Some New Results in Nearest Neighbour Classification and lung sound analysis

A thesis submitted to the Faculty of Engineering

of the University of Glasgow

for the degree of Doctor of Philosophy

Andrew Luk

February, 1987

This work was supported by the Croucher Foundation, Hong Kong.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

This thesis is dedicated to

my parents, my only sister and Anita.

Preface and Acknowledgements

Work in the Department of Electronics and Electrical Engineering, University of Glasgow on human lung sounds was originally begun in the mid-1970s by Mr. J. McGhee on a suggestion of Dr. A. Pack, and was continued by Dr. R.B. Urquhart. This thesis is the present author's contribution and represents the results of research carried out over the period 1983 - 1986. All parts are the author's original work as agreed in discussion with Dr. J.E.S. Macleod, except that parts of chapter 6 are based on contributions by Professor D.M. Titterington. These contributions are acknowledged in the text.

I would like to express my thanks to Professor J. Lamb of the Department of Electronics and Electrical Engineering, University of Glasgow, for the provision of research facilities and also nominating myself for a scholarship from the Croucher Foundation, Hong Kong. My deepest gratitude is given to Lord Todd for accepting me as a recipient for a scholarship from the Croucher Foundation, Hong Kong.

I would also like to thank Dr. J.E.S. Macleod for supervising and encouraging my research. In particular, I must thank him for teaching me the virtue of the English language. I would also like to thank Professor D.M. Titterington of the Department of Statistics, University of Glasgow, for many sessions of useful discussion in nearest neighbour classification. Thanks are also due to Mr. D.D. Campbell for

developing hardware and assembly level programmes for the lung sound work. The assistance of Dr. K.T. Macfarlane, Dr. R.B. Urquhart and Miss A. Mackinnon in programming the PDP-11, GEC 4080 and GEC 4180 is also deeply appreciated.

This research would not have been possible without the provision of research facilities at Glasgow Royal Infirmary by Dr. F. Moran. The clinical part of the collaborative research was carried out by Dr. K. Anderson and my thanks are given for his assistance and advice over a long period.

Finally, I would like to thank my Ph.D. colleages for all their wonderful jokes and occasional good-natured insults. I would also like to thank the academic staff and technicians of this department for their frequent support and encouragement, especially Mr. J. Young and Mr. D. Atkins for their help in bringing the instrumentation tape recorder back and forth from Glasgow Royal Infirmary, and to the staff of the Department of Computing Science, University of Glasgow for teaching me the secret of computing. My thanks should also be extended to Dr. E.Y.B. Pun, Dr. H.D. Liu, Mr. J. Cheng, Miss I. Tang, Mr. and Mrs. Tam for their hospitality and their kind advice. Last but not least, I would like to thank Mr. D. Guy, Mr. M. MacGrath, Mr. P. Mckeown, Mr. N. Goldsmith and Mr. A. Brown for introducing myself to the underworlds of dungeons, the perils of fire breathing dragons, the unholy cultists of the twenties and thirties, the blood-lusted future cops, and the understanding of deep-space battles with alien races like Klingon and Gorn.

Summary

This thesis describes the results of the author's contribution to a collaborative research programme between the Department of Respiratory Medicine, Glasgow Royal Infirmary, and the Department of Electronics and Electrical Engineering, University of Glasgow. After the first six months of research, it was decided that the project would be limited to demonstrating the feasibility of developing a non-invasive examination system for patients exposed to asbestos dusts. The development of this system led to a growing interest in nearest neighbour (NN) classification algorithms and to the investigation of a number of interesting problems in this area.

In particular, it is argued that when the size of the prototype set is finite, a weighted k-NN rule may, in some cases, has a smaller probability of error than an unweighted k-NN rule. Analytical solutions to a simple example and experimental results are presented to support this argument.

A new NN classification scheme is also described in this thesis. Again a number of modifications for the case of a finite prototype set are suggested, and experimental results are given.

The remainder of the work in this thesis concerns the development of the proposed non-invasive examination system which uses lung sound as its input. Due to the complexity of the

proposed system and the initial small data set, only part of the system has been implemented. However, preliminary experimental results on three groups of patients, namely (a) patients with asbestosis, (b) patients who have exposed to asbestos dust, and (c) healthy subjects, have shown that it is possible to discriminate these three groups of patients. More extensive studies are required before the system can be used in clinical conditions. Suggestions for these continuing studies are made.

Contents

Introduction

This work was supported by the Croucher Foundation, Hong Kong.

Chapter 1: Introduction

1.1 A brief historical account of the lung sound research

Ever since Laennec (1819) invented his stethoscope, lung sound has been used as a diagnostic tool. Unfortunately, in those earlier days, lung sounds were identified by a proliferation of different terminologies which were subjective and depended on the interpretations of individual physicians. Therefore, as science and technology advanced, the stethoscope was slowly replaced in importance by other non-invasive investigative techniques, such as radiography and more recently nuclear magnetic resonance imaging. Perhaps it was the work of Forgacs (1978) in the 1960s and early 1970s, who stressed the merit of a scientific approach, that stimulated a growing interest in lung sound analysis in the mid-seventies.

It was at about this time that the lung sound research programme also started in Glasgow University. It has been going on for nearly ten years and is a collaboration between the Department of Respiratory Medicine, Glasgow Royal Infirmary, and the Department of Electronics and Electrical Engineering at the University. McGhee (1978) designed a suitable transducer for subsequent lung sound research. Urquhart et al (1981) and Banham et al (1984) investigated the significance of low frequency lung sound signals (0 to 50Hz). Four groups of patients were used in their studies: (a) patients with asbestosis, (b) patients with interstitial pulmonary oedema (IPO), (c) patients with

cryptogenic fibrosing alveolitis (CFA), and (d) healthy non-smoking (or normal) subjects. The application of Urquhart's (1980, 1982) new graph-theoretical clustering algorithms and the Karhunen-Loeve transformation (Watanabe, 1965) has yielded three slightly overlapping clusters, namely (a) patients with IPO, (b) patients with fibrosis (asbestosis and CFA), and (c) normal subjects. From this result, the authors concluded that lung sound signals analysis may contribute to non-invasive diagnosis.

This thesis reports the recent research on nearest neighbour classification (Luk and Macleod, 1986; Macleod, et al, 1987) and on lung sound signal analysis (Anderson et al, 1986). The original aim of the project was to analyse lung sound automatically by using pattern recognition and signal processing techniques. However, after the first six months of research, a decision was taken to limit the project to demonstrating the feasibility of developing a non-invasive system that can routinely be used to examine patients exposed to asbestos dusts, using lung sound as its input. Further, it was hoped that the proposed system could be able to distinguish between patients with asbestosis and patients without. The development of this system led to a growing interest in nearest neighbour (NN) classification algorithms and to the investigation of a number of interesting problems in this area.

1.2 A general introduction to pattern recognition and signal processing

In a sense, pattern recognition is being performed daily by every individual. Every individual uses his/her five senses to detect, interpret and learn about the surroundings. Suppose for example that a person wants to find a bar of chocolate that he/she knows is inside a room. That person will use his/her eyes as a pair of visual input transducers to scan that room. If, say, there is one bar of chocolate lying on a table, it is up to the brain to segment that part of visual information and then to identify that the object lying on the table is a bar of chocolate. This example also demonstrates the important of learning. That person must have been taught in the past what a bar of chocolate looked like and how it differed from a table. He/she must also have the a priori knowledge that there may be a bar of chocolate in that room.

Although it is not possible to imitate all human recognition abilities completely, it is possible, nevertheless, to perform some of these abilities artificially. Two approaches are generally available in an artificial system: one, the statistical (or decision-theoretical) approach, is based on mathematical models (Devijver and Kittler, 1982; Duda and Hart, 1973; Fukunaga, 1972; Patrick, 1972); and the other, the syntactic (or structural) approach ,is based on linguistic models (Fu, 1974; Gonzalez and Thomason, 1978; Pavlidis, 1977). As in the above example of human recognition, the operation of an

artificial system can usually be divided into a learning and a recognition phase. At the learning phase, the system will attempt to generate and select a set of useful properties or attributes from the information presented by a transducer. In syntactic methods the attributes are the "primitives" which together constitute the complete pattern. In statistical pattern recognition, on the other hand, the properties are "features" which together form a different type of description of the pattern and are treated as the components of a feature vector. Learning algorithms, which can be supervised or unsupervised in nature (Anderberg, 1973), can then be applied to study the underlying relationships for these patterns, such as the class membership of each pattern or the grammar that consitutes the patterns. During the recognition phase, the selected features or primitives from an input pattern will be identified as belonging to one of the predetermined classes or families of grammars respectively through the use of suitable classifiers (such as those using the nearest neighbour type of algorithm).

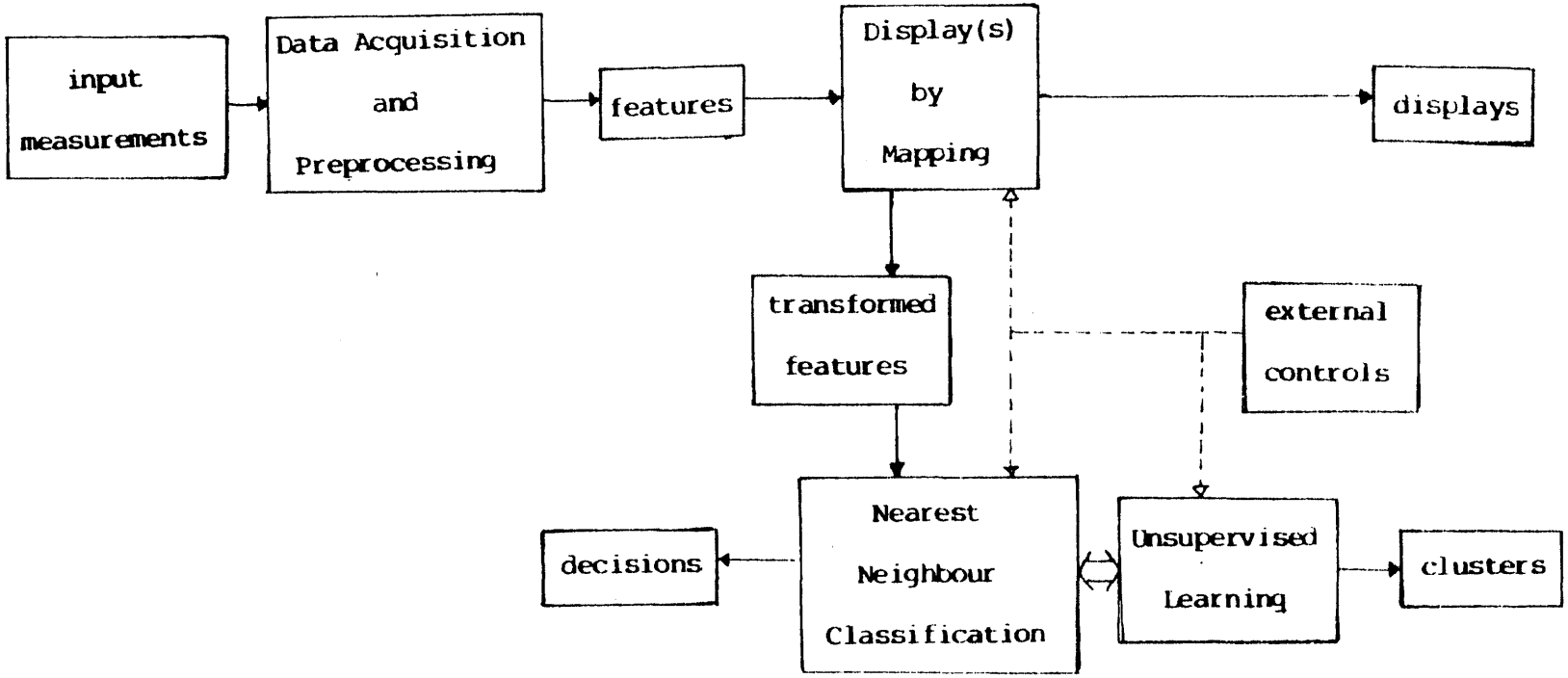
Pattern recognition techniques have been applied to numerous problems. These include character recognition, speech processing, remote sensing, fault detection, medical signal analysis and industrial inspection (Macleod, 1985). Some of these applications have been reviewed in a book edited by Fu (1982).

Generally, some sort of signal preprocessing or conditioning is required on the input signals. For example, it

may be necessary to remove high frequency noise from the lung sound signals by filtering. Other less trivial operations, such as spectrum estimation, may also be required before the input signal becomes usable by a pattern recognition system. On the whole, the amount of signal preprocessing needed is dependent on the type of application.

1.3 Aims of research

One of the aims of this project is to demonstrate the feasibility of developing a non-invasive system that can routinely be used to examine patients who have been exposed to asbestos dusts, using lung sound as its input. In particular, the system should be capable of discriminating patients with asbestosis from those without. A modular approach was adopted at the beginning of this project. The major modules in the proposed system are shown in figure 1.1. The data acquisition and preprocessing module is essentially an interface between the patient and the proposed system. In this case, the function of this module is to acquire lung sound from a human chest wall and then perform a number of computational operations, such as digitization, so that a set of features can be generated for the other modules to analyse. These features are regarded as being elements of a multi-dimensional feature vector. The purpose of the display module is to project these higher dimensional lung sound data onto a lower dimensional (sub)space and, for example, to plot the data as a point set in a plane. Physicians or para-medical personnel could then interpret relationships in a set of



Figurel.1 Major modules in the proposed non-invasive system for examinating patients exposed to asbestos dust.

lung sound data, and could decide whether these recordings were consistent with asbestosis or not. This module was implemented by using mapping algorithms, to the extent that off-line displays can be produced on the GEC 4180 mini-computer.

