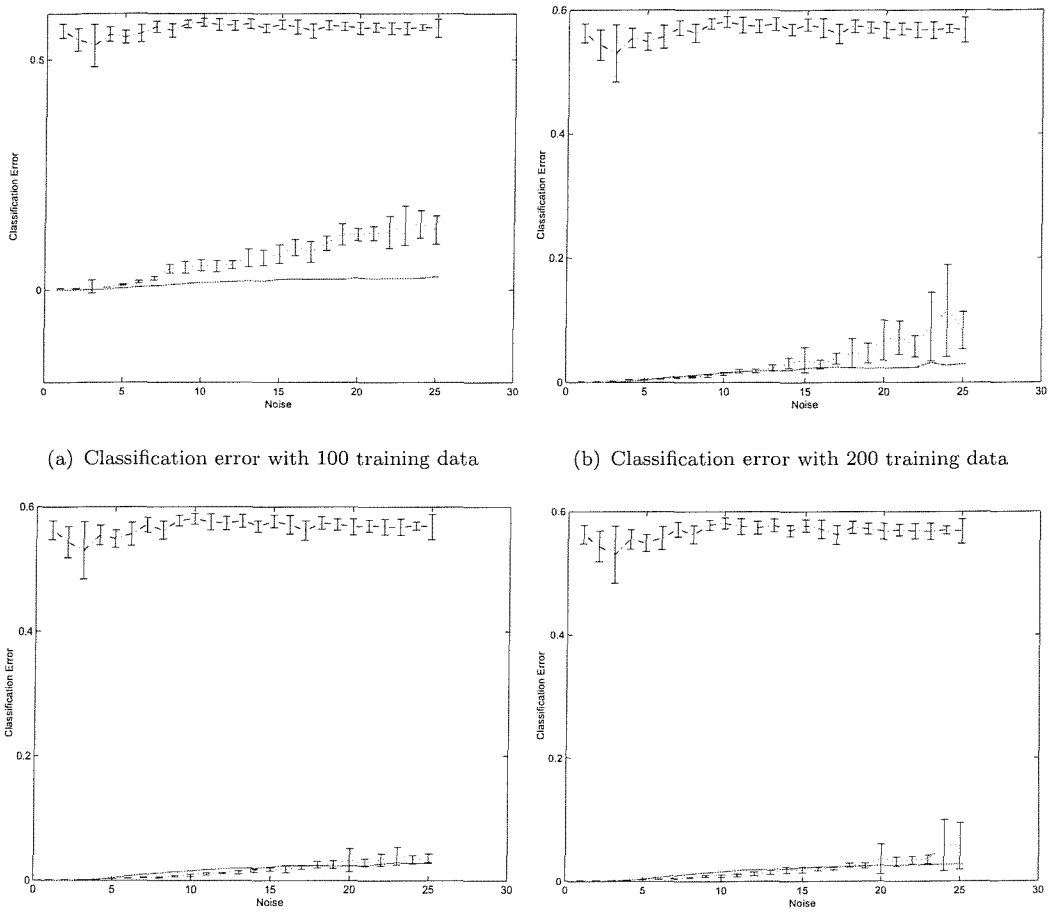
(d) Binding potential error with 1000 training data

FIGURE 6.4: Simulation results on binding potential accuracy, with 4 clusters in unsupervised classification (Red solid : Error bar for semi-supervised classification; Blue dotted: Error bar for supervised classification; Black dashdot: Error bar for unsupervised classification.)

mean value from error bars in Fig. 6.3. Both supervised and unsupervised segmentation using a MRF model has improved non-reference region classification error, while the performance is slightly worse in the semi-supervised case.

Fig. 6.10 shows the difference of the BP value of the extracted reference TAC, generated by the mean value of each variable from error bars in Fig. 4.9 minus the corresponding mean value from error bars in Fig. 6.4. The BP is the main interest. The semi-supervised segmentation using a MRF model achieves improved results compared to the independent case. The BP value difference for both supervised and unsupervised segmentation are not very stable, although better BP accuracy is obtained in the MRF model-based supervised segmentation with high noise level in the 500 and 1000 training data case.

Fig. 6.11, 6.12, 6.13 show the corresponding results with 10 clusters in unsupervised and



(a) Classification error with 100 training data

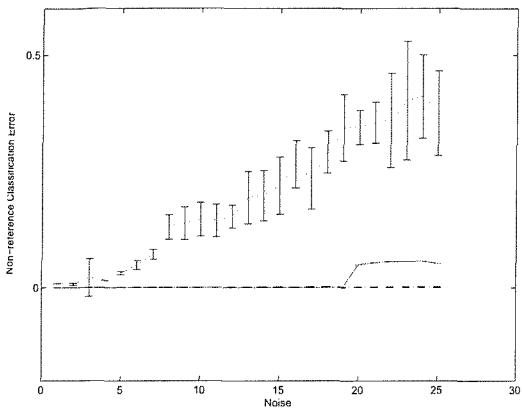
(b) Classification error with 200 training data

(c) Classification error with 500 training data

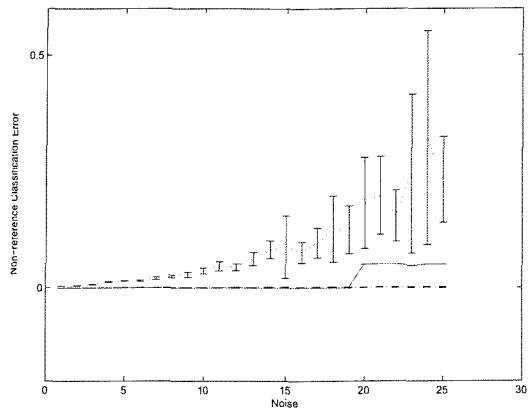
(d) Classification error with 1000 training data

FIGURE 6.5: Simulation results on classification accuracy, with 10 clusters in unsupervised classification (Red solid : Error bar for semi-supervised classification; Blue dotted: Error bar for supervised classification; Black dashdot: Error bar for unsupervised classification.)

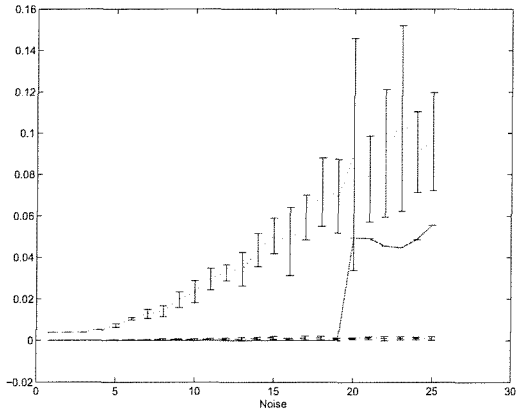
semi-supervised segmentation. Similarly, Fig. 6.11 is generated by the mean value of each variable from error bars in Fig. 4.10 minus the corresponding mean value from error bars in Fig. 6.5. Fig. 6.12 is generated by the mean value of each variable from error bars in Fig. 4.11 minus the corresponding mean value from error bars in Fig. 6.6. Fig. 6.13 is generated by the mean value of each variable from error bars in Fig. 4.12 minus the corresponding mean value from error bars in Fig. 6.4. The performance difference in the 10-cluster segmentation case follows the similar pattern as the performance difference in the above 4-cluster segmentation case.



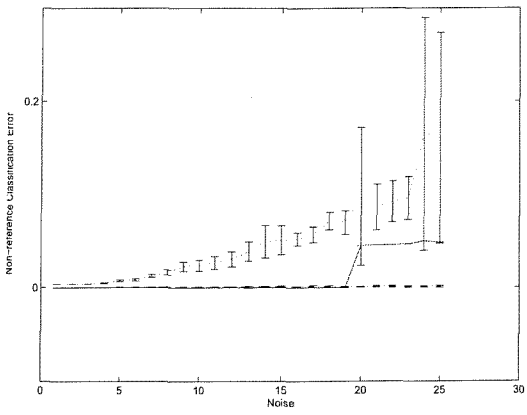
(a) Non-reference region classification error with 100 training data