The next module in the system is a classification module which at this stage of development is implemented by using the nearest neighbour type of algorithm. The knowledge of the designer is limited to the amount of information available during the design stage, and it is possible that this information is insufficient to cope with unseen data or that some of it is inaccurate. Therefore, if this module produces excessive error, the knowledge within the system may be incomplete and hence it may be necessary to "re-learn" some missing information by invoking the unsupervised learning module. Furthermore, the classifier can also be used to provide some very simple information about the condition of a subject for the user, and thus may assist physicians or para-medical personnel in reaching a decision. The last module, the unsupervised learning module, can be implemented by using clustering algorithms such as those of Urquhart (1982). This module is used mainly for data exploration. (Due to the small data set available, no unsupervised learning module was implemented in this project.)

The second aim is to investigate the finite sample behaviour of nearest neighbour classification rules. In particular, two nearest neighbour rules will be compared (Macleod et al, 1987), and another new nearest neighbour classification

scheme will be investigated (Luk and Macleod, 1986).

1.4 Organisation of the thesis

The materials covered in this thesis are relevant to research in both biomedical engineering and nearest neighbour classification. A reader who is interested in the lung sound research should read chapters 1 to 4 and part of chapter 5, then skip chapter 6, and continue to read the rest of the thesis. On the other hand, if the reader is interested only in nearest neighbour classification, only chapters 5 and 6 need be read.

Chapter 2 is primarily intended to provide the background on both the material asbestos and the disease asbestosis. Some background information is given on the basic anatomy of the respiratory tract and lung sound. Similarly chapters 3 and 4 are intended to provide brief background knowledge on the signal preprocessing and mapping algorithms used in the proposed system.

Chapter 5 presents a "more than usual textbook" review on nearest neighbour (NN) classification. In particular, a number of NN algorithms is reviewed. A brief summary of the properties of NN classification is also presented.

Chapter 6 reports some new results in nearest neighbour classification. The performances of two NN rules on a prototype set of finite size are compared. Analytical solutions to a

simple example and experimental results are also presented to support the argument. Finally, a new NN classification scheme is described. Again a number of modifications for the case of a finite prototype set are suggested, and experimental results are given.

Chapter 7 describes the work on the proposed non-invasive system. Each individual module is described in more detail. Difficulties encountered during the research and experimental results are also presented.

Finally, chapter 8 gives some general conclusions on the whole research project and also suggests some possible future research.

1.5 Remarks

In this project, lung sounds from over 50 subjects have been recorded over a period of eighteen months, with the help of Dr. Anderson at Glasgow Royal Infirmary. For each recording, several time consuming data transfer operations are required before the lung sound signal can be used (for details see section 7.3). Furthermore, the resulting computer files are very large and cannot all be simultaneously held on disc: hence to do certain further experiments on the recorded data, some of these transfer operations have to be repeated. The overall process is very time-consuming, and any failure on any part of the equipment (e.g. on one of the three computers involved) leads to still

further delay. Consequently, the signals from only 15 subjects have been used and analysed by the proposed system. The results and conclusions presented in chapter 7 should therefore be considered as preliminary.

Work to develop improved, faster data transfer facilities is now in progress (Macleod and Aitken, 1987).

References

[01] Anderberg, M.R. (1973). **Cluster Analysis for Applications**, Academic Press, New York.

[02] Anderson, K., Luk, A., Macleod, J.E.S. and Moran, F. (1986). "The application of pattern recognition and signal processing techniques in the diagnosis of asbestosis". **Thorax**, **41**, 715.

[03] Banham, S.W., Urquhart, R.B., Macleod, J.E.S. and Moran, F. (1984). "Alteration in the low frequency lung sounds in respiratory disorders associated with crackles". **European Journal of Respiratory Disease**, **65**, 58-63.

[04] Devijver, P.A. and Kittler, J. (1972). **Pattern Recognition: a statistical approach**, Prentice-Hall International, Inc., London.

[05] Duda, R.O. and Hart, P.E. (1973). **Pattern Classification and Scene Analysis**, John Wiley and Sons, New York.

[06] Forgacs, P. (1978). **Lung Sounds**, Balliere Tindal, London.

[07] Fu, K.S. (1974). **Syntactic Methods in Pattern Recognition**, Academic Press, New York.

[08] Fu, K.S. (1982), Editor. **Applications of Pattern Recognition**, CRC Press, Inc., Florida.

[09] Fukunaga, K. (1972). **Introduction to Statistical Pattern Recognition**, Academic Press, New York.

[10] Gonzalez, R.C. and Thomason, M.G. (1978). **Syntactic Pattern Recognition - An Introduction**, Addison-Wesley, Reading, Massechutte.

[11] Laennec, R.T.H. (1819). **De l'Auscultation Mediate, ou traite du diagnostic des maladies des poumons et du coeur, fonde principalement sur le nouveau moyen d'exploration**, Brosson and Chaude.

[12] Luk, A. and Macleod, J.E.S. (1986). "An alternative nearest neighbour classification scheme". **Pattern Recognition Letters**, **4**, 375-381.

[13] Macleod, J.E.S. and Aitken, J.S. (1987). Private communication.

[14] Macleod, J.E.S., Luk, A. and Macfarlane, K.T. (1985). "Applications of pattern recognition in respiratory medicine and industrial inspection". Presented in the Image Processing Symposium, University of Glasgow, 12 December, 1985.

[15] Macleod, J.E.S., Luk, A. and Titterington, D.M. (1987). "A re-examination of the distance-weighted k-nearest neighbour rule", **IEEE Trans. on Systems, Man, and Cybernetics**, accepted for publication.

[16] McGhee, J. (1978). Unpublished results.

[17] Patrick, E.A. (1972). **Fundamentals of Pattern Recognition**, Prentice-Hall International, Inc., London.

[18] Pavlidis, T. (1977). **Structural Pattern Recognition**, Springer-Verlag, Berlin.

[19] Urquhart, R.B. (1980). "Algorithms for computation of the relative neighbourhood graph". **Electronics Letters**, **14**, 556-557.

[20] Urquhart, R.B., McGhee, J., Macleod, J.E.S., Banham, S.W. and Moran, F. (1981). "The diagnostic value of pulmonary sounds: a preliminary study by computer-aided analysis". **Computers in Biology and Medicine**, **11**, 129-139.

[21] Urquhart, R.B. (1982). "Graph theoretical clustering based on limited neighbourhood sets". **Pattern Recognition**, **15**, 173-187.

[22] Watanabe, S. (1965). "Karhunen-Loeve expansion and factor analysis - theoretical remarks and applications". **Trans. on the 4th Prague Conference on Information Theory, Statistics, Decision Functions, and Random Process**, Prague, 635-660.

Asbestos and Asbestosis

This work was supported by the Croucher Foundation, Hong Kong.

Chapter 2: Asbestos and Asbestosis

Summary

This chapter introduces the necessary biomedical background for this thesis. The basic anatomy of the respiratory tract is reviewed. The material asbestos is introduced and its biomedical effects are outlined. The theory of the pathology of asbestosis is described and diagnostic methods are reviewed.

2.1 Introduction

Ever since human beings developed the skill of constructing and manufacturing tools from materials in their environment, they have encountered a large number of materials that are useful and, perhaps, essential for survival. Unfortunately some of these materials may cause disease and one such group of disorders, known as occupational lung diseases, is caused by the inhalation of dusts or fumes or noxious substances (Crofton and Douglas, 1975). This group of disorders may range from minor ailments, such as irritation of the air passages by ammonia, to chronic disabling diseases like asbestosis, baritosis, siderosis and silicosis. Some of these diseases are obstructive (i.e. a partial blockage of some of the airways) in nature while others including asbestosis are restrictive (i.e. lung volume is reduced but there is no airway obstruction).

One of these potentially hazardous dusts is formed by

asbestos fibres. As Paul Kotin put it in his foreward for Selikoff and Lee's book (1978):

> "Asbestos, a naturally occuring fibrous material, is a startling example of a material at once uniquely useful, because of its physical and chemical properties, and at the same time potentially hazardous to man. Evidence linking inhalation of asbestos fiber to the development of a group of diseases is generating ever growing concern."

The hazardous nature of asbestos has been suspected for many years. The annual report for 1898 in England (Woman Inspectors of Factories, 1899) gave the following account:

> "The evil effects of asbestos dust have also attracted my attention, a microscopic examination of this material dust which was made by HM Medical Inspector clearly revealed the sharp, glass-like, jagged nature of the particles, and where they are allowed to rise and to remain suspended in the air of a room, in any quantity, the effects have been found to be injurious, as might have been expected."

Unfortunately, as with many other occupational lung disorders, it took more than 30 years before the hazard of asbestos was finally recognized and precautionary measures were introduced (Cooke, 1924; Cooke et al, 1927; Merewether and Price, 1930; Wood and Gloyne, 1930). From there onward, numerous research papers and extensive literature have been published covering the adverse

effects of asbestos and related products. It has become increasingly difficult to write a full review or survey on either asbestos or its related diseases. Therefore, the aim of this chapter is to provide a very general background on asbestos and on the disease asbestosis and its link with lung sound research. For those readers who are unfamiliar with the respiratory system, a brief description is given in the next section. It is then followed, in section 2.3, by a summary on asbestos and its related diseases. Section 2.4 will introduce some common terminology used in lung sound analysis and discuss the various possible lung sound generating mechanisms. Finally, in section 2.5, the disease asbestosis will be discussed.

2.2 Basic anatomy of the respiratory system

The main function of the respiratory system is to provide a suitable medium for the exchange of oxgyen and carbon dioxide between the circulatory system and the environment. Roughly, the respiratory tract can be divided into (a) the upper respiratory tract which comprises the nose, the paranasal sinuses, the eustachian tube, the pharynx and the larynx, and (b) the lower respiratory tract which includes all the conducting airways and the respiratory air units or acini (Crofton and Douglas, 1975; Parkes, 1982). The upper respiratory tract, beside its many other physiological functions, is responsible for filtering, warming and humidifying the inspired air, and at the same time together with the lower respiratory tract for conducting air to and from the acini.

The conducting airways start with the trachea which extends from the larynx and bifurcates into the left and right main bronchi. The left main bronchus gives rise to 2 lobar bronchi and the right gives rise to 3; which also are the numbers of lobes in the left and right lungs. Each lobar bronchus is further divided into many generations of segmental bronchi and each segmental bronchus gives rise to a few generations of bronchioles. The bronchi are characterized by the presence of variable amounts of cartilage inside their muscular wall, which is made up of blood vessels, connective tissues, epithelium, lymphatic vessels, muscle tissues and other specialized cells; whereas there is no cartilage in the bronchioles. The last generation of bronchioles without any alveoli attaching to its wall is known as the terminal bronchioles. Each terminal bronchiole gives rise to a variable number of acini which begins with one to three generations of respiratory bronchioles (that is, bronchioles with alveoli attached to their walls), which are further subdivided into a few generations of alveolar ducts. Each alveolar duct is terminated with two or more air sacs, each containing a number of alveoli which are in close contact with the alveolar capillaries so as to facilitate gaseous exchange. The average number of subdivisons from the trachea to the alveoli is about 23 generations in human beings. The acini, which constitute the lung tissues and gaseous exchange tissues of the lung, together with the terminal bronchioles are alternatively referred as the lung parenchyma (Crofton and Douglas, 1975; Parkes, 1982).

An alveolus has an average diameter of about 0.25 mm, and there are about 200 to 600 million alveoli in human beings, making a total of roughly 40 to 120 m^2 of respiratory surface. The alveolar wall is made up of a number of specialized cells responsible for (a) gaseous exchange, (b) disposal of inhaled foreign material, and (c) immunological activity within the lung. Figure 2.1 is a diagrammatic view of the interface between an alveolus and an alveolar capillary. The surface of the alveolus is covered mainly by "type I" cells i.e. cells which are extremely flat and thin to facilitate function (a); the rest of the surface is occupied by one or more "type II" cells i.e. cells which are capable of replacing an injured type I cell (a type I cell is not capable of regenerating itself when damaged by, say, an asbestos fibre - Spencer, 1985). Both these types of cells are attached to an alveolar basement membrane. Inside the alveolus (i.e. in the alveolar lumen) are white blood cells (leucocytes) of a particular kind called alveolar macrophages, which play an important part in functions (b) and (c).

The capillary surface is made up of a layer of flat and thin capillary endothelial cells which are attached to an endothelial basement membrane. At certain places between the capillary and the alveolus, the basement membranes are in close contact to provide a minimum barrier for gaseous diffusion between the circulating capillary red blood cells (or erythrocytes) and the alveolar gas content (Parkes, 1982). Elsewhere between the two basement membranes is the interstitial alveolar space (or septal space), which contains different types