(b) Non-reference region classification error with 200 training data



(c) Non-reference region classification error with 500 training data



(d) Non-reference region classification error with 1000 training data

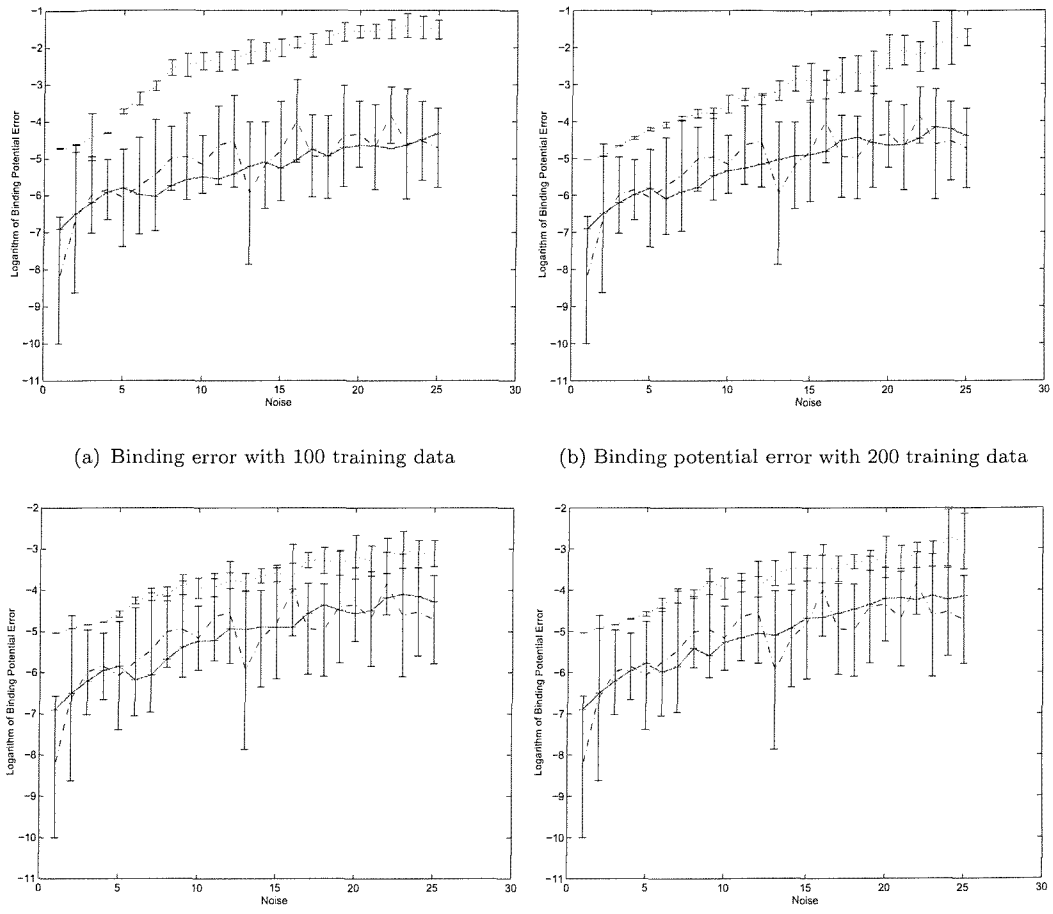
FIGURE 6.6: Simulation results on non-reference region classification accuracy, with 10 clusters in unsupervised classification (Red solid : Error bar for semi-supervised classification; Blue dotted: Error bar for supervised classification; Black dashdot: Error bar for unsupervised classification.)

6.2 PET Data Studies

6.2.1 Experiment Description

The developed algorithms are also applied to eighteen $[^{11}C](R)$-PK11195 PET data, which has been used in the experiments in Chapter 4. Same pre-processing and data normalisation are used. As in the previous simulation, three different spatio-temporal segmentation methods are compared to extract the reference region from the binding region:

- Unsupervised image segmentation using EM with a pseudo-likelihood prior approximation (Equation (5.41));



(a) Binding error with 100 training data

(b) Binding potential error with 200 training data

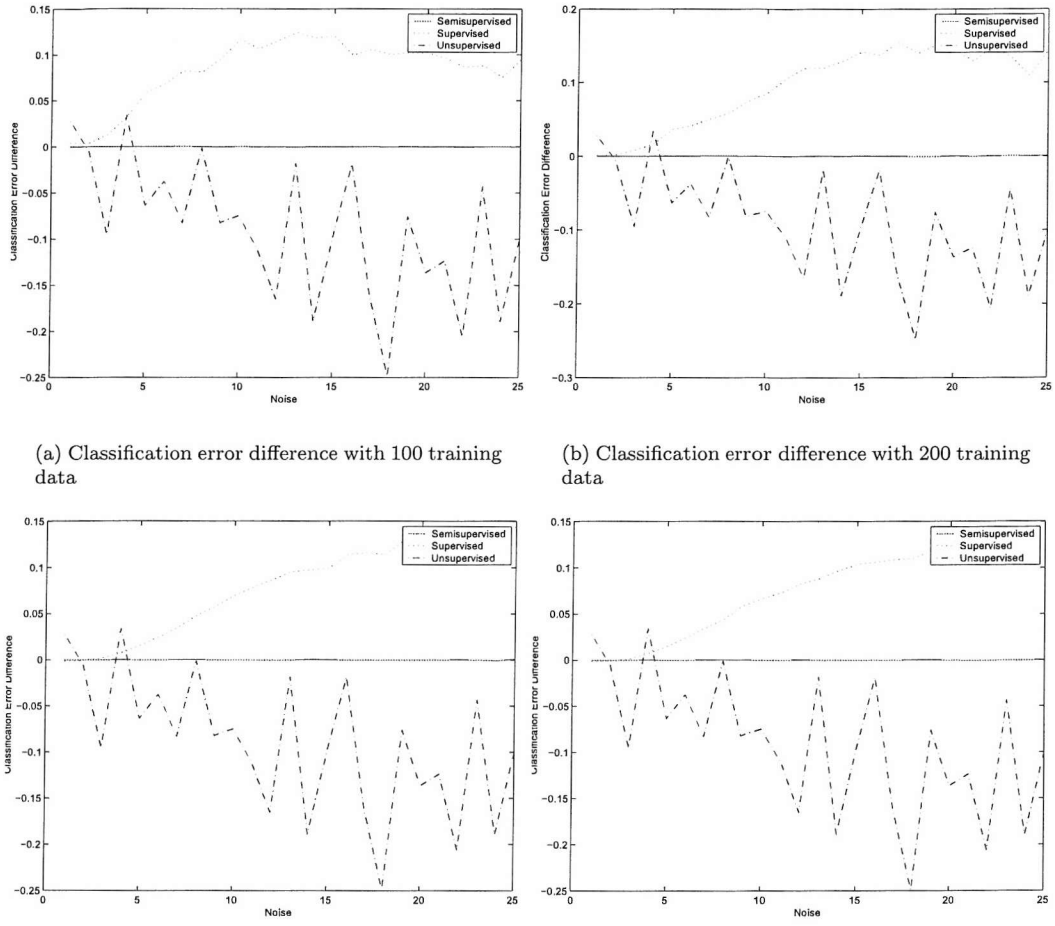
(c) Binding potential error with 500 training data

(d) Binding potential error with 1000 training data

FIGURE 6.7: Simulation results on binding potential accuracy, with 10 clusters in unsupervised classification (Red solid : Error bar for semi-supervised classification; Blue dotted: Error bar for supervised classification; Black dashdot: Error bar for unsupervised classification.)

- Supervised segmentation by hierarchically using trained neural networks and ICM as described in Section 5.4;
- Semi-supervised method proposed in Section 5.5.

In the unsupervised and semi-supervised segmentation, the cluster number is set as 10. In supervised classification, four hidden nodes were found to be suitable for this segmentation problem. In supervised and semi-supervised algorithms, the final classification is carried out automatically by assigning every voxel to the class that has the highest posterior probability. In unsupervised segmentation, the final classification is carried out by choosing the cluster whose centre has lowest BP as the reference class, while the rest are treated as the other class.



(a) Classification error difference with 100 training data

(b) Classification error difference with 200 training data

(c) Classification error difference with 500 training data

(d) Classification error difference with 1000 training data

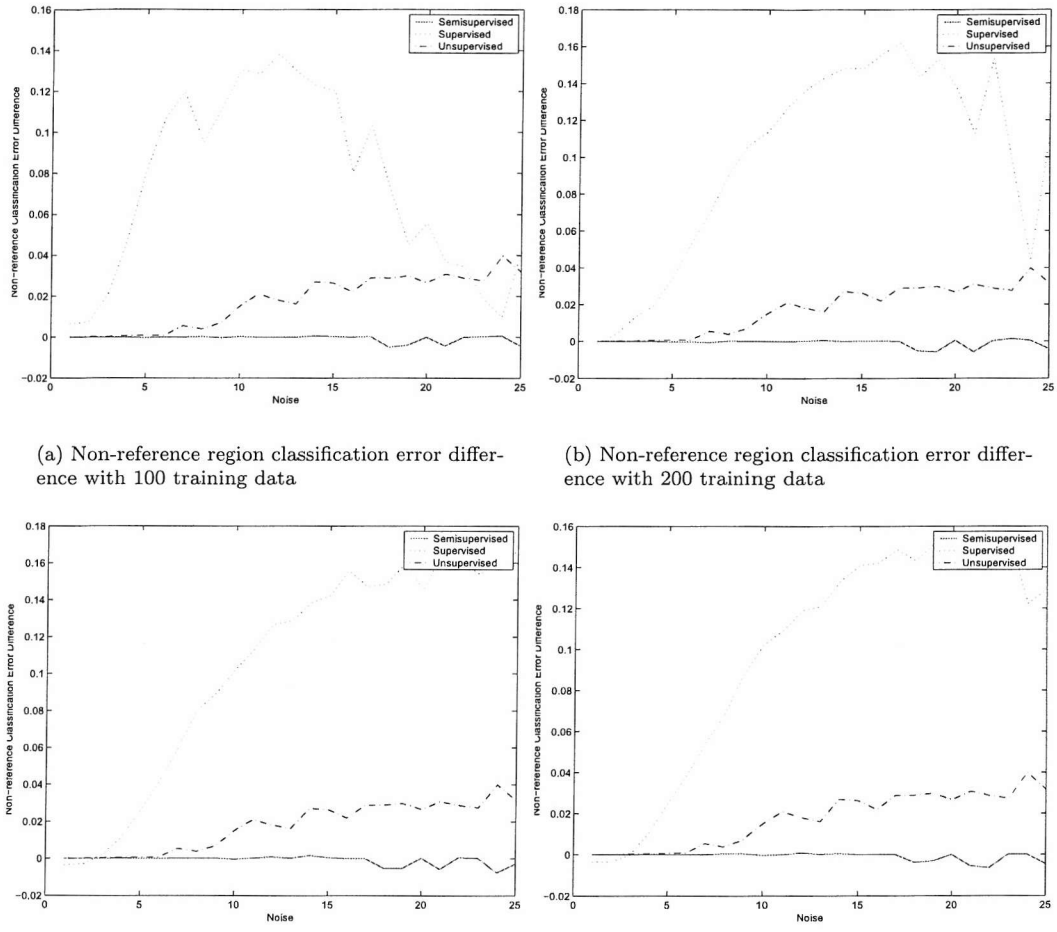
FIGURE 6.8: Simulation results difference between temporal and spatial-temporal modelling on classification accuracy, with 4 clusters in unsupervised and semi-supervised classification

The labelled data are needed for supervised and semi-supervised segmentation. The labelling data are the same as the experiments in Chapter 4, 2800 TACs from the cortex region and 2100 TACs from the scalp, thalamus and cerebellum regions were randomly sampled from seven scans as the labelled data. For comparison convenience, they are exactly the same as in Chapter 4. The same pre-processing and input normalisation is also used.