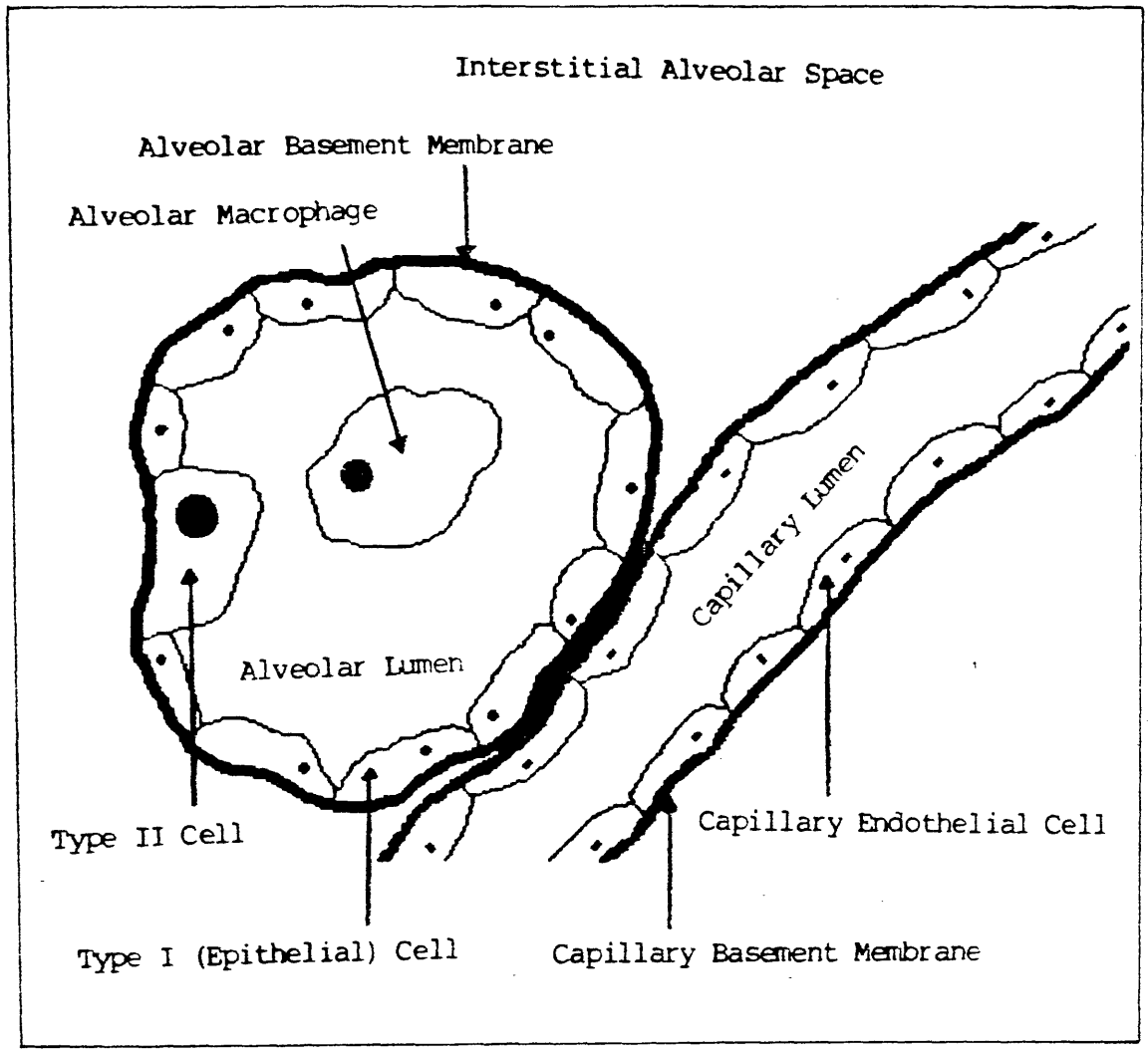


Figure 2.1 A diagrammatic view of an alveolar capillary unit.

of interstitial cells (such as fibroblasts and another type of leucocytes called lymphocytes), various types of fibres (for example collagen fibres) and other substances (like fibronectin and reticulin).

The whole lower repiratory tract, except part of the upper trachea, is enclosed within the thoracic cavity which is the space formed by the chest wall (composed of the ribs and other tissues); whereas the lung parenchyma is completely enclosed inside two pleural cavities with the left lung lying within the left pleural cavity and the right lung inside the right cavity. The wall of each pleural cavity is composed of two layers of connective tissue, known as the visceral pleura and the parietal pleura, which are joined together (or continuous) at the root of the lung parenchyma. The space between the two pleurae is filled with a thin film of lubricative fluid to provide a frictionless surface between the lung parenchyma and the thoracic cavity (Basmajian, 1982; Last, 1984).

2.3 Asbestos

2.3.1 Introduction

The "unquenchable" or "inextinguishable" material - asbestos - has been used, whether by accident or by choice, for many years. Asbestos was incorporated in Finnish pottery as early as 2500 B.C. (Noro, 1968). Sporadic use and mining of asbestos has also been reported in Africa, Asia and Europe long

before the first commercial mining in Quebec (Canada) in 1879 (Selikoff and Lee, 1978). It is not surprising that, in those early superstitious days, asbestos was sometimes misrecognised as a kind of magical material from a mythical creature called the salamander; it simply cannot be destroyed by household fire.

Asbestos has many desirable properties, such as its mechanical strength and its ability to resist fire, acid and alkali, making it a valuable industrial material for many applications. Therefore, ever since the first mining in Canada, the world production of asbestos has been steadily rising despite the fact that the material may be harmful if not under sufficient and effective hygiene control. This is because there is still no suitable substitute for some of its applications. Figure 2.2 shows the world production of asbestos from the year 1960 to 1985. It can be seen that throughout this time the world production of asbestos has increased although there is a tendency to level off in the past few years. In the United Kingdom, on the other hand, there has been a steady drop in imports of raw asbestos ore in the past ten years (figure 2.3). This does not necessarily reflect a decline in the use of asbestos. Rather, as Crofton and Douglas (1975) put it, this may be due to other industrial usages "which are not directly subject to statutory regulations".

2.3.2 Types and Applications

Asbestos is a common name for six different types of

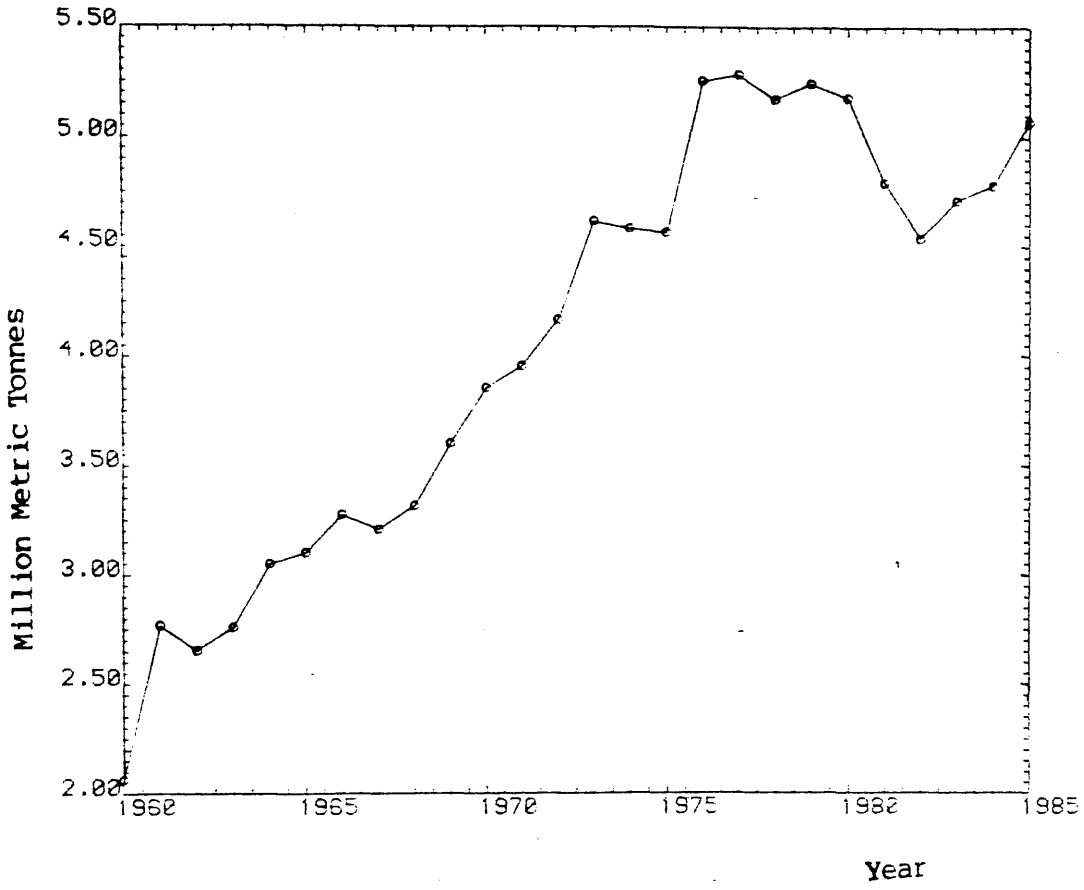


Figure 2.2 World production of asbestos (all types) between 1960 and 1985. (Data are kindly supplied by the United States Department of the Interior, Bureau of Mines, Washington D.C.)

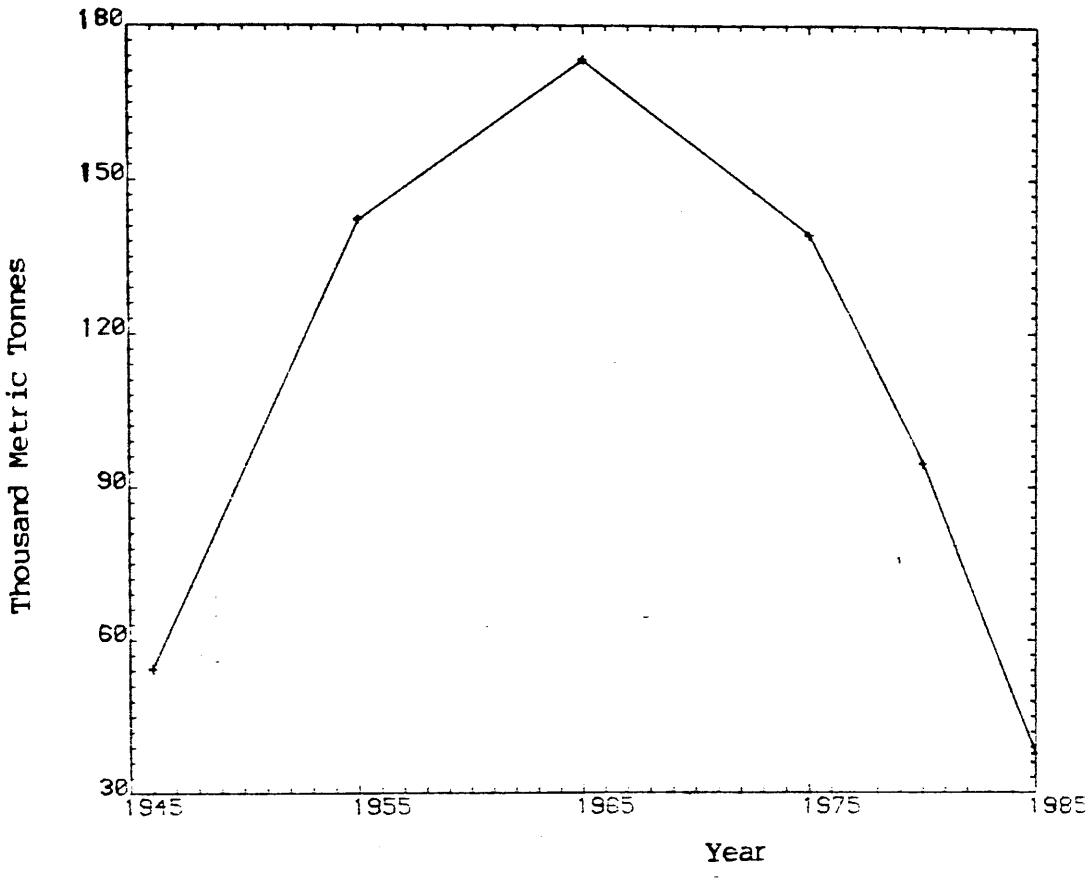


Figure 2.3 Importation of asbestos in United Kingdom. (Data are extracted from various government reports.)

naturally occuring mineral fibres of silicates. One of these materials (chrysotile) is referred to as a serpentine asbestos, and the other five (actinolite, amosite, anthophyllite, crocidolite and tremolite) as amphibole asbestos. The distinction between serpentine and amphibole asbestos depends on the light microscopic appearance which results from the crystal structure (Michaels and Chissick, 1979).

Chrysotile [$3MgO.2SiO_2.2H_2O$], also known as white asbestos because its fibres are usually yellowish or greenish white in colour, is mined mainly in Canada and Russia, and accounts for over 90% of the world production. Its fibre tends to be long, flexible, soft and very fine. It has good fire and alkali resistance but is susceptible to attack by mineral acids. Its applications include many asbestos-cement products for building, insulation and fire proofing, asbestos-textile products such as brake linings, clutch plates and fire protective clothing, and other asbestos-paper products such as filters for many types of fluids (Michaels and Chissick, 1979; Parkes, 1982; Selikoff and Lee, 1978).

Crocidolite [$NaFe(SiO_3)_2.FeSiO_3.H_2O$] or blue asbestos is typically bluish in colour (Michaels and Chissick, 1979). It is mined mainly in South Africa. Its fibre tends to be fine and straight. Its good acid resistant property renders it extremely useful in manufacturing asbestos-cement pressure pipes for chemical plant and other asbestos-textile products where acid resistancy is required. Crocidolite has been most commonly

implicated in the development of asbestos related disease, particularly mesothelioma. For this reason the importation of raw crocidolite has therefore been banned in the United Kingdom since 1970. Nevertheless, substantial quantities of this material still remain in some old installations and may release fibres into the surrounding air if they are damaged.

Amosite [$(FeMg)SiO_3$] is usually pale brown in colour and is often referred to as brown asbestos. It is mined only in South Africa. Its fibre tends to be brittle, thick, straight and very long. Its main application is in fire resistant products. Anthophyllite [$(MgFe)_7.Si_8O_{22}.(OH)_2$] fibre tends to be white in colour. It is mined mainly in Italy. Like amosite, anthophyllite fibre tends to be brittle, straight and long but is much thicker than amosite, chrysotile or crocidolite. It has good fire and chemical resistance. Unfortunately its usage is limited, and in the United Kingdom it is mainly used in friction materials and in fillers and reinforcement products (Health and Safety Executive, 1979).

The other two types of amphibole asbestos, namely actinolite [$CaO.3(MgFe)O.4SiO_2$] and tremolite [$Ca_2Mg_5Si_8O_{22}.(OH)_2$], are rarely used. Both fibres usually occur as a contaminant in other asbestos fibres and in talc. A small amount of tremolite is mined in Italy and in Japan.

2.3.3 Fate and medical effects of the inhaled asbestos fibres

Most of the large airborne particles will be deposited in the upper respiratory tract through interception by the nasal hairs. Smaller airborne particles can be deposited in the lower repiratory tract by three mechanisms:

(a) sedimentation, where a particle is deposited under the influence of gravity,

(b) inertial impaction, where a particle is deposited near the junction of a bifurcation in the airways because of its high momentum, and

(c) diffusion, where a particle is deposited under the influence of Brownian motion.

Both mechanisms (a) and (b) depend on the diameter and the density of the particles (Morgan and Seaton, 1984). Other factors that will generally affect the deposition of asbestos fibres include (i) the possible co-existence of other diseases, such as chronic bronchitis and (ii) minor variations of lung physiology due to different ethnical origin, sex, age, height and even weight. Particles deposited in the conducting airways will be removed by mucociliary clearance (Morgan and Seaton, 1984), a process known as the "ciliary escalator" (Parkes, 1982), in which a layer of mucus (secreted by "goblet cells", in the conducting airways) is propelled slowly towards the larynx by the ciliated epithelial cells also in the conducting airways.

The likelihood that an airborne fibre will deposit in the conducting airways and subsequent be removed by the "ciliary

escalator" depends entirely on its diameter. It has been shown (Timbell, 1965) that fibres with a diameter less than 3 μm are unlikely to deposit in the conducting airways either by sedimentation or by inertial impaction, and hence have a higher probability of depositing in an acinus, where clearance depends on cellular mechanisms (see below) and not on the "ciliary escalator" which terminates at the bronchiolar level.

Asbestos fibres which penetrate into an acinus and are deposited there can be removed by

(a) a nonabsorptive process which involves phagocytosis of the fibres by the alveolar macrophages and later the migration of these leucocytes into the conducting airways so that subsequently they can be removed by the "ciliary escalator", and

(b) an absorptive process which involves selective migration of the asbestos fibres through the alveolar wall either

  (i) via the capillary wall into the bloodstream from which they are removed by other leucocytes in the circulatory system, or

  (ii) the alveolar interstitial space, from which they are later removed by the lymphocytes via the lymphatic vessels.

The amount of dust that each of these two clearance processes can remove has a certain threshold. Above this threshold, the excess fibres simply lie freely in the alveoli. Also, both the absorptive and the nonabsorptive process are restricted by the length and diameter of the fibres deposited. In section 2.5.1,

it will be made clear that asbestosis is a result of excessive immunological response in the acinus. (Other parts of the respiratory tract can also be damaged by these inhaled asbestos fibres. Here also, natural immunological responses will be triggered to repair the damaged tissue.)

Two recent government reports (Health and Safety Executive, 1979; Doll and Peto, 1985) have established that prolonged exposure to a high "dose" of asbestos fibres may cause (a) a number of benign conditions of the pleura such as pleural effusions, diffuse pleural thickening and/or the formation of pleural plaques; and/or (b) a number of possibly fatal diseases such as asbestosis, lung cancer, mesothelioma and/or laryngeal cancer. Other types of cancer (like gastro-intestinal cancer, ovarian cancer and renal cancer) had been reported (Selikoff and Lee, 1978). However, Doll and Peto (1985) were uncertain about the evidence reported so far and claimed that some of the published medical cases may be due to misdiagnosis. The reports also pointed out that smoking will aggravate fibrogenic disorders, like asbestosis, but not necessarily carcinogenic diseases. It has also established that amphibole asbestos is more carcinogenic and fibrogenic than serpentine asbestos. Crocidolite fibre has been single out as both carcinogenic and fibrogenic even under a relatively short and low dose exposure; this is partly due to its physical appearance: long and straight with a small diameter. However the dose-response relationship, which is a measure of the likelihood of occurence of a certain disease with respect to the amount and/or duration of exposure to

fibres, is still not completely clarified for different types of asbestos fibres. Doll and Peto (1985) also argued that the current hygiene standard of 0.5 fibre per ml should be reduced to 0.25 fibre per ml to lower the expected risk of any ailments caused by the exposure to asbestos fibres, especially chrysotile fibres which account for over 80% of the imported asbestos.

2.3.4 Remarks: Difficulties in assessing quantitative effects of asbestos and Dose-response relationships

The difficulties, and sometimes inabilities, in assessing quantitatively the biomedical effects due to exposure to asbestos fibres, as well as the dose-response relationships of asbestos, stem from the following reasons:

(a) there exist more than one type of asbestos fibre, each with different physical and chemical properties, and hence different biomedical responses are to be expected;

(b) the proportion of asbestos fibres having a specific configuration will vary with the environment (for example, in mines or in mills) in which the material is being handled;

(c) the amount of asbestos contained varies with the type of end product and the asbestos may be contaminated by other types of asbestos fibre which are more fibrogenic and carcinogenic (for example commerical chrysotile is usually contaminated by tremolite);

(d) the different methods and equipment that have been used by different countries for estimating the amount of fibres

present in a given environment render some published results difficult to interpret; and

(e) asbestos-related disorders may be complicated by other more fatal disorders, such as circulatory or renal disorders.

2.4 Lung Sound

Ever since Laennec invented the stethoscope in 1816 for mediate auscultation (Kligfield, 1981), the different types of lung sound heard from the chest have been used as a diagnostic tool for many respiratory and circulatory disorders. One of the main drawbacks of this listening approach is that the diagnosis is rather subjective and requires a lot of experience to differentiate the different types of sound. In occupational lung disorders, its importance is now generally replaced by radiology and pulmonary function tests. Nevertheless, change in lung sound is still an important clinical sign for some respiratory disorders (Murphy and Holford, 1980). Indeed, with the help of modern electronics and computing facilities, lung sound may once again be useful in medical research and in diagnosis (Anderson et al, 1986; Murphy et al, 1977; Urquhart, et al, 1981). More important, since this research project concentrates on using lung sound as the input signal for the proposed system for examinating patients exposed to asbestos dusts, a brief review of the different types of sound that can be heard over the chest is necessary.

Lung sound heard over the chest is usually divided into breath sound which is always present during breathing and adventitious (or added) sounds which are not normally present. Breath sound is generally classified into normal or vesicular breath sound and abnormal bronchial breath sounds (Forgacs et al, 1971). The former sound is louder in inspiration than in expiration; whereas the later sounds are associated with the consolidation of the lung tissue (so that the tissue becomes airless between the chest wall and the conducting airways), the transmitted sound being much attenuated and filtered. Forgacs (1978) suggested that breath sound is caused by a central turbulence in the conducting airways and has a flat frequency distribution between 200 and 2000 Hz.

Adventitious sounds, on the other hand, are abnormal lung sounds that usually occur in a number of lung disorders such as asbestosis. They are generally divided into crackles (or rales or crepitations) which are short explosive sounds and wheezes (or rhonchi) which are continuous sounds with well defined frequency characteristics (Bunin and Loudon, 1979; Murphy, 1981). Forgacs (1978) suggested that crackles might be caused by abrupt opening of small airways in the lung; while wheezes are caused by a mechanism similar to that of "a reed in a child's toy trumpet". Wheezes and early inspiratory crackles are usually associated with obstructive disorders while late inspiratory crackles are usually associated with restrictive diseases such as asbestosis (Nath and Capel, 1974).

2.5 Asbestosis

2.5.1 Theory of the pathology of asbestosis

Asbestosis is a chronic fibrosing disorder caused by inhalation of asbestos fibres. It belongs to a group of lung disorders known as diffuse interstitial collagenous fibrosis, i.e. the excessive production of collagen fibres. Its development is related to both duration and quantity of exposure to the fibres. The disease usually takes 20 or more years to develop after the initial exposure (Parkes, 1982).

It is now generally believed that the initial pathological change of asbestosis, which is the derangement of the alveolar structure, is the result of an alveolitis rather than an interstitial fibrosis (Wagner, 1965). Alveolitis is a disorder characterised by (a) a marked increase in the total number of inflammatory and immune effector cells (such as alveolar macrophages and lymphocytes) within the alveolar structures; (b) a change in the relative proportion of one or more effector cell populations; and (c) an activation of one or more effector cell types, such as the presence of neutrophils (another type of leucocyte) in the alveolar structure (Keogh and Crystal, 1982).

The accumulation and the activation of these effector cells will alter (or derange) the natural composition of the alveolar structures and hence may affect the lung function by the

loss of these functional alveolar-capillary units. To quote Keogh and Crystal (1982), "the specific effects of such physical deformations on lung function is unknown, but they probably influence critical mechanical processes involved in the respiratory cycle". (In chapter 7, this may partly explain why patients who have been exposed to asbestos but in whom asbestosis has not developed form a separate group from patients with asbestosis and from normal patients without any exposure to asbestos.)

It has been shown (Crystal et al, 1981; Gadek et al, 1981) that the alveolitis in asbestosis is characterised by an increase in the number of alveolar macrophages and a large number of neutrophils. These neutrophils are extremely dangerous because they will release a number of inflammatory mediators (such as oxidants and connective tissue specific proteases) that can injure the parenchymal tissues (for example, both type I cells and the basement membranes in the alveolus) and disorganise the content of the septal spaces. Gadek et al (1980) have obtained evidence that the asbestos fibre stimulates the alveolar macrophage to release a chemotactic factor which attracts these neutrophils from the circulatory system. The future development of the disease (and hence the probability that a patient who has been exposed to asbestos will subsequently develop asbestosis) depends on the intensity of the alveolitis.

Furthermore, the interaction between the asbestos fibre and the alveolar macrophages also results in the release of a

fibrogenic factor which causes proliferation of fibroblasts and, hence, an increase in the synthesis of collagen fibres. These collagen fibres will form a fine network around the damaged alveolus. If the disease is subsequently developed, the collagen fibres will continue to expand into the alveolar ducts and the surrounding tissues, and eventually link some of these separate inflammatory units together (deShazo, 1982).

2.5.2 Diagnostic methods

Perhaps some of the most important clues for detecting asbestosis come from the occupational history of a patient. Having noted that a patient has been exposed to asbestos, a medical practitioner can then look for the following symptoms (Parkes, 1982):

(a) abnormal physical signs. The most important sign is persistent, bilateral, basal late-inspiratory crackles. Clubbing of fingers and toes, breathlessness, and cyanosis may occur in some patients but are generally not reliable signs;

(b) abnormalities in pulmonary function tests which show a restrictive lung defect with reduced gas transfer, for example, a decrease in carbon monoxide diffusing capacity; and

(c) radiographic abnormalities. This usually involves fine reticular shadowing at the lower lobes of the lung in the chest X-ray. However, there is usually some disagreement in the visual interpretation of the X-ray film.

Lung biopsy, bronchioalveolar lavage, gallium-67 scanning and other computerized tomography, and blood serum tests can all be added for especially difficult cases or as extra confirmatory evidences.

It is important to note that none of these clinical signs by itself is sufficient to indicate the existence of asbestosis. The use of lung sound may therefore be a valuable addition to the range of techniques available for examination of patients exposed to asbestos dusts.

2.5.3 Remarks: prognosis and prevention

The progression of asbestosis depends very much on the individual. Generally, the progression will be slow and in some cases the patient's condition may cease to deteriorate. However, the corresponding risk of developing complications such as lung cancers and other pulmonary disorders also increases for these patients (Parkes, 1982). Fortunately, nowadays very few people who have been exposed to asbestos dusts will develop asbestosis because the better hygiene standards imposed in this country will reduce the quantity of fibres inhaled.

Since there is no known drug that can arrest or retard the progress of asbestosis, prevention is the only answer for employees who are working with asbestos or asbestos-related products. The reports by the Health and Safety Executive (1979) list a number of recommendations for storage and disposal of

asbestos, conditions of a factory using asbestos and its products, level of airborne dust within a factory, and individual safety measures, which if satisfactory complied with, should provide adequate protection for each individual employee.

References

[01] Anderson, K., Luk, A., Macleod, J.E.S. and Moran, F. (1986). "The application of pattern recognition and signal processing techniques in the diagnosis of asbestosis". **Thorax**, **41**, 715.

[02] Basmajian, J.V. (1982). **Primary Anatomy** (Eighth Edition), Williams and Wilkins, Baltimore.

[03] Bunin, N.J. and Loudon, R.G. (1979). "Lung sound terminology in case reports". **Chest**, **76**, 690-692.

[04] Cooke, W.E. (1924). "Fibrosis of the lungs due to inhalation of asbestos dust". **British Medical Journal II**, 147.

[05] Cooke, W.E., McDonald, S. and Oliver, T. (1927). "Pulmonary asbestosis". **British Medical Journal II**, 1024-1027.

[06] Crofton, J. and Douglas, A. (1975). **Respiratory Diseases** (Second Edition), Blackwell Scientific Publications, Oxford.

[07] Crystal, R.G., Gadek, J.E., Ferrans, V.J., Fulmer, J.D., Line, B.R. and Hunninghake, G.W. (1981). "Interstitial lung disease: current concepts of pathogenesis, staging and therapy". **The American Journal of Medicine**, **70**, 542-568.

[08] deShazo, R.D. (1982). "Current concepts about the pathogenesis of silicosis and asbestosis". **The Journal of Allergy and Clinical Immunology**, **70**, 41-49.

[09] Doll, R. and Peto, J. (1985). **Asbestos: Effects on Health on exposure to asbestos**, Health and Safety Commission, HM Stationery Office, London.

[10] Forgacs, P. (1978). **Lung Sounds**, Bailliere Tindall, London.

[11] Forgacs, P., Nathoo, A.R. and Richardson, H.D. (1971). "Breath sounds". **Thorax**, **26**, 288-295.

[12] Gadek, J.E., Hunninghake, G.W., Zimmerman, R.L. and Crystal, R.G. (1980). "Regulation of the release of alveolar macrophage-derived neutrophil chemotactic factor". **American Review of Respiratory Disease**, **121**, 723-733.

[13] Gadek, J.E., Hunninghake, G.W., Schoenberger, C.I., Fells, G. and Crystal, R.G. (1981). "Pulmonary asbestosis and idiopathic pulmonary fibrosis: pathogenetic parallels". **Chest**, **80** (Supplement), 63S-64S.

[14] Health and Safety Executive (1979). **Asbestos: Final report of the Advisory Committee**, HM Stationery Office, London.

[15] Keogh, B.A. and Crystal, R.G. (1982). "Alveolotis: the key to the interstitial lung disorders (editorial)". **Thorax**, **37**, 1-10.

[16] Kligfield, P. (1981). "Laennec and the discovery of mediate auscultation". **The American Journal of Medicine**, **70**, 275-278.

[17] Last, R.J. (1984). **Anatomy: Regional and Applied** (Seventh Edition), Churchill Livingstone, Edinburgh.