6.2.2 Results

6.2.2.1 Extracted Reference Region TAC

Fig. 6.14 shows the extracted reference TACs in the 7 scans where part of data are used in training supervised and semi-supervised classifiers. Three curves are shown in each



(a) Non-reference region classification error difference with 100 training data

(b) Non-reference region classification error difference with 200 training data

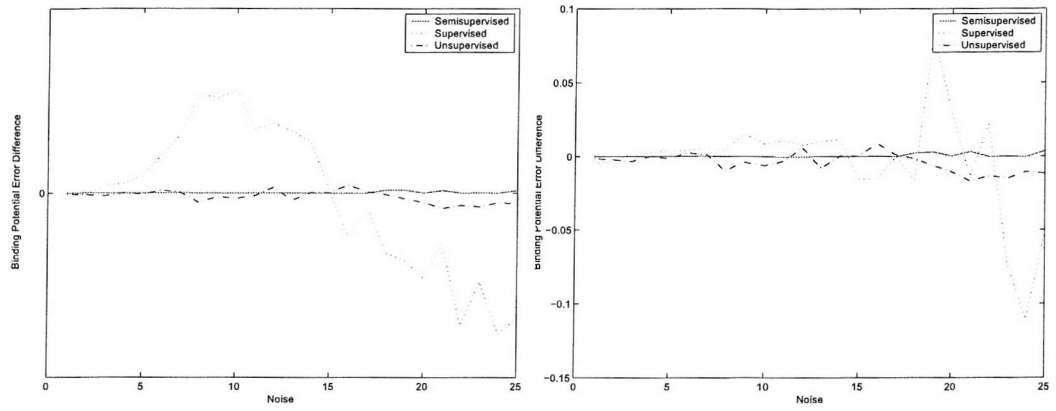
(c) Non-reference region classification error difference with 500 training data

(d) Non-reference region classification error difference with 1000 training data

FIGURE 6.9: Simulation results difference between temporal and spatial-temporal modelling on non-reference region classification accuracy, with 4 clusters in unsupervised and semi-supervised classification

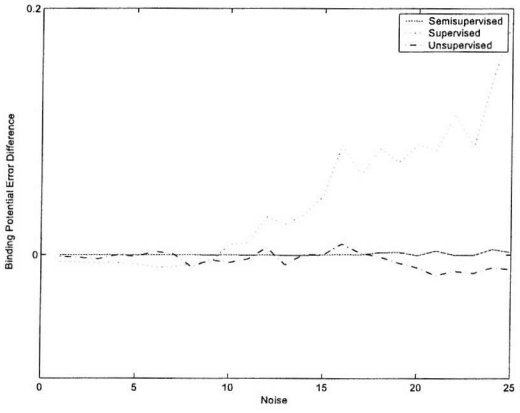
figure: the curve extracted from the semi-supervised classification; the curve extracted from the supervised neural network and the curve extracted from the unsupervised clustering. In unsupervised classification, the mean curve is obtained by choosing the largest cluster, which is the current method used at Hammersmith hospital.

The performance across PET scans is assessed by testing the performance in the other 11 independent PET scans, shown in Fig. 6.15. In both Fig. 6.14 and Fig. 6.15, the curves extracted from supervised and semi-supervised classification catch the shape of the reference region curves very well. However, the unsupervised classification performs inconsistently. It fails to capture the features of reference region curve in most scans. It fails to capture the features of reference region in Fig. 6.14(c) for scan n02816. It also fails to capture the shape and magnitude of the refernce region TAC in most scans with an exception of scan n03694 in Fig. 6.15(g).

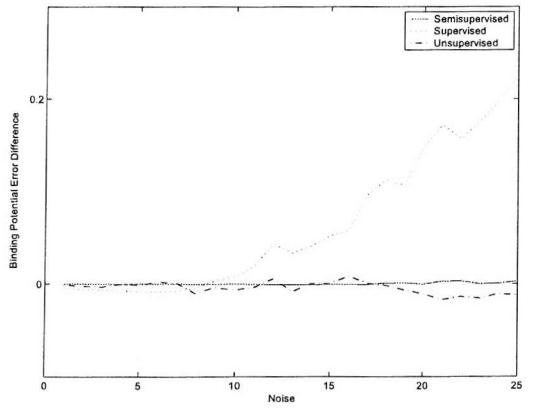


(a) Binding potential error difference with 100 training data

(b) Binding potential error difference with 200 training data



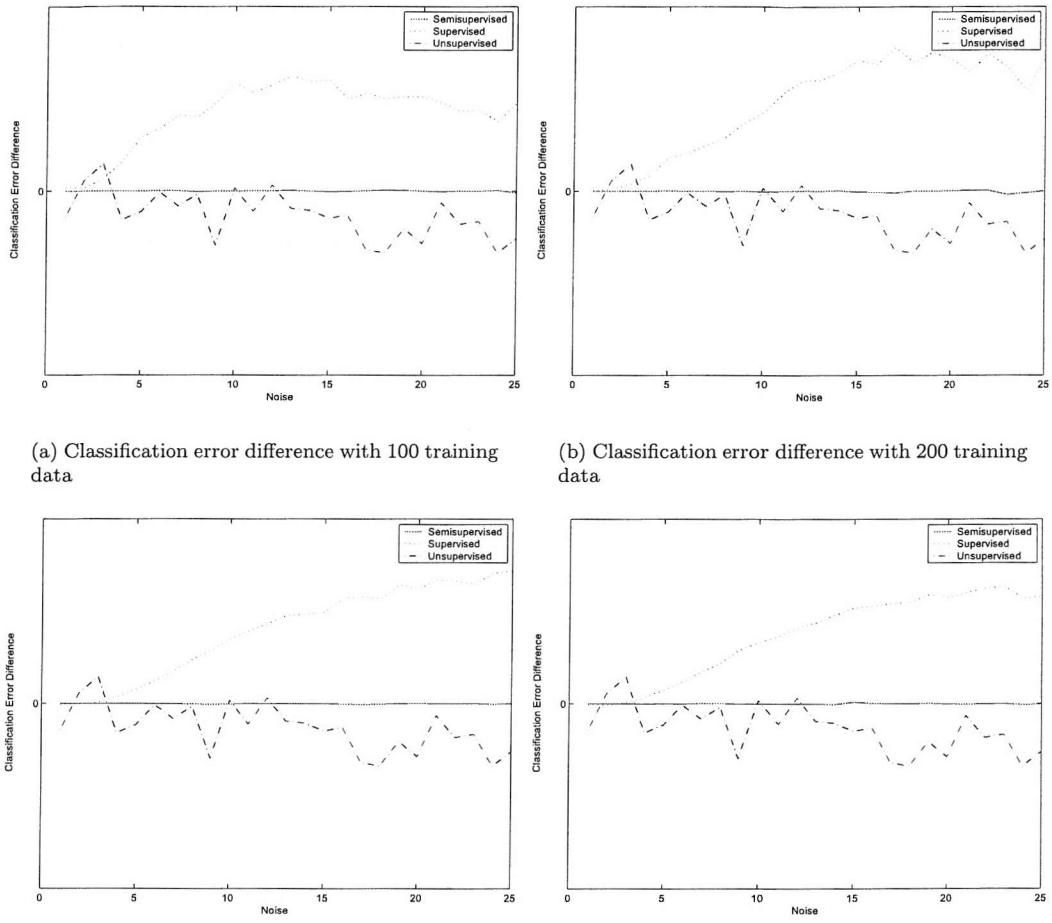
(c) Binding potential error difference with 500 training data



(d) Binding potential error difference with 1000 training data

FIGURE 6.10: Simulation results difference between temporal and spatial-temporal modelling on binding potential accuracy, with 4 clusters in unsupervised and semi-supervised classification

The supervised and semi-supervised classification have successfully learnt the information to discriminate the reference region from the other regions based on the shape of their TACs. Their performances are consistent across subjects. In the figure for each scan, the curve extracted from the semi-supervised method has lower tracer concentration value in the last several time instants, compared to the curve extracted from the supervised method, indicating the semi-supervised methods extracted reference region TACs have a lower binding potential value than the supervised method extracted reference region TACs.