[18] Merewether, E.R.A. and Price, C.V. (1930). **Report on Effects of Asbestos Dust on the Lungs and Dust Suppression in the Asbestos Industry**, HM Stationery Office, London.

[19] Michaels, L. and Chissick, S.S. (1979). **Asbestos: Properties, Applications, Hazards**, **1**, John Wiley and Son, Belfast.

[20] Morgan, W.K.C. and Seaton, A. (1984). **Occupational Lung Diseases** (Second Edition), W.B. Saunders Company, Philadephia.

[21] Murphy, R.L.H. (1981). "Auscultation of the lung: past lessons, future possibilities". **Thorax**, **36**, 99-107.

[22] Murphy, R.L.H., Holford, S.K. and Knowler, W.C. (1977). "Visual lung sound characterization by time-expanded waveform analysis". **The New England Journal of Medicine**, **296**, 968-971.

[23] Murphy, R.L.H. and Holford, S.K. (1980). "Lung sounds". **Basics of RD**, **8**, (4), 1-6.

[24] Nath, A.R. and Capel, L.H. (1974). "Inspiratory crackles - early and late". **Thorax**, **29**, 223-227.

[25] Noro, L. (1968). "Yant memorial lecture: occupational and non-occupational asbestosis in Finland". **American Industrial Hygiene Association Journal**, **29**, 195-201.

[26] Parkes, W.R. (1982). **Occupational Lung Disorders** (Second Edition), Butterworths, London.

[27] Selikoff, I.J. and Lee, D.H.K. (1978). **Asbestos and Disease**, Academic Press, New York.

[28] Spencer, H. (1985). **Pathology of the lung** (Fourth Edition), Pergamon Press, Oxford.

[29] Timbrell, V. (1965). "The inhalation of fibrous dusts". **Annals of the New York Academy of Sciences**, **132**, 255-273.

[30] Urquhart, R.B., McGhee, J., Macleod, J.E.S., Banham, S.W. and Moran, F. (1981). "The diagnostic value of pulmonary sounds: a preliminary study by computer-aided analysis". **Computers in Biology and Medicine**, **11**, 129-139.

Preprocessing

This work was supported by the Croucher Foundation, Hong Kong.

Chapter 3: Preprocessing

Summary

In this chapter, the development of the Fourier transform from its continuous forms to the discrete form is briefly mentioned. The two main problems, those of aliasing and leakage, with the discrete Fourier transform are discussed. One of the applications of the Fourier transform is in estimating a spectrum. This application is reviewed and the problem associated with the variance of the estimation is outlined. This leads to the development of weighted overlapped segment averaging algorithms which reduce both the leakage and the variance of the estimate. Three of these algorithms, those of Welch, of Carter and Nuttall and of Yuen, are briefly described.

3.1 Introduction

In some very simple template matching applications, all input measurements (or signals) can immediately be used for identification purposes. However, in most practical situations, some sort of preprocessing or conditioning will be required to enable the input measurements to be usable by other operations. This is necessary because the input measurements may be corrupted by noise, distorted by the input transducer, and/or affected by interference from other external sources. In other situations, the number of input measurements may not be constant: an example arises (assuming constant sampling rate) from the fact that the

duration of a breath cycle will vary from individual to individual or even from breath to breath for one individual. In communications engineering, some of these contaminations can be reduced by increasing the power of the source of the input measurements or through the use of different modulation and filtering techniques. Unfortunately, in lung sound analysis, there is not much one can do: one cannot ask the patient to breath harder and harder. Besides, if the patient breathes too hard, additional adventitious sounds will be generated and superimposed on the natural lung sound (Forgacs, 1978). Thus, it is clear that a lot of work has to be done at the "receiving" end in lung sound analysis. To achieve this, ideally the coupling between the chest wall and the input transducer should be matched. Then, the input measurements have to be filtered to remove the high frequency noise (section 3.2.4). Afterwards, signal analysis techniques can be applied to detect possible periodicities in the input measurements. Some of these techniques can operate directly on the input measurements in the time domain and are potentially useful, but in time domain analysis it is often difficult to form suitable features for further analysis. This study therefore concentrates on frequency domain analysis. Spectrum estimation based on the discrete Fourier transform (DFT) is one of the most commonly used methods of frequency domain analysis. Other spectrum estimation techniques, such as the maximum entropy method, are described in a review paper by Kay and Marple (1981).

In the next section, the Fourier transform and its

discrete version will be reviewed briefly. The two major problems, namely leakage and aliasing, associated with the DFT will also be introduced. In section 3.3, spectrum estimation based on the discrete Fourier transform will be discussed in greater detail and the method known as weighted overlapped segment averaging (WOSA) will be described.

3.2 Fourier Transform (FT)

3.2.1 Introduction

About a hundred and eighty years ago, Jean Baptiste Joseph de Fourier proposed his famous Fourier analysis on any arbitrarily shaped function. Essentially, his theorem can be restated as: any arbitrary periodic function $x(t)$ with a period equal to T can be approximated by the Fourier series (Lynn, 1973).

If $x(t)$ is a non-periodic (or aperiodic) function, it can be expressed in terms of the Fourier transform $X(\omega)$. The relevant equation pair is

$$x(t) = (1/2\pi)\int_{-\infty}^{\infty} X(\omega)\, \exp(j\omega t)\, d\omega \qquad (1)$$

where

$$X(\omega) = \int_{-\infty}^{\infty} x(t)\, \exp(-j\omega t)\, dt. \qquad (2)$$

The Fourier transform $X(\omega)$ is sometimes known as a frequency density function. In addition, equations (1) and (2) illustrate

the dualism between the time domain and the frequency domain. Equation (1) shows that a time varing function x(t) is composed of an infinite orthogonal exponential series defined in the frequency domain. The reverse is denoted in equation (2).

With the rapid advancement in digital computing facilities in the past few decades, input measurements are now usually stored and processed digitally. This led to the development of the discrete Fourier transform (DFT) which is suitable for sampled data or measurements.

3.2.2 Discrete Fourier Transform (DFT)

Assume that outside the region (-T/2, T/2) the function x(t) is zero and that x(t) is sampled n times within the region [-T/2, T/2] at equal sampling intervals. Let x_0, x_1, ..., x_{n-1} be the n sampled measurements. The discrete version of the Fourier transform (equations 1 and 2) can be defined as

$$x_k = (T/n) \sum_{m=0}^{n-1} X_m \exp(j2\pi mk/n) \qquad (3)$$

and

$$X_k = \sum_{m=0}^{n-1} x_m \exp(-j2\pi mk/n) \qquad (4)$$

where X_k is the k-th Fourier coefficient of x(t) and k = 0,

1, ..., (n - 1). Each of these X_k represents a certain component (in the frequency domain) of the original sampled measurements (in the time domain). Equation (3) is usually referred as the n-point DFT and equation (4) is the corresponding n-point inverse discrete Fourier transform (IDFT). Yuen and Fraser (1979) have shown that given an n-point DFT of a function, a Fourier series can always be reconstructed such that the function is exactly recovered at the n sampling instants.

Sampling frequency is important when the DFT is used. From the sampling theorem, the "sampling interval" Δt between any two sampled measurements should be smaller than $1/2f_{max}$, where f_{max} is the highest frequency of interest in a given problem. Thus, the following relationships can be written

$$\Delta t = T/n < 1/2f_{max} \quad \text{or} \quad n > 2\, f_{max}\, T. \tag{5}$$

Equation (4), if used directly to compute an n-point DFT, requires n^2 mutiplications, i.e. n mutiplications per X_k. A careful inspection of the exponential terms will reveal that the same product $x_m \exp(-j2\pi mk/n)$ is formed many times for different combination of m and k. This fact was noticed by Cooley and Tukey (1965) who proposed the fast Fourier transform (FFT) algorithm, which reduces such repetitive calculation of products from n^2 to $n \log_2 n$. The idea is to divide and reshuffle the input measurements into a number of smaller subgroups, transform each subgroup individually, and combine the results to produce the DFT for the n sampled measurements. A number of verified FFT

implementations have been published in a symposium entitled "Programs for digital signal processing" which was edited by the digital signal processing committee (1979).

3.2.3 Leakage

Leakage is an inherent problem with the discrete Fourier transform. The finite number of sampled measurements in the DFT means that the function x(t) is truncated abruptly at some point in the time domain. This causes the k-th Fourier coefficient (k = 0, 1, 2, ..., n - 1) in the frequency domain to oscillate, i.e. the values of all X_p, where $p \in (-\infty, \infty)$ and $p \neq k$, are in general affected. Hence X_k "leaks out" into the neighbouring Fourier coefficients (Yuen and Fraser, 1979). Thus if there is a peak at the k-th Fourier coefficient then, instead of only one peak, a series of peaks of varying sizes will be observed because of the leakage. It follows that two very close peaks in the frequency domain may be masked by this leakage effect. Unfortunately, leakage cannot be prevented; it can only be reduced. Techniques to achieve this are known as windowing techniques. Essentially windowing reduces the order of discontinuity at the two ends (or boundaries) of the truncated time function. Harris (1978) reviews various windows for the DFT, such as the minimum 4-sample (or term) Blackman-Harris and 4-sample Kaiser-Bessel windows. (See also the paper by Nuttall, 1981.)

3.2.4 Aliasing

Another important problem associated with DFT is aliasing. This is due to inability to distinguish time functions having high frequency components greater than f_{max}. In effect, sampling causes these high frequency components to "fold back" into (or overlap with) the low frequency components within the range $[0, f_{max}]$ (Lynn, 1973). Therefore the input signal has to be filtered by a low pass filter to remove these high frequency components (often due to the undesirable contaminations mentioned in section 3.1), and then sampled at a frequency which must be higher than $2f_{max}$ so that the time function can be reconstructed (section 3.2.2).

Unfortunately, as mentioned in section 3.2.3, leakage cannot be prevented. Thus, even if it is possible to filter away all the higher frequency components actually present in x(t), there are still some high frequency components due to leakage (which can produce both the low and the high frequency components). Therefore, a good window is necessary to reduce the leakage and hence reduce this combined effect of leakage and aliasing. Leakage is thus one reason why in pratice the sampling frequency must usually be considerably greater than the theoretical minimum of $2f_{max}$ (Yuen and Fraser, 1979).

3.3 Spectrum Estimation

3.3.1 Introduction

One of the many applications of the Fourier transform is in spectrum estimation (i.e. estimation of the power at particular frequencies). Spectra are of great importance in many fields including quantum mechanics and the study of the internal structures of different elements as well as pattern recognition. Their usefulness in pattern recognition, stems from the ease with which features can be generated from a spectrum and has led to their use in a number of applications such as speech recognition (Fu, 1982). Robinson (1983) in his book (appendix 9, pages 345-407) has given an excellent historical perspective of pioneering work in spectrum estimation and related topics. As mentioned in section 3.1, only spectrum estimation techniques that are based on the discrete Fourier transform will be reviewed here.

It must however be mentioned that Fourier analysis is only one of several methods of spectrum estimation (Kay and Marple, 1981). One other very commonly used spectrum estimation technique is the maximum entropy method (MEM), introduced by John Burg in 1967 (Robinson, 1983), which is based on extrapolating a segment of a known autocorrelation function (section 3.3.2.1) so that the entropy in the time domain is maximized. Interested readers can refer to a book edited by Smith and Grandy (1985) which gives a collection of very useful papers on this subject

and its applications. The MEM technique was investigated by Urquhart (1983) in relation to lung sounds. However his preliminary study indicated that it is very difficult to select the right model order, as is essential in MEM (Kay and Marple, 1981).

3.3.2 Spectrum estimation based on the Fourier transform

3.3.2.1 The continuous case

The total energy of an aperiodic function x(t) is related to its Fourier transform $X(\omega)$ by Parseval's energy theorem which states that

$$\int_{-\infty}^{\infty} |x(t)|^2 \, dt = \int_{-\infty}^{\infty} |X(\omega)|^2 \, d\omega \qquad (6)$$

where $|X(\omega)|^2 = S(\omega)$ is sometimes referred to as the energy spectral density and $|.|$ is the absolute operator. Equation (6) states that the total energy of the time domain signal is equal to the total energy in the frequency domain (Kay and Marple, 1981).

For a periodic function x(t) with a period equal to T, the (power) spectrum $P(\omega)$ (sometimes referred as the periodogram) of x(t) can be found by

$$P(\omega) = \lim_{T \longrightarrow \infty} E \{ (1/T) \, S(\omega) \}. \qquad (7)$$

Another indirect way of finding the spectrum of x(t), related with the autocorrelation function $r_{xx}(\tau) = \{ x(t+\tau)x^*(t) \}$ (where * is the conjugate operator), is given by the Wiener-Khinchin relations

$$P(\omega) = \int_{-\infty}^{\infty} r_{xx}(\tau) \exp(-j\omega\tau)\, d\tau \qquad (8)$$

and

$$r_{xx}(\tau) = \int_{-\infty}^{\infty} P(\omega) \exp(j\omega\tau)\, d\tau. \qquad (9)$$

3.3.2.2 The discrete case

Until Cooley and Tukey (1965) proposed their fast Fourier transform algorithm, the estimated spectrum $\hat{P}(\omega)$ was usually found by using the discrete version of the Wiener-Khinchin relations because of the inherent n^2 product computation in the DFT. Again, x(t) is assumed to be defined over [-T/2, T/2] (section 3.2.2). The discrete version is sometime known as the Blackman-Tukey algorithm and can be written as

$$\hat{P}_k = (1/T) \sum_{m=0}^{p-1} \hat{r}_{xx}(m) \exp(-j2\pi mk/n) \qquad (10)$$

where $p \leq n-1$ and the autocorrelation function $\hat{r}(m)$ is estimated as

$$\hat{r}(m) = (1/n) \sum_{q=0}^{n-m-1} x_{q+m} x_q^* \qquad (11)$$

where x_k (k = 0, 1, ..., n - 1) is the sampled measurment as defined in section 3.2.2.

Since Cooley and Tukey introduced the FFT, spectra have usually been estimated using the direct method which can be written as

$$\hat{P}_k = (1/T) \, |X_k|^2 = (1/T) \, \|X_k\|^2 \qquad (14)$$

where $\|.\|$ is the modulus operator.