(a) Classification error difference with 100 training data

(b) Classification error difference with 200 training data

(c) Classification error difference with 500 training data

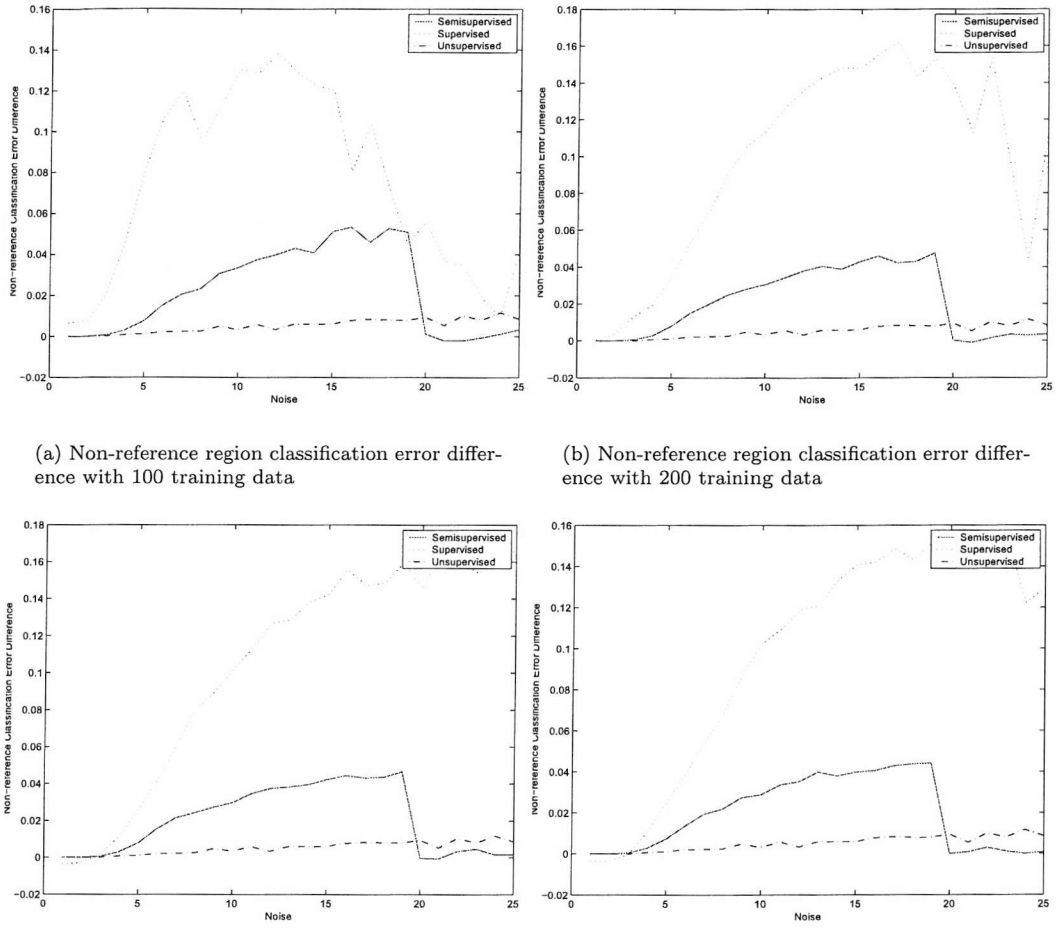
(d) Classification error difference with 1000 training data

FIGURE 6.11: Simulation results difference between temporal and spatial-temporal modelling on classification accuracy, with 10 clusters in unsupervised and semi-supervised classification

6.2.2.2 Parametric Images

Binding potential images can be generated by applying the learnt TAC to the reference region model. Fig. 6.16 shows an example of binding potential image generated for plane no.20 of scan n03578, a healthy subject. These three figures correspond to using reference region TAC generated from the unsupervised, supervised and semi-supervised Markov random field model based image segmentation methods respectively. They are generated in the same way as in Section 4.6.3. First, voxels outside the scalp with low TACs are filter out to avoid unnecessary computation, using:

$$\text{if } \sum_{j=1}^{18} x^j > 0.5 \sum_{j=1}^{18} x_r^j, \text{ calculate } BP \text{ value;}$$
$$\text{if } \sum_{j=1}^{18} x^j \leq 0.5 \sum_{j=1}^{18} x_r^j, \; BP = 0.$$



(a) Non-reference region classification error difference with 100 training data

(b) Non-reference region classification error difference with 200 training data

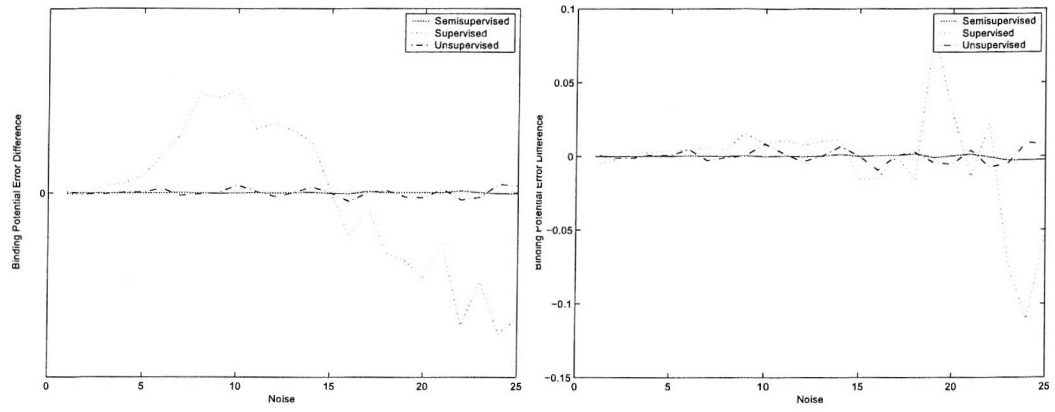
(c) Non-reference region classification error difference with 500 training data

(d) Non-reference region classification error difference with 1000 training data

FIGURE 6.12: Simulation results difference between temporal and spatial-temporal modelling on non-reference region classification accuracy, with 10 clusters in unsupervised and semi-supervised classification

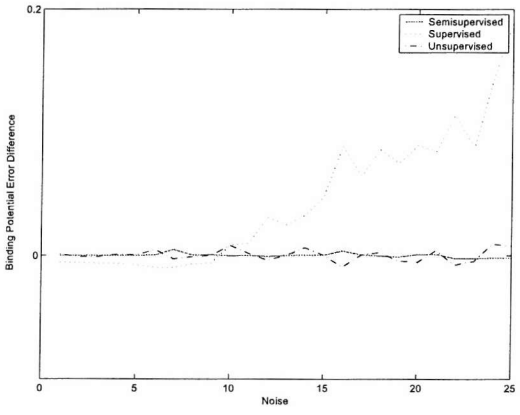
The calculation of BP value is realised by applying the simplified reference region model (Equation (2.3)) and the basis function method.

The binding potential images for plane no. 20 of a patient scan n02904 using the reference region model and the extracted reference region TAC for the three different methods are shown in Fig. 6.17. The binding potential images are generated in the same procedure as for the above scan n03578. The reference region TAC extracted by the unsupervised segmentation has a large error and fails to extract the reference region TAC. The binding potential image using the unsupervised extracted reference TAC is dramatically different from the others, with the binding potential value in the range $[-15, 45]$. From the binding potential image, it can be seen that the scan contains a strong binding area around the middle-right region.

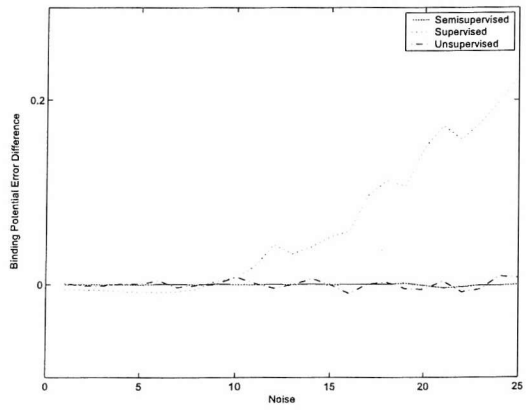


(a) Binding potential error difference with 100 training data

(b) Binding potential error difference with 200 training data



(c) Binding potential error difference with 500 training data



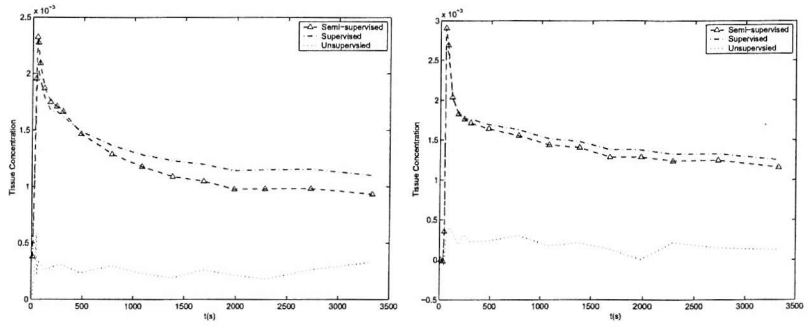
(d) Binding potential error difference with 1000 training data

FIGURE 6.13: Simulation results difference between temporal and spatial-temporal modelling on binding potential accuracy, with 10 clusters in unsupervised and semi-supervised classification

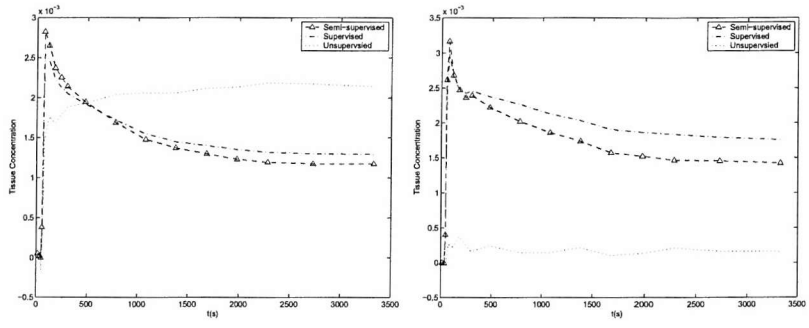
6.2.2.3 Improved Test

The lack of ground truth for performance evaluation in real PET data experiment, the test-retest scheme are used for performance evaluation. The four test-retest scan pair from four healthy subjects (as shown in Table 4.2) are used here. The difference between the segmentation results for each scan pair is measured and used as a criterion for evaluating the performance of the segmentation method. The test-retest experiment for the three reference region extraction techniques is illustrated in Fig. 6.18.

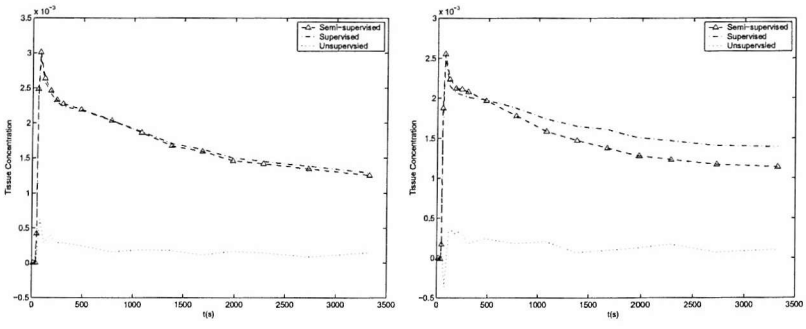
The test-retest results are shown in Fig. 6.19. Each sub-figure shows the test-retest difference between a scan pair, where the binding potential differences for thalamus, cerebellum and cortex are displayed. The binding potential of a region is calculated by using the extracted reference region TAC and the simplified reference region model.



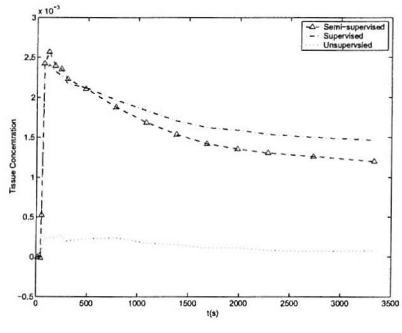
(a) n02791 (b) n02805



(c) n02816 (d) n02833

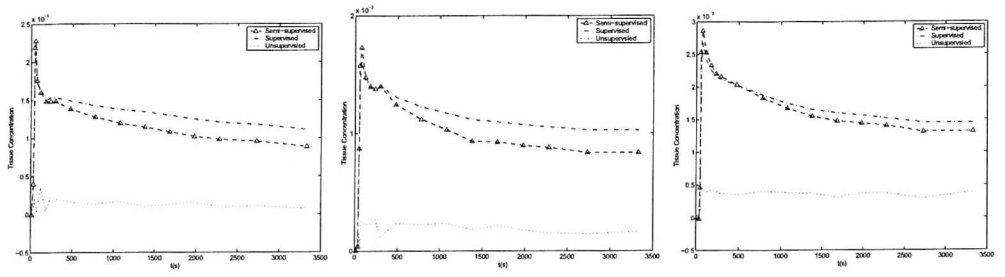


(e) n02870 (f) n03637

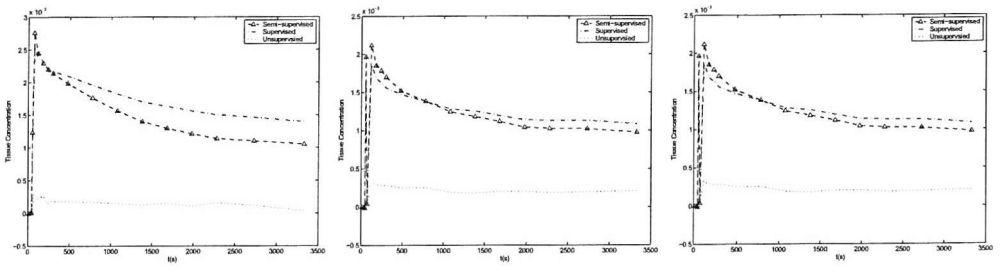


(g) n03657

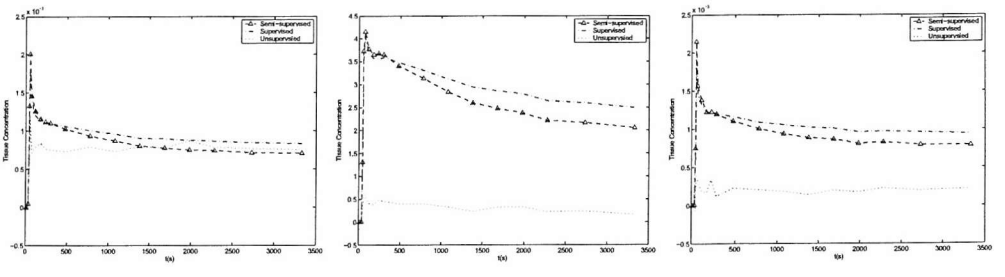
FIGURE 6.14: Results in seven different planes (Each figure shows the mean TACs from the reference region extracted in three different ways.)



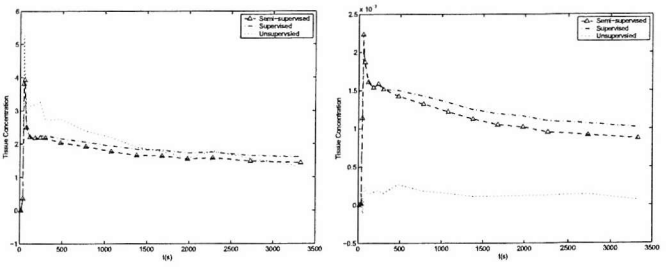
(a) n02907 (b) n02938 (c) n03578



(d) n03642 (e) n03661 (f) n03689

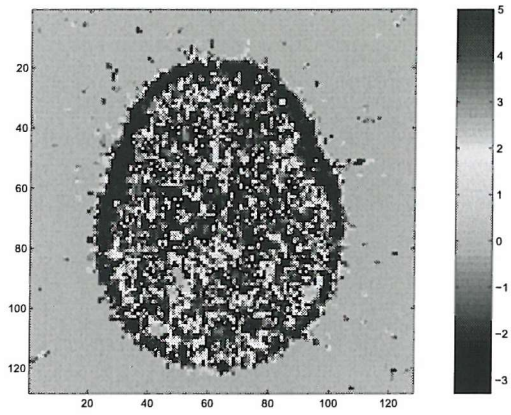


(g) n03694 (h) n04071 (i) n04073

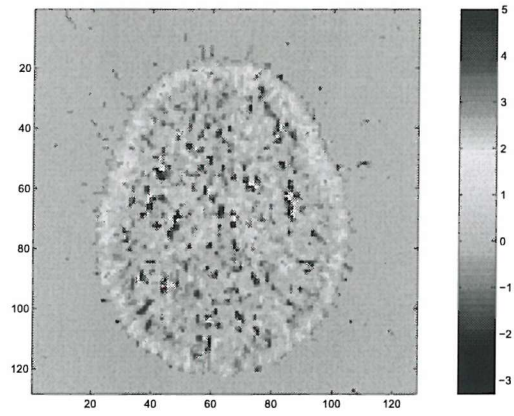


(j) n04128 (k) n02904

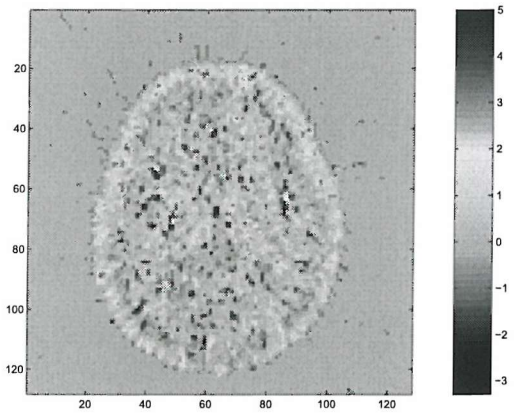
FIGURE 6.15: Results in ten planes from independent scans (Each figure shows the mean TACs from the reference region extracted in three different ways.)



(a) BP image using unsupervised extracted reference region

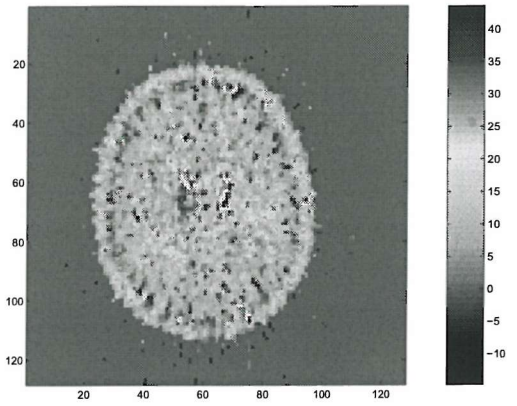


(b) BP image using supervised neural network extracted reference region

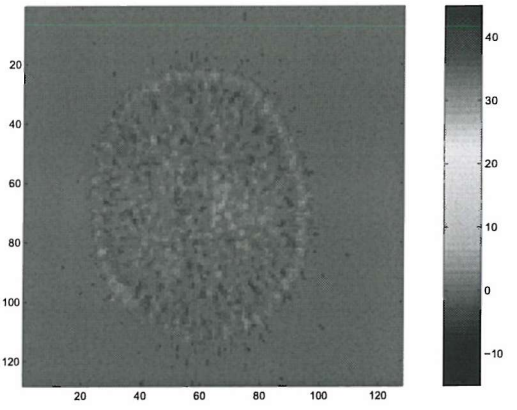


(c) BP image using semi-supervised extracted reference region

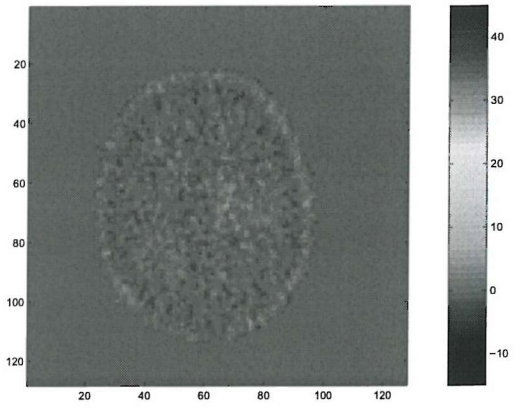
FIGURE 6.16: Parametric image of binding potential (BP) for a healthy subject's PET scan n03578 plane 20