One of the main problem associated with the above two estimation methods is the variance of the estimated spectrum. It has been shown (Yuen and Fraser, 1979, pages 72-73) that the variance can be as much as $P(\omega)$ itself. One way of reducing the variance is to window the sampled measurements before the estimation. Another method of reducing the variance is known as weighted overlapped segment averaging which is more suitable when there is a large number of sampled measurements and will be introduced in the next section.

3.3.3 Weighted Overlapped Segment Averaging Spectral Estimation

This method can be viewed as a development of an idea due to Welch (1967), who proposed a spectrum estimation method which reduces the variance of the estimated spectrum and is extremely useful for large numbers of input measurements. His idea is to divide the n sampled measurements into p segments,

window each segment by a linear window (such as the minimum 4-sample Blackman-Harris window mentioned in section 3.2.3), and finally average the magnitude-square of the p segments to produce an averaged spectrum. Both spectral leakage and variance are thereby reduced because of the windowing operation and the averaging operation respectively.

Welch's idea was considerable extended by Carter, Nuttall and Yuen in a series of papers (Yuen, 1977; Yuen, 1978; Yuen, 1979; Nuttall and Carter, 1980; Carter and Nuttall, 1980; Yuen, 1983). The essential difference between the algorithms of Carter and Nuttall and of Yuen is in whether the segments should be overlapped and windowed prior to being transformed into the frequency domain by the FFT method. Later in chapter 7, it will be shown experimentally that for the lung sound data, both methods produced very similar results. Since both methods will be deployed in chapter 7, a brief outline of both methods is given in the following two paragraphs.

Yuen's algorithm (1983) is very similar to Welch's. Essentially his method is to divide the n sampled measurements into p non-overlapping segments, Fourier transform each segment, average the squared magnitudes of the p transformed segments, and finally apply a quadratic (or lag) window to the averaged spectrum.

The algorithm proposed by Carter and Nuttall (1980) is more complicated but is claimed by the authors to produce

slightly better results than Yuen's method (Yuen, 1983) for processes with "large dynamic range" spectra. Their idea is to divide the n sampled measurements into q overlapped segments, window each overlapped segment using a linear (or tapering) window (such as the minimum 4-sample Blackman-Harris window), Fourier transform each windowed segment and average the squared magnitudes of the q transformed segments. This averaged spectrum is then transformed back into a lag domain via the FFT (equation 13), multiplied by a lag window and then returned to the frequency domain by another FFT operation (equation 14). (Note that the last three steps are essentially the Blackman-Tukey method of calculating the spectrum).

3.4 Remarks

One thing worth mentioning at this stage is that the weighted overlapped segment averaging (WOSA) techniques enable a large number of input measurements, say n', in the time domain to be compressed into a relatively smaller number of measurements, say n, in the frequency domain. This data compression technique is very useful in the lung sound signal analysis described in chapter 7.

Other advantages of using the WOSA techniques when the number of input measurements n is large are (a) the possibility of reducing computation time and (b) the amount of main storage, provided the number of segments p is less than the square-root of n (Welch, 1967). Furthermore, Welch (1967) has shown that a

small reduction in variance can be achieved when overlapping is used in the WOSA methods. It has also been shown (Nuttall and Carter, 1980; Nuttall, 1981) that the amount of overlapping is dependent on the particular window employed in the WOSA algorithms. Finally, the question of whether linear or quadratic windowing should be used is largely academic. Nevertheless, Mathews and Youn (1984) have shown that both linear and quadratic windowing in the WOSA methods provide asymptotically the same leakage suppression, under the assumption that segments relatively far apart are uncorrelated.

References

[01] Cooley, J.W. and Tukey, J.W. (1965). "An algorithm for the machine calculation of complex Fourier series". **Mathematics of Computation**, **19**, 297-301.

[02] Carter, G.C. and Nuttall, A.H. (1980). "On the weighted overlapped segment averaging method for power spectral estimation". **Proceedings of the IEEE**, **68**, 1352-1354.

[03] Digital Signal Processing Committee, Editors (1979). **Programs for Digital Signal Processing**. IEEE Press, New York.

[04] Forgacs, P. (1978). **Lung Sounds**. Bailliere Tindall, London.

[05] Fu, K.S. (1982), Edited. **Applications of Pattern Recognition**, CRC Press, Inc., Florida.

[06] Harris, F.J. (1978). "On the use of windows for harmonic analysis with the discrete Fourier transform". **Proceedings of the IEEE**, **66**, 51-83.

[07] Kay, S.M. and Marple, S.L. (1981). "Spectrum analysis: a modern perspective". **Proceedings of the IEEE**, **69**, 1380-1419.

[08] Lynn, P.A. (1973). **An Introduction to the Analysis and Processing of Signals**. Macmillan, London.

[09] Mathews, V.J. and Youn, D.H. (1984). "Spectral leakage suppression properties of linear and quadratic windowing". **IEEE Trans. on Acoustics, Speech, and Signal Processing**, **ASSP-32**, 1092-1095.

[10] Nuttall, A.H. (1981). "Some windows with very good sidelobe behaviour". **IEEE Trans. on Acoustics, Speech, and Signal Processing**, **ASSP-29**, 84-91.

[11] Nuttall, A.H. and Carter, G.C. (1980). "A generalized framework for power spectral estimation". **IEEE Trans. on Acoustics, Speech, and Signal Processing**, **ASSP-28**, 334-335.

[12] Robinson, E.A. (1983). **Multichannel Time Series Analysis with Digital Computer Programs** (Second Edition). Goose Pond Press, Houston.

[13] Smith, C.R. and Grandy, W.T., Edited (1985). **Maximum-Entropy and Bayesian Methods in Inverse Problems**. D. Reidel Publishing Company, Dordrecht.

[14] Urquhart, R.B. (1983). "Some new techniques for pattern recognition research and lung sound signal analysis". Ph.D. Thesis, University of Glasgow.

[15] Welch, P.D. (1967). "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms". **IEEE Trans. on Audio and Electroacoustics**, **AU-15**, 70-73.

[16] Yuen, C.K. (1977). "A comparison of five methods for computing the power spectrum of a random process using data segmentation". **Proceedings of the IEEE**, **65**, 984-986.

[17] Yuen, C.K. (1978). "Quadratic windowing in the segment averaging method for power spectrum computation". **Technometrics**, **20**, 195-200.

[18] Yuen, C.K. (1979). "Comments on modern methods for spectrum estimation". **IEEE Trans. on Acoustics, Speech, and Signal Processing**, **ASSP-27**, 298-299.

[19] Yuen, C.K. (1983). "Comments on spectral estimation using combined time and lag weighting". **Proceedings of the IEEE**, **71**, 535-536.

[20] Yuen, C.K. and Fraser, D. (1979). **Digital Spectral Analysis**. CSIRO-Pitman, London.

Mapping

This work was supported by the Croucher Foundation, Hong Kong.

Chapter 4: Mapping

Summary

Mapping techniques are briefly surveyed in this chapter. In particular, four commonly used linear mapping algorithms, those of Karhunen-Loeve, of Kittler-Young, of Fisher and of Fukunaga-Mantock are introduced. Modifications to the Fukunaga-Mantock transformation are described. The difficulties in choosing the best linear transformation is also outlined.

4.1 Introduction

In chapter 3, it has been shown that a datum comprising a variable number n'' of input measurements (in the time domain) can be preprocessed (e.g. using the WOSA technique described in section 3.3.3) to form a datum with a fixed number n' of transformed measurements (in the frequency domain). In some applications, these n' transformed measurements can readily be interpretated and be used as features for, say, a discrimination (or classification) task. However, in some problems (including the one investigated in this thesis), although these transformed measurements are meaningful, the dimensionality n' of the datum is still too large for practical pattern recognition. Therefore, it is sensible to somehow compress or reduce the dimensionality of the data from an n'-space to an n-dimensional (sub)space (where $n < n'$) without losing too much information. It is certainly very useful if the dimensionality n of the compressed

data is less than or equal to 3 so that the original set of N n'-dimensional data could be displayed in some conventional device, such as a visual display unit or a plotter, for subsequent interpretation such as analysis or classification of the set of data.

Two principal families of methods, namely feature selection and feature extraction, are available for reducing the dimensionality of data. In feature selection a subset comprising n of the original n' measurements is selected in such a way that the redundant or less useful measurements can be discarded without a significant loss in the original information. Very often feature selection is achieved by optimising a criterion function that is related to classification error. Such criterion functions can be based on probabilistic distance measures (for example Mahalanobis distance), dependence measures or Euclidean distance. Global optimisation by exhaustive search is rarely feasible if n' is large, because the criterion function has to be evaluated $\binom{n'}{n}$ times. For this reason, many suboptimal "top down" and "bottom up" approaches have been proposed which involve much less computation. Devijver and Kittler (1982) give an intensive treatment on this subject and it will not be discussed further in this chapter.

In feature extraction, on the other hand, all the n' measurements are utilized and a new transformed space is formed through the use of some transformation operations. A subset of n variables in the transformed space is then extracted or chosen

(Kittler, 1975; Urquhart, 1983). Let **x** be a random n'-dimensional feature vector (here it is assumed that all the n' measurements are taken as features). A feature extraction algorithm then consists of a mapping or a transformation F, where

$$\mathbf{y} = F(\mathbf{x}) \qquad (1)$$

such that the resulting transformed random feature vector **y** will be of lower dimensionality than **x**. There are a number of mapping techniques which can achieve this goal. Mapping is usually dependent on the optimization of a certain criterion function H such as the entropy of a system. Some mapping techniques are linear while others are nonlinear: within each category, some of the algorithms are iterative while others are non-iterative. Iterative algorithms involve iterative optimization of a criterion function which compares high and low dimensional representations of the data. In non-iterative mapping, on the other hand, the form of F can usually be calculated directly and will therefore be unique. See the review by Urquhart (1983).

Non-iterative nonlinear mappings may have a rather complicated form for the mapping function F and, in some cases, may require a complete knowledge of the underlying distributions of each measurement (Young and Calvert, 1974, pages 255-258). These mappings are therefore rarely used in any application. On the other hand, the form of F in some iterative nonlinear mappings may not be possible to determine explicity. One example is Sammon's nonlinear mapping algorithm (1969). There are a

number of variants of this method (Calvert and Young, 1969; Kruskal, 1971; Urquhart, 1983; Wang, 1983). Unfortunately, these methods depend on the choice of the initial configuration of the subspace, and it is necessary to perform the whole mapping procedure from the beginning every time a new feature vector becomes available. Consequently, these methods will not be of particular use for the proposed system. They will however remain as a useful research tool for analysing the structure of a data set.

For non-iterative linear mapping algorithms, equation (1) may be rewritten as

$$\mathbf{y} = U^T\mathbf{x} \qquad (2)$$

where T is the transpose operator. The mapping matrix U can usually be evaluated, and its evaluation is usually less computationally demanding than any of the above mentioned data reduction methods. Therefore, these algorithms are often employed in many different applications (Fukunaga and Koontz, 1970; Kulikowski, 1970; Sammon, 1970). Because of their importance, a brief review is presented in section 4.2.

4.2 Linear Mapping

4.2.1 Introduction

Linear mapping has been the subject of research and

application for a number of decades in various disciplines, such as statistics, communication theory and pattern recognition. In essence, it is an expansion of the random vector $\mathbf{x}$ in terms of the eigenvectors of a matrix R, that is the transformation matrix U is formed from n eigenvectors $\mathbf{u}_j$ (j = 1, 2, ..., n), each of dimensionality n', of R (usually associated with the n largest eigenvalues λ_j). Each of the n eigenvectors is usually normalized so that its magnitude is unity, i.e. $\|\mathbf{u}_j\| = 1$, and the eigenvectors are made uncorrelated, i.e. $\mathbf{u}_j^T\mathbf{u}_j = 1$ and $\mathbf{u}_k^T\mathbf{u}_j = 0$ where $k \neq j$ (i.e. the n eigenvectors are orthonormal to each other). This matrix R can be a between class scatter matrix (or class conditional between class matrix) S_B or a within class scatter matrix (or class-conditional covariance matrix) S_W or a total scatter (i.e. covariance) matrix S_T or a combination of them. Let $\mathbf{x}_i$ be a prototype sample (where i = 1, 2, ..., N and N = number of prototype samples available) and let $\mathbf{x}_p^q$ be a prototype sample from class ω_q (where p = 1, 2, ..., N_q, q = 1, 2, ..., c, N_q = number of prototype samples in class ω_q and c = number of classes) both of dimemsionality n'. Then the three scatter matrices can be estimated as:

$$S_T = S_B + S_W \tag{3a}$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\{\mathbf{x}_i - \mathbf{m}\}\{\mathbf{x}_i - \mathbf{m}\}^T \tag{3b}$$

$$S_B = \sum_{q=1}^{C} p(\omega_q)\{m_q - m\}\{m_q - m\}^T \qquad (4)$$

$$S_W = \sum_{q=1}^{C} \frac{p(\omega_q)}{N_q - 1} \sum_{p=1}^{N_q} \{x_p^q - m_q\}\{x_p^q - m_q\}^T \qquad (5)$$

$\mathbf{m}$ = estimated total mean

$$= \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

$\mathbf{m}_q$ = estimated mean for class ω_q

$$= \frac{1}{N_q} \sum_{p=1}^{N_q} \mathbf{x}_p^q$$

$p(\omega_q)$ = estimated a priori probability of class ω_q

$= N_q / N.$

The following subsections will introduce some commonly used linear mapping algorithms. (The derivation of the following transformation will not be shown but can be found in Devijver and Kittler, 1982, chapter 9, or in the original papers).