(a) BP image using unsupervised extracted reference region



(b) BP image using supervised neural network extracted reference region



(c) BP image using semi-supervised extracted reference region

FIGURE 6.17: Parametric image of binding potential (BP) for patient PET scan n02904 plane 20

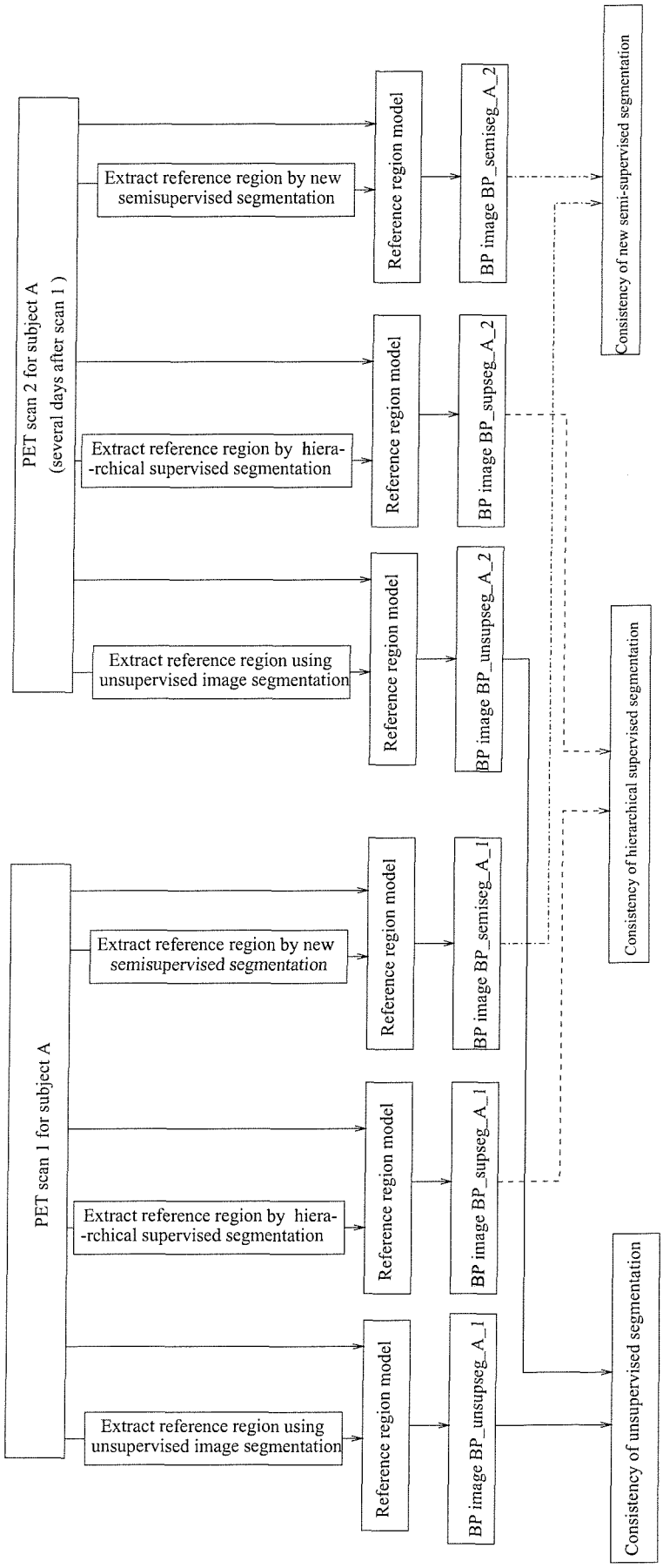


FIGURE 6.18: Improved test scheme

Three bars for the same region correspond to three different classification methods.

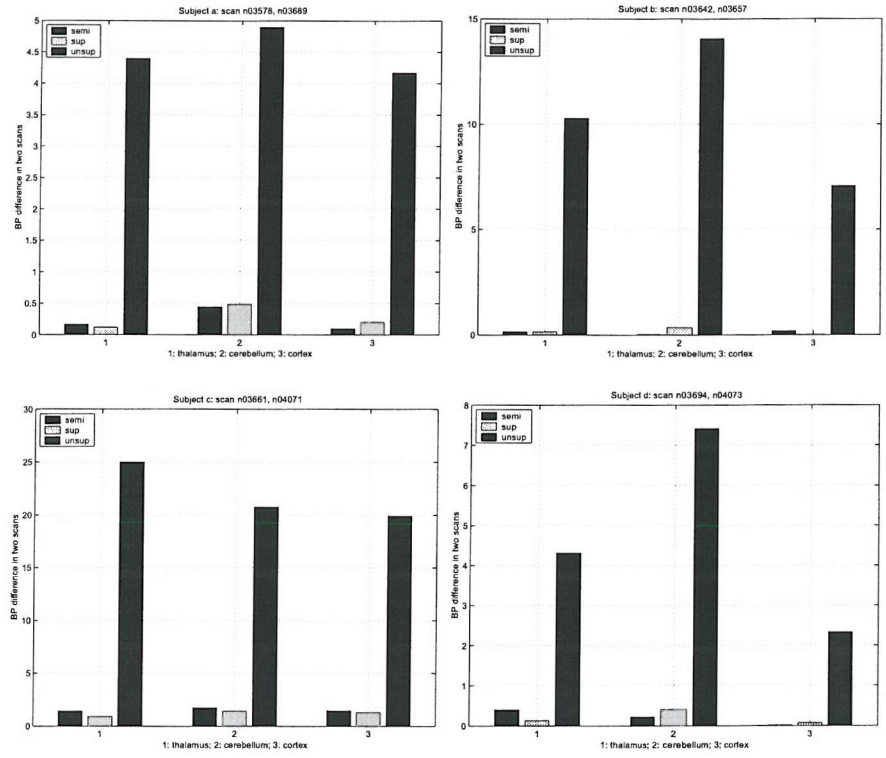
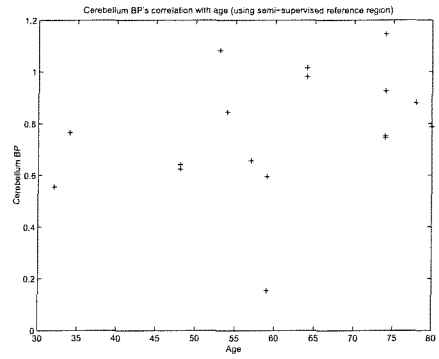


FIGURE 6.19: Test-retest result

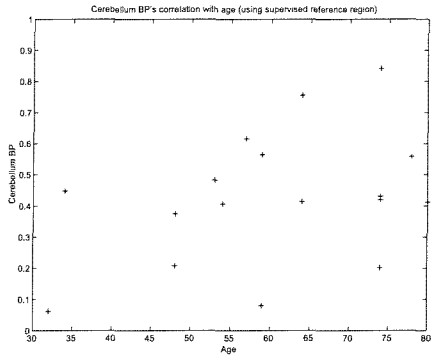
6.2.3 Cerebellum Binding's Correlation with Age

To find possible connections between age and brain function, the age-associated variations in the receptor binding will to be examined using PET data from a group of heathy human subjects at different ages.

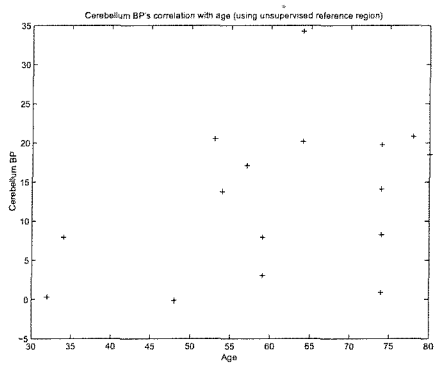
The binding of the cerebellum in 17 healthy scans are examined. Fig. 4.23 shows the estimation of cerebellum in these scans with extracted reference region TAC and the simplified reference region model. Three sub-figures correspond to the results with three different reference region extraction methods. Cerebellum binding's correlation with age is examined using the simplified reference region model. Fig. 6.20 shows the extracted reference TACs via different methods. Table 6.1 shows the R^2 statistics analysis of binding potential's variation explained by age for the three methods. Again, no obvious correlation of age with cerebellum binding potential value can be found in these 17 scans.



(a) Using semi-supervised extracted reference region



(b) Using supervised extracted reference region



(c) Using unsupervised extracted reference region

FIGURE 6.20: Cerebellum binding's correlation with age

Method	R^2 statistics
Unsupervised	0.08%
Supervised	18.12%
SemiSupervised	9.67%

TABLE 6.1: R^2 statistics

6.2.4 Summary

The spatio-temporal extraction of the reference region TAC for 18 [^{11}C](R)-PK11195 PET data has been described and analogised. The importance of using labelled data (i.e. expert knowledge)in the segmentation process has been justified . The extracted TAC using methods using labelled data, supervised and semi-supervised image segmentation, leads to more stable and reliable results than using unsupervised image segmentation, which justifies the two proposed image segmentation methods. Although the performance evaluation is very difficult for real PET data, the test-retest scheme show that the supervised and semi-supervised extracted reference region TACs are relatively consistent.

6.3 Conclusions

Spatio-temporal reference region extraction with simulated and real PET data are performed in this chapter, using MRF model-based image segmentation. The unsupervised image segmentation and the two new methods developed in chapter 5 is used. It shows that the inclusion of expert knowledge and image models greatly reduces the uncertainty in the segmentation. The new semi-supervised framework achieves substantial performance gains over the other methods.

Chapter 7

Conclusions and Future Work

7.1 Summary of the Thesis

This thesis has focused on modern data classification techniques for automatic positron emission tomography image segmentation. The aim of this thesis is to explore reference region localisation in PET with the unlabelled data and expert labelled data to achieve higher accuracy. A hierarchical supervised image segmentation enables labelled data to be used in image segmentation. A combined learning framework for model-based image segmentation is proposed to tackle the general image segmentation problem with the aid of both labelled and unlabelled data. The efficiency of these two methods are tested on both simulated and real PET data.

The background knowledge of PET imaging is reviewed in Chapter 2. The process of the PET experiment to obtain data in both spatial and time domain is introduced. With the consideration of the quality of the data and the complexity of the noise sources, the analysis of the PET data is very important. Different compartmental models have been developed, while the simplified reference region model is our main interest in this thesis. The simplified reference region model is superior to other models in that no blood sampling is needed, based on the assumption that there is a region devoid of specific binding. Finally the problem of efficiently and accurately extracting this PET reference region is proposed.

Since the extraction of the reference region in PET is inherently a problem of learning, or more specifically, classification from data, the problem of pattern recognition is formulated in chapter 3. Learning and generalisation is introduced. The mathematical formulations of supervised classification, unsupervised classification and semi-supervised classification are given, with a review of related methods and optimisation processes.

Chapter 4 shows the result of using different pattern recognition techniques to the PET reference region exaction, for both synthetic and $[^{11}C](R)$-PK11195 PET data sets. The

experimental comparison is also given. In simulations, the error bars generated by semi-supervised method show less error and display more stability than those for unsupervised and supervised method. In [^{11}C](R)-PK11195 PET data application, as no ground truth is available, a test-retest scheme is used to estimate the performance. The simplified reference region model is used for generating binding potential images, which give important information on the scanned subject.

Various statistical models are widely applied in image segmentation since the paper of Geman and Geman (1984). The models and techniques for low-level image segmentation are reviewed in chapter 5. Markov random field models are discussed to model the spatial correlation of image pixels or voxels, which is not considered in chapter 3 and 4. The use of image models like MRFs in image segmentation is equivalent to regularisation in the spatial domain, overcoming the data independence assumption in most data classification techniques, which is more suitable for modelling images. However, the introduction of MRF models in the image segmentation makes the optimisation difficult. The unsupervised image segmentation problem is discussed. The problem of learning in image segmentation with labelled data is also discussed. A new hierarchical image segmentation with supervised learning is proposed. Additionally, a new framework of learning from both labelled and unlabelled data as well as modelling the voxel's spatial correlations is proposed. An approximate EM algorithm is also proposed to solve the related difficult optimisation problem.

Chapter 6 shows the application of the image segmentation with labelled data. The unsupervised image segmentation, as well as the two proposed supervised and semi-supervised image segmentation methods, are applied to the synthetic and [^{11}C](R)-PK11195 PET data set. Similar to chapter 4, error bars are displayed to compare three segmentation results. Semi-supervised image segmentation achieves best accuracy and has more stable performance than the rest. These spatio-temporal segmentation results also show performance gains over the temporal segmentation results in chapter 4. In [^{11}C](R)-PK11195 PET data application, the test-retest scheme is used to estimate the performance. The simplified reference region model is used for generating binding potential images.

In conclusion, this thesis shows that the inclusion of expert knowledge greatly reduces the uncertainty in the PET segmentation. The new semi-supervised framework achieving substantial performance gains over the other methods.

7.2 Suggestions for Future Work

As is often the case in research, a lot of problems remain unsolved and many new problems arise as the result of the research conducted. Some possible avenues for future research are outlined here.

- Application of the combined learning image segmentation framework to other contexts, such as function MRI, sonar images, remote sensing images and text. In this thesis, different image segmentation methods have been used for PET reference region localisation and modelling. There are various medical imaging techniques available, such as functional MRI, EEG, MEG etc. From the view of statistical learning, the data generated from these imaging methodologies are in the similar form as PET data, i.e. data are spatial-temporally connected. This enables the image segmentation methods proposed in this thesis to be used in these wider medical imaging fields. Apart from potential applications in various types of medical images, these methods can also be applied to images from other sources such as remote sensing. Text segmentation is another field where the segmentation result depends heavily on context. The segmentation methods proposed in this thesis may have application here.

- The development of new combined learning methods and the quantitative analysis of the efficiency of the combined classification models. The combined learning using both labelled data and unlabelled data lies between two well-founded areas: supervised learning and unsupervised learning. Thus there are naturally two ways to treat the combined learning problem: supervised learning methods based techniques with the additional unlabelled data to provide extra prior knowledge or a source of information to add more generalisation to the model; unsupervised learning category with the labelled data treated as the complete (or partially complete) data set while the unlabelled data as the incomplete data set. Further research into a more general combined framework, which sees these two avenues as particular approaches would be elegant.

References

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.

Ashburner, J., J. Haslam, C. Taylor, and V. Cunningham (1996). A cluster analysis approach for the characterization of dynamic PET data. *Quantification of Brain Function Using PET*, 301–306.

Banati, R., G. Goerres, and R. Myers (1999). [^{11}C](*R*)-PK11195 positron emission tomography imaging of activated microglia in vivo in rasmussen's encephalitis. *Neurology 53*, 2199–2203.

Barber, D. and C. K. I. Williams (1996). Gaussian processes for bayesian classification via hybrid monte carlo. *Advances in Neural Information Processing Systems 9, MIT Press, Cambridge, MA.*

Barker, S. and P. Rayner (2000). Unsupervised image segmentation using markov random field models. *Pattern Recognition 33*, 587–602.

Barron, A., J. Rissanen, and B. Yu (1998). The minimum description length princilple in coding and modeling. *IEEE Transactions on Information Theory 44*(6), 2743–2760.

Benavides, J., P. Cornu, and T. Dennis (1988). Imaging of human brain lesion siwth an omega 3 site radioligand. *Ann Neurol 24*, 708–712.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of Royal Statistical Society B 48*(3), 259–302.

Bishop, C. (1995). *Neural Networks for Pattern Recognition.* UK: Oxford University Press.

Blum, A. and S. Chawla (2001). Learning from labeled and unlabeled data using graph mincuts. *International Conference on Machine Learning*, 19–26.

Blum, A. and T. Mitchell (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, 92–100.

Boudraa, A., J. Champier, L. Cinotti, and J. Bordet (1996). Delineation and quantitation of brain lesions by fuzzy clustering in positron emission tomography. *Computerized Medical Imaging and Graphics 20*(1), 31–41.

Bouman, C. A. and M. Shapiro (1994, March). A multiscale random field model for Bayesian image segmentation. *IEEE Trans. Image Processing 3*(2), 162–177.

Broomhead, D. and D. Lowe (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems 2*, 321–355.

Burnham, K. and D. Anderson (1998). *Model Selection an Inference.* Springer.

Castelli, V. and T. Cover (1996). The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Trans. Information Theory 42*(6), 2102–2117.

Chan, H., K. Doi, S. Galhotra, C. Vyborny, H. MacMahon, and P. Jokich (1987). Image feature analysis and computer-aided diagnosis in digital radiography: Part 1 automated detection of microcalcifications in mammography. *Medical Physics 14*, 538–548.

Chandler, D. (1987). *Introduction to Modern Statistical Mechanics.* Oxford University Press.

Cherkassky, V. and F. Mulier (1998). *Learning From Data: Concepts, Theory, and Methods.* John Wiley & Sons, Inc. , New York.

Clifford, P. (1990). Markov random fields in statistics. *In G. Grimmett and D.J. Welsh editors, Disorder in Physical Systems.*

Cohn, D., Z. Ghahramani, and M. Jordan (1996). Active learning with statistical models, , (4): 129-145. *Journal of Artificial Intelligence Research* (4), 129–145.

Courant, R. and D. Hilbert (1953). *Methods of Mathematical Physics.* Interscience.

Cristianini, N. and J. Shawe-Taylor (2000). *An Introduction to Support Vector Machines (and other kernel-based learning methods).* Cambridge University Press.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B39 (1)*, 1–38.

Derin, H., H. Elliott, R. Cristi, and D. Geman (1984, November). Bayes smoothing algorithms for segmentation of binary images modeled by Markov random fields. *IEEE Trans. Pattern Analysis and Machine Intelligence 6*(6), 707–719.

Devroye, L., L. Gyorfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition.* Springer.

Doulamis, D., D. Doulamis, and K. D.S. (2000). On-line retrainable neural networks: Improving the performance of neural networks in image analysis problems. *IEEE Trans. Neural Networks 11*(1), 137–155.

Fox, P. T., M. A. Mintun, M. E. Raichle, and P. Herscovitch (1984). A noninvasive approach to quantitative functional brain mapping with h2 ^{15}o and positron emission tomography. *J. Cereb. Blood Flow Metab. 4*(4), 329–333.

Friston, K., A. Holmes, K. Worsley, J. Poline, C. Frith, and R. Frackowiak (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping 2*, 189–210.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence 6*(6), 721–741.

Genovese, C. R. (2000). A Bayesian time-course model for functional magnetic resonance imaging data. *Journal of American Statistical Association*.

Ghahramani, Z. and M. I. Jordan (1994). Supervised learning from incomplete data via an em approach. *NIPS 6*.

Green, P. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika 82*, 711–732.

Greiger, D. and F. Firosi (1991). Parallel and deterministic algorithms for MRF's: surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13*(5), 401–412.