4.2.2 Karhunen-Loeve (K-L) transformation

This is perhaps one of the most commonly used linear mapping algorithms and is also referred as the principal components transformation. In the generalised version of Chien and Fu (1968) the algorithm assumes that the prototype samples have been normalized so that the estimated total mean is a zero vector, i.e. $\mathbf{m} = \mathbf{0}$. The transformation matrix U is then constructed using the n uncorrelated eigenvectors of $R = S_W$ associated with the n largest eigenvalues of S_W, i.e.

$$U = [u_1, u_2, \ldots, u_n]^T \qquad (6)$$

where u_j is associated with the j-th largest eigenvalues of S_W. It has been shown (Gower, 1966) that these uncorrelated (or orthonormal) eigenvectors define a coordinate system that is optimal in the least square sense. It is of interest to note that the coordinate axes with the largest variances are selected by the K-L transformation, which thus also maximizes the amount of information retained in the lower dimensional space (Devijver and Kittler, 1982). A number of variations of the K-L transformation are possible, for instance, instead of using S_W, the n uncorrelated eigenvectors can be derived from the matrix S_T.

4.2.3 Kittler-Young (K-Y) transformation

Kittler and Young (1973) noticed that the most significant information for classification purposes is usually contained in the matrix S_B rather than S_W or S_T. They proposed an algorithm which will optimally compress the class mean information. To achieve this, both the between-class and within-class scatter matrices are prewhitened by a prewhitening matrix B such that

$$B^T S_W B = I = S'_W \tag{7}$$

and

$$B^T S_B B = S'_B \tag{8}$$

where I is the identity matrix and S'_B and S'_W are respectively the between-class and within-class scatter matrix after the prewhitening operation. In this prewhitening space, S'_W is uncorrelated and with unit variance. It will therefore be invariant under any orthonormal transformation. The authors have shown that the prewhitening matrix is obtained from the n' uncorrelated eigenvectors u'_a (a = 1, 2, ..., n') and the eigenvalues λ'_a of S_W and is given by

$$B = U' \Lambda'^{-1/2} \tag{9}$$

where

$$U' = [u'_1, u'_2, \ldots, u'_{n'}]^T$$

and

$$\Lambda' = \begin{bmatrix} \lambda_1', & 0, & \ldots, & 0 \\ 0, & \lambda_2', & \ldots, & 0 \\ : & : & & : \\ : & : & & : \\ 0, & 0, & \ldots, & \lambda_n', \end{bmatrix}.$$

The transformation matrix U is then formed from B and the uncorrelated eigenvectors g_j (j = 1, 2, ..., n and n = c - 1) of S_B' (equation 8). Note that only (c - 1) non-zero eigenvalues exist because of the singularity of the matrix S_B. Thus, U is given by

$$U = BG = U'\Lambda'^{-1/2}G \qquad (10)$$

where

$$G = [g_1, g_2, \ldots, g_{(c-1)}].$$

Thus, this algorithm compresses the mean information into a feature space of dimensionality (c - 1). (It is worth noting that Babu (1972) obtained the same equation as in equation 10 when applying a probabilistic distance measure on feature selection.)

4.2.4 Fisher (F-S) transformation

Fisher (1936) proposed an algorithm to construct the transformation matrix U by using the (c - 1) uncorrelated eigenvectors u_j (j = 1, 2, ..., c - 1) associated with the (c - 1) non-zero eigenvalues of $R = S_W^{-1}S_B$. (Again, only c - 1 non-

zero eigenvalues exist because of the singularity of S_B.)

When there are only two classes (c = 2), only one non-zero eigenvalue exists, and hence only $\mathbf{u}_1$ can be obtained by the F-S transformation. Sammon (1970) proposed finding a second vector (say $\mathbf{u}_2$ to simplify the notation) from the first derivative of R (with respect to $\mathbf{u}_1$) such that it is orthogonal to $\mathbf{u}_1$. $\mathbf{u}_1$ and $\mathbf{u}_2$ are given by the following equations

$$\mathbf{u}_1 = \alpha_1 S_W^{-1} \Delta \tag{11a}$$

$$\mathbf{u}_2 = \alpha_2 \left[S_W^{-1} - \frac{\Delta^T [S_W^{-1}]^2 \Delta}{\Delta^T [S_W^{-1}]^3 \Delta} [S_W^{-1}]^2 \right] \Delta \tag{11b}$$

where $\Delta = [\mathbf{m}_1 - \mathbf{m}_2]$ and α_1 and α_2 are normalization constant such that both $\mathbf{u}_1$ and $\mathbf{u}_2$ are unit vector. The Sammon's proposed method would define an "optimal discriminant plane". Note that it is not necessary to calculate the between-class scatter matrix. Sammon's algorithm was extended by Foley and Sammon (1975), who have used a similar approach as Sammon to recursively derive n orthogonal eigenvectors, which they refer as "optimal discriminant vectors". However, Longstaff (1985) has pointed out that if the decision surface at the higher dimensional space of the data is a curved hyperplane, the data will not project uniquely on the Fisher axis. He proposed to find the second "radius" vector by transforming the data to a spherical symmetry in the subspace normal to Fisher's first axis.

4.2.5 Fukunaga-Mantock (F-M) transformation

The singularity problem of the between-class scatter matrix S_B is overcome by Fukunaga and Mantock (1983), who have constructed a nonparametric form S_{Bk} of S_B which is guaranteed to be of full rank. The two class case can be written as

$$S_{Bk} = (1/N)\sum_{p=1}^{N_1} w_p(x_p^1 - m_k^2(x_p^1))(x_p^1 - m_k^2(x_p^1))^T$$

$$+ (1/N)\sum_{p=N_1+1}^{N} w_p(x_p^2 - m_k^1(x_p^2))(x_p^2 - m_k^2(x_p^2))^T \qquad (13)$$

where $m_k^a(x_p^q)$ is the mean vector of the k nearest neighbours of x_p^q that belong to class ω_a (a = 1,2, q =1,2 and a $\neq$ q), and w_p is a weighting function. The definition used by Funkunaga and Mantock for the weighting function is

$$w_p = \frac{\min\{ d_1^h(x_p^q - z_k^1),\ d_2^h(x_p^q - z_k^2) \}}{d_1^h(x_p^q - z_k^1) + d_2^h(x_p^q - z_k^2)} \qquad (14)$$

where $d_a^h(x_p^q - z_k^a)$ is the distance between x_p^q and its k-th nearest neighbour z_k^a from class ω_a in the Minkowski metric of order h.

The author now proposes a modification to equation (14) because the weighting function suggested by Fukunaga and Mantock

(1983) is intended to deemphasize prototype samples far away from the classification boundary. Suppose $\mathbf{x}_p^q$ belongs to class ω_1 in figure 4.1 and is close to the k nearest neighbour mean vector of class ω_1; then in equation (14) $d_1^h(\mathbf{x}_p^q - \mathbf{z}_k^1)$ will be small. If equation (14) is used as the weighting function, $(\mathbf{x}_p^q - \mathbf{m}_k^2(\mathbf{x}_p^q))$ in equation (13) will be small and deemphasized because of the minimum operator in equation (14).

Now suppose $\mathbf{x}_p^q$ is an "outlier" far away from both classes. In this case, both the distances $d_1^h(\mathbf{x}_p^q - \mathbf{z}_k^1)$ and $d_1^h(\mathbf{x}_p^q - \mathbf{z}_k^2)$ will have similarly large values and the sample will be emphasized by the weighting function in equation (14). Moreover $(\mathbf{x}_p^q - \mathbf{m}_k^a(\mathbf{x}_p^q))$ will be large for both classes in equation (13). The resulting S_{Bk} matrix may therefore be dominated by these outliers, particularly if the number of prototype samples is small, and hence a subspace with poor class separation may be found. From this simple example, it is clear that equation (14) does not achieve the goal set out by Fukunaga and Mantock. To handle this undesirable property of equation (14), one way is to use a maximum operator in equation (14).

The author also proposes that the distance from sample $\mathbf{x}_p^q$ to the mean of the k-nearest neighbours belonging to class ω_a, $\mathbf{m}_k^a(\mathbf{x}_p^q)$ $(a \neq q)$, rather than the distance to just one point namely the k-th nearest neighbour $\mathbf{z}_k^a$ as in equation (14), should be used in the weighting function. This is desirable because

(a) equation (13) is primarily concerned with the separability between $\mathbf{x}_p^q$ and the local mean vectors of the other

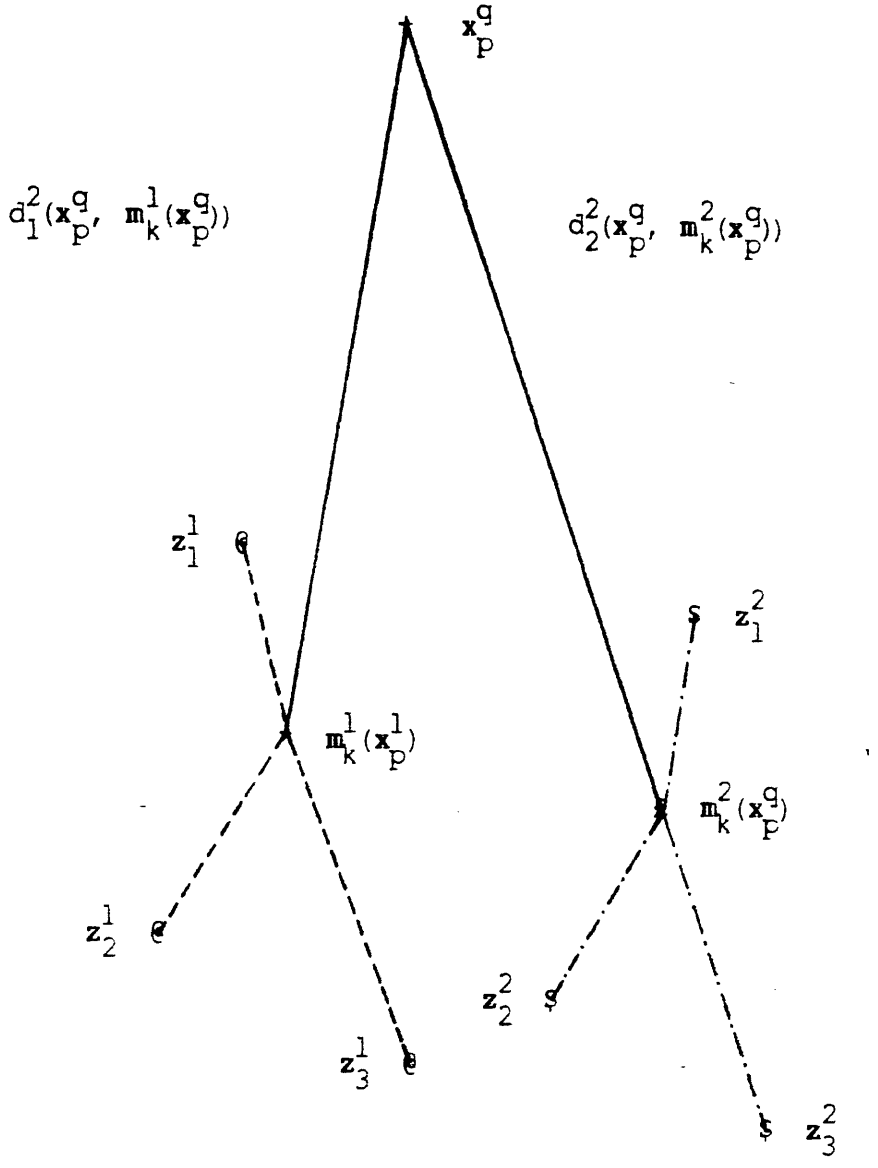


Figure 4.1 A visual interpretation of the weighting function suggested in equation (15) in the text. x_p^q is the p th element in class ω_q to be considered. $m_k^1(x_p^q)$ and $m_k^2(x_p^q)$ are the local mean vectors of classes ω_1 and ω_2 respectively. In the above diagram, k is equal to 3. z_f^a is the f-th nearest neighbour of x_p^q in class ω_a.

classes (c.f. sections 4.2.3 and 4.2.4), and

(b) Short and Fukunaga (1981) have shown that $\mathbf{m}_k^a(\mathbf{x}_p^q)$ is an important parameter of the weighting function in the optimal local distance measure for nearest neighbour classification in a finite sample set.

Therefore, the new weighting function can be written as

$$w'_p = \beta \cdot \frac{\max \{ d_1^h(\mathbf{x}_p^q - \mathbf{m}_k^1(\mathbf{x}_p^q)),\ d_2^h(\mathbf{x}_p^q - \mathbf{m}_k^2(\mathbf{x}_p^q)) \}}{d_1^h(\mathbf{x}_p^q - \mathbf{m}_k^1(\mathbf{x}_p^q)) + d_2^h(\mathbf{x}_p^q - \mathbf{m}_k^2(\mathbf{x}_p^q))} + \gamma \quad (15)$$

where β and γ are two real constants. When $\beta = 1.0$ and $\gamma = 0.0$, the weighting function w'_p has a range between 0.5 to 1.0. (The range can be altered by using the two constants. For example when $\beta = 1.0$ and $\gamma = -0.5$, w'_p has a range similar to equation 14.) Near the decision boundary it has values close to 0.5. As we move further away from the boundary and approach the vicinity of the centres of the two clusters (assuming, for simplicity, unimodal distributions for the two classes), w'_p increases towards 1.0 then decreases slowly after the cluster centre is passed, tending towards 0.5 again at all outlying points. Isometric plots of w_p and w'_p are shown in figures 4.2 and 4.3 for a data set comprising two subsets (classes) in two dimensions, each subset being uniformly distributed in a square region, one on either side of the decision boundary (by symmetry, only half of the decision region of one of the subsets need be shown). The noisiness of w_p compared with w'_p may be due to the