Gunn, R., A. Lammertsma, and V. Cunningham (1996). Parametric imaging of ligand-receptor interactions using a reference tissue model and cluster analyse. *Quantification of Brain Function Using PET*, 401–406.

Gunn, R. N. (1996). *Mathematical modelling and identifiability applied to positron emission tomography data.* Ph. D. thesis, University of Warwick.

Gunn, R. N., S. R. Gunn, and V. J. Cunningham (2001). Positron emission tomography compartmental models. *Journal of Cerebral Blood Flow and Metabolism 21*, 635–652.

Gunn, S. R. (1998). Support vector machines for classification and regression. Technical Report ISIS-1-98, Department of Electronics and Computer Science, University of Southampton.

Gutfinger, D. and J. Sklansky (1991). Robust classifiers by mixed adaptation. *IEEE Trans. Pattern analysis and Machine Intelligence 13*(6), 552–567.

Hammersley, J. and P. Clifford (1971). Markov field on finite graphs and lattices. unpublished.

Hampshire, J. and B. Pearlmutter (1990). Equivalence proofs for multi-layer perceptron classifiers and the Bayesian discriminant function. *Proceedings of the 1990 Connectionist Models Summer School*, 159–172. San Mateo, CA: Morgan Kaufmann.

Hartigan, J. (1975). *Clustering Algorithms.* New York: Wiley.

Haykin, S. (1999). *Neural Networks: a comprehensive foundation.* Prentice Hall Press, USA.

Iidaka, T., N. Anderson, S. Kapur, R. Cabeza, C. Okamoto, and F. Craik (1999). Age-related differences in brain activation during encoding and retrieval under divided attention: A positron emission tomography (PET) study. *Brain and Cognition 39*(1), 53–55.

Ivanov, Y., B. Blumberg, and A. Pentland (2001). Expectation maximization for weakly labeled data. *International Conference on Machine Learning.*

Jaakkola, T. and D. Haussler (MIT Press, 1998). Exploiting generative models in discriminative classifiers. *NIPS 11*, 487–493.

Jacquez, J. (1985). *Compartmental Analysis in Biology and Medicine* (2nd ed.). Ann Arbor: The University of Michigan Press.

Jain, A. and K. Karu (1996). Learning texture discrimination masks. *IEEE Trans. Pattern Analysis and Machine Intelligence 18*(2), 195–205.

Jebara, T. and A. Pentland (1998). Maximum conditional likelihood via bound maximization and the CEM algorithm. *NIPS 11*, 494–500.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. *Proceeding of the 16th International Conference on Machine Learning (ICML)*, 200–209.

Jordan, M. and R. Jacobs (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation 6*, 181–214.

Jordan, M. I. (1998). *Learning in Graphical Models.* Kluwer Academic Publishers.

Kak, A. and M. Slaney (1988). *Principles of Computerized Tomographic Imaging.* New York: IEEE Press,.

Kety, S. and C. Schmidt (1948). The nitrous oxide method for the quantitative determination of cerebral blood flow in man: theory, procedure and normal values. *J Clin Invest* (27), 476–483.

Kiebel, S., J. Ashburner, J. Poline, and K. Friston (1997). MRI and PET coregistration - a cross validation of statistical parametric mapping and automated image regression. *Neuroimage 5*, 271–279.

Kimura, Y., H. Hsu, H. Toyama, M. Senda, and N. M. Alpert (1999). Improved signal-to-noise ratio in parametric images by cluster analysis. *NeuroImage 9*, 554–561.

Kuhl, D. E. and R. Q. Edwards (1963). Image separation radioisotope scanning. *Radiology 80*, 653–661.

Lammertsma, A. and S. Hume (1996). Simplified reference tissue model for PET receptor studies. *Neuroimage 4*, 153–158.

Langer, S. and S. Arbilla (1988). Limitations of the benzodiazepine receptor nomenclature: a proposal for a pharmacological classification as omega receptor subtypes. *Fund Clin Pharmacol 2*, 159–170.

Li, S. (1995). *Markov Random Field Modelling in Computer Vision.* Springer-Verlag.

Mackay, D. J. C. (1992a). Bayesian interpolation. *Neural Computation 4*(3), 415–447.

Mackay, D. J. C. (1992b). The evidence framework applied to classification networks. *Neural Computation 4*(5), 720–736.

MacQueen, J. (1967). Some methods for classificaiton and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability I*, 281–297.

MaLachlan, G. J. and K. E. Basford (1988). *Mixture Models: Inference and Applications to Clustering.* New York: Marcel Dekker.

Miller, D. and H. Uyar (1996). A mixture of experts classifier with learning based on both labelled and unlabelled data. *NIPS 9*, 571–577.

Mintun, M., M. Raichle, and M. Kilbourn (1984). A quantitative model for the in vivo assessment of drug binding sites with positron emission tomography. *Ann. Neurol 15*, 217–227.

Neal, R. and G. Hinton (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. *in Learning in Graphical Model M.I. Jordan (editor).*

Neal, R. M. (1993). Probabilistic inference using markov chain monte carlo methods. Technical report crg-tr-93-1, Dept. of Computer Science, University of Toronto.

Neal, R. M. (1994). *Bayesian Learning for Neural Networks.* Ph. D. thesis, University of Toronto, Canada.

Nigam, K. and R. Ghani (2000). Analyzing the effectiveness and applicability of co-training. *Ninth International Conference on Information and Knowledge Management*, 86–93.

Nigam, K., A. K. Mccallum, S. Thrun, and T. Mitchell (2000, May). Text classification from labeled and unlabeled documents using em. *Machine Learning* (39(2/3)), 103–134.

Niyogi, P. and F. Girosi (1996). On the relationship between generalization error, hypothesis complexity and sample complexity in regulaization networks. *Neural Computation 8*, 819–842.

Patterson, D. W. (1996). *Artificial neural networks: theory and applications.* Prentice Hall.

Phelps, M. and S. Gambhir (1993). Let's play pet! Available from http://www.crump.ucla.edu/lpp/lpphome.html.

Phelps, M., J. Mazziotta, and H. S. (eds) (1986). *Positron Emission Tomography and Autoradiography: Principles and Applications for the Brain and Heart.* New York: Raven Press.

Richardson, S. and P. Green (1997). On Bayesian analysis of mixtures with unknown number of components. *Journal of Royal Statistical Society B 59*, 731–792.

Ripley, B. and A. Sutherland (1990). Finding spiral structures in images of galaxies. *Philosophical Transactions of the Royal Society A*(332), 477–485.

Rissanen, J. (1987). Stochastic complexity. *Journal of Royal Statistical Society B 49*, 223–239.

Ruanaidh, J. O. and W. Fitzgerald (1996). *Numerical Bayesian Methods Applied to Signal Processing.* Springer-Verlag, New York.

Schuurmans, D. and F. Southey (2001). Metric-based methods for adaptive model selection and regularization. *Machine Learning, Special Issue on New Methods for Model Selection and Model Combination.*

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* (6), 461–464.

Sclove, S. (1983). Application of the conditional population mixture model to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 5*(4), 428–433.

Shahshahani, B. and D. Landgrebe (1994.). The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *IEEE Trans. Geosci. Remote Sensing 32*(5), 1087–1095.

Sokoloff, L., M. Reivich, C. Kennedy, M. DesRosiers, C. Patlak, K. Pettigrew, O. Sakurada, and M. Shinohara (1977). The ^{14}C-deoxyglucose method for the measurement of local cerebral glucose utilisation: theory, procedure and normal values in the conscious and anaesthetized albino rat. *J. Neurochem* (28), 897–916.

Svensen, M., F. Kruggel, and D. Y. V. Cramon (2000). Probabilistic modeling of single-trial fMRI data. *IEEE Trans. Medical Imaging 19*(1), 25–35.

Titterington, D., A. Smith, and U. Makov (1985). *Statistical Analysis of Finite Mixture Distributions.* Chichester, U.K.: John Wiley & Sons.

Vapnik, V. (1998). *The Nature of Statistical Learning Theory.* John Wiley & Sons, Inc. , New York.

Vowinckel, E., D. Reutens, and B. Becher (1997). Pk 11195 binding to the peripheral benzodiazepine receptor as a marker of microglia activation in multiple sclerosis and experimental autoimmune encephalomyelitis. *J Neurosci Res 50*, 345–353.

Werbos, P. J. (1994). *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting.* New York: Wiley.

Whindham, M. and A. Cutler (1992). Information ratios for validating mixture analysis. *Journal of American Statistical Association 87*, 1188–1192.

Williams, C. K. I. (2000). A mcmc approach to hierarchical mixture modelling. *Advances in Neural Information Processing Systems* (12), 680–686.

Xu, L. and M. Jordan (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation 8*, 129–151.

Yamamura, H. (1990). *Methods in Neurotransmitter Receptor Analysis.* Lippincott-Raven Press.

Yap, J., C. M. Kao, M. Cooper, C. Chen, and W. M. (1996). Sinogram recovery of dynamic PET using principal component analysis and projection onto convex sets. *Quantification of brain function using PET*, 109–112.

Zhang, J. (1992). The mean field theory in EM procedures for Markov random fields. *IEEE Trans. Signal Processing 40*(10), 2570–2583.

Zhang, J., J. Modestino, and D. Langan (1994). Maximum-likelihood parameter estimation for unsupervised stochastic model-based image segmentation. *IEEE Trans. Image Processing 3*(4), 405–419.

Zubieta, J., R. Dannals, and J. Frost (1999). Gender and age influences on human brain mu-opioid receptor binding measured by PET. *American Journal of Psychiatry 156*(6), 842–848.