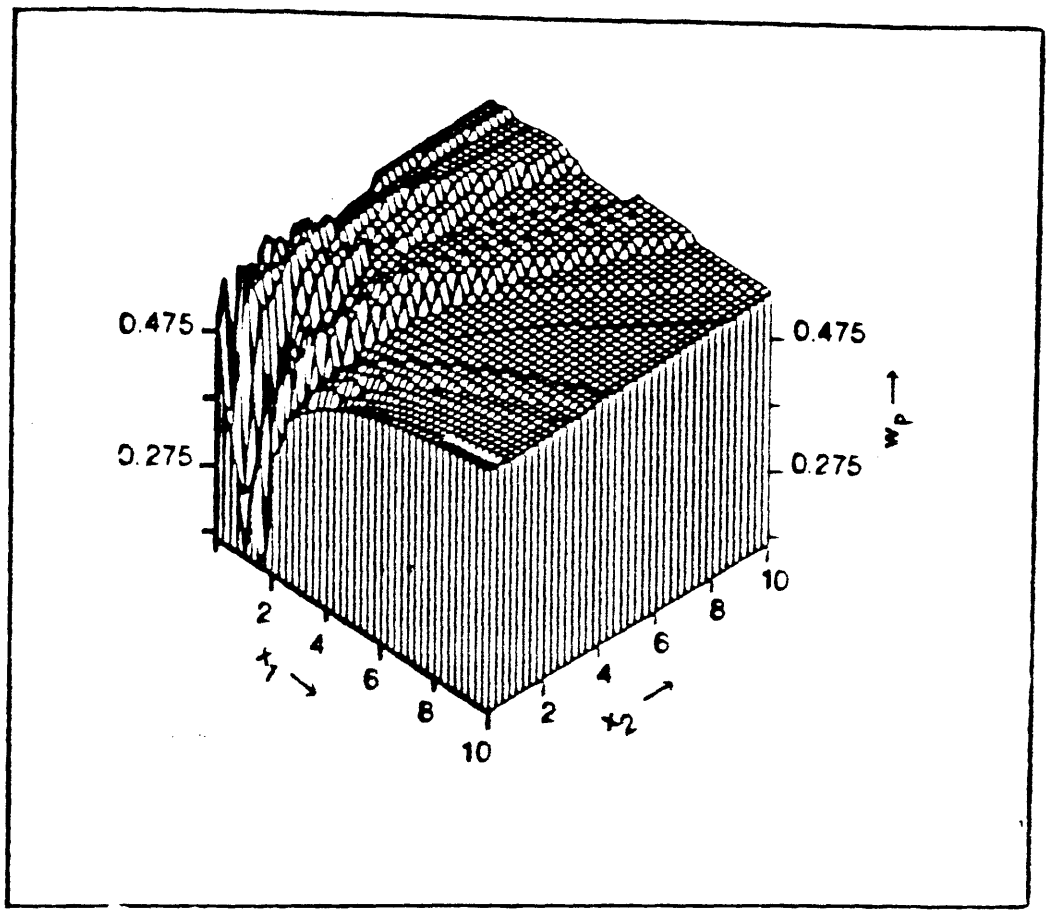


Figure 4.2 Isometric plot of the original weighting function w_p (equation 14) due to Fukunaga and Mantock (1983) for a set comprising two subsets (classes) in 2 dimensions. Each subset is uniformly distributed in a square region. The two squares lie symmetrically one on each side of the decision boundary and the distribution is roughly symmetrical about a line joining the centres of the squares: hence only half of the region on one side of the decision boundary is represented. The square containing this subset is bounded by the lines $x_1 = 0.5$, $x_1 = 1.5$ and $x_2 = \pm 0.5$.

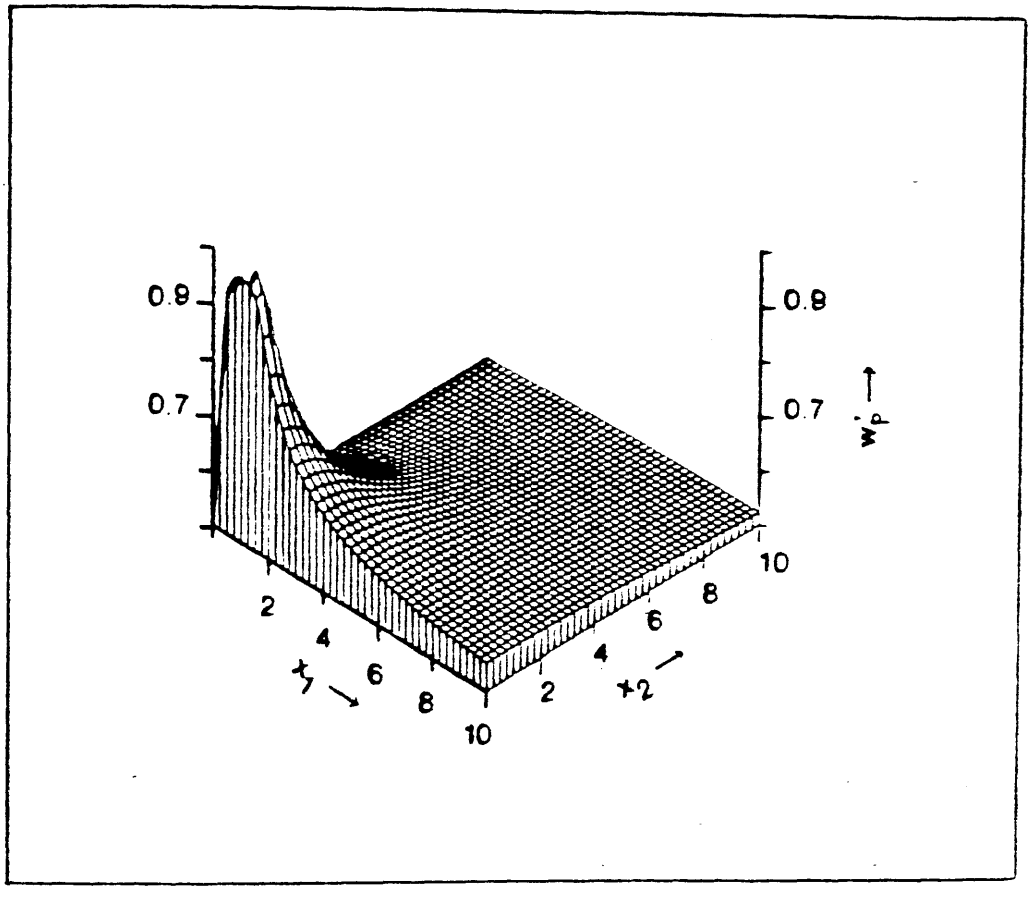


Figure 4.3 Isometric plot of the proposed weighted function w'_p (equation 15) for the same data set as in figure 4.2.

use of $\mathbf{z}_k^a$ alone in equation (14) rather than the mean $\mathbf{m}_k^a(\mathbf{x}_p^q)$ as in equation (15).

From the above description and as illustrated in figures 4.2 and 4.3, the new weighting function w_p' will not only deemphasize samples near to the decision boundary but will also deemphasize all outliers. It will therefore reduce the danger that outliers may dominate the S_{Bk} matrix.

Figures 4.4 and 4.5 are displays obtained by the F-M transformation using respectively w_p and w_p' on a three class, 9-dimensional real data set (with 50 samples per class) extracted from eddy current signals from flaws in heat-exchanger tubing (Macleod, 1982; Macfarlane, 1987). It can be seen that the weighting function in equation (15) has given a better output display and fewer misclassification. Macfarlane has found that this subspace achieves better class separation than the K-L and K-Y transformations. See also chapter 7 for an application of this weighting function to results from the lung sound data.

4.2.6 Remarks

The above four sections have introduced four of the most commonly used linear mapping algorithms. Many variations of these algorithms exist (Urquhart, 1983). The real problem is to select which algorithm to use. Unfortunately, none of the algorithms is universally applicable to all data sets. A classical example is the transformed subspace obtained by K-L and

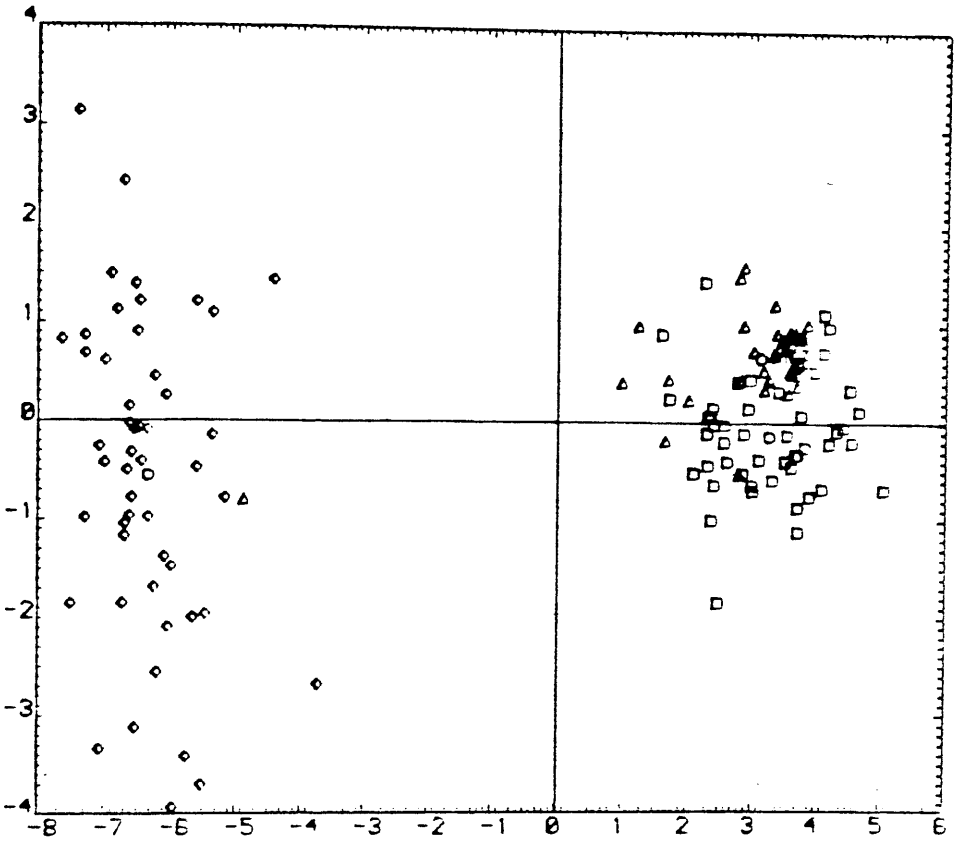


Figure 4.4 Output subspace of nonparametric discriminant analysis using the weighting function w_p of equation (14) on 3 classes of 9-dimensional data, each with 50 samples. Plot symbols △, □, and ◇ denote classes 1, 2 and 3 respectively.

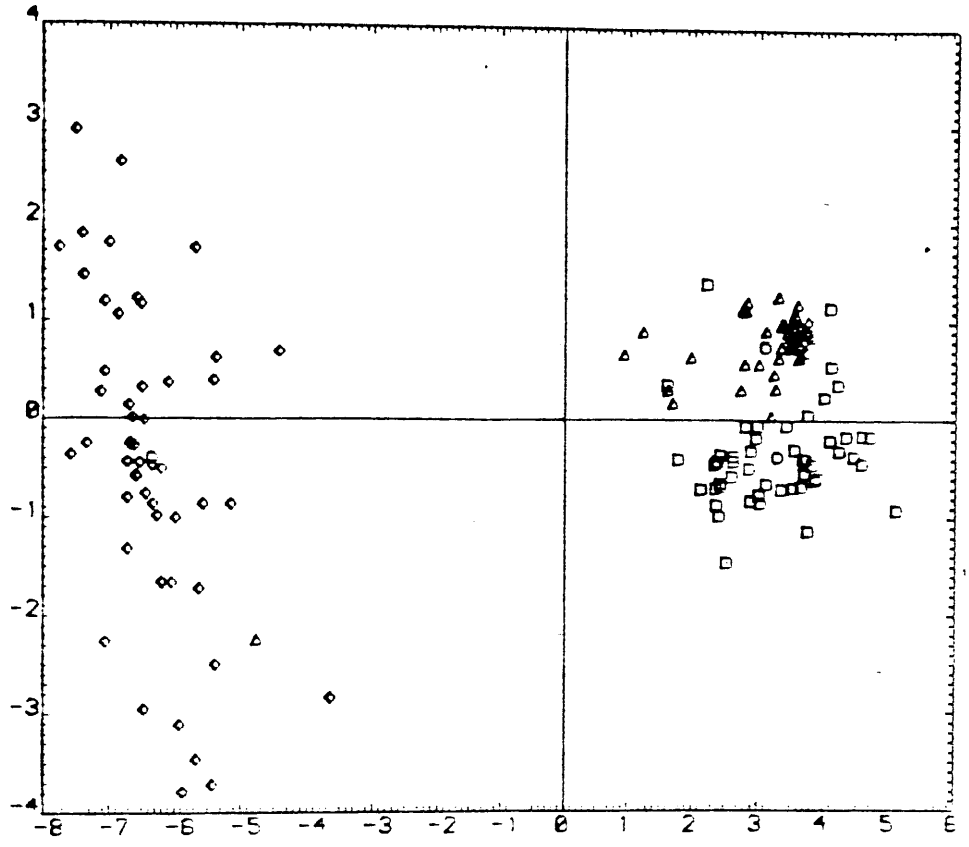


Figure 4.5 Output subspace of nonparametric discriminant analysis using the weighting function w'_p of equation (15) on the same data set as in figure 4.4.

K-Y transformation on two parallel ellipses as in figure 4.6. The K-L transformation will select u_1 as the best axis for mapping because the variance along u_1 is the greatest. However, this axis cannot discriminate between the two classes at all. On the other hand, the K-Y transformation will select u_2 for mapping, which in this example is the axis that will provide maximum discrimination. Thus, it is clear that each of these mapping algorithms will only be suitable for certain types (or structures) of data. Without a priori knowledge of a given data set, the best strategy is to experiment with different mapping algorithms and select the one that provides the best output display (i.e. the subspace with the best separability or discrimination between different classes). A good example can be found in chapter 7 when the high dimensional lung sound data is mapped, by different linear mapping algorithms, onto a lower dimensional space.

Another problem associated with mapping in general is the variability of the transformation matrix. As the transformation matrix is derived from the estimated scatter matrix (section 4.2.1), it follows that the variance of the transformation matrix is proportional to that of the estimated scatter matrix. By deriving a prediction criterion $\hat{Q}$ from an estimated scatter matrix with a gaussian distributed data set, Kalayeh and Landgrebe (1983) have shown that

$$\mathrm{var}(\hat{Q}) = 2n' / (N - 1) \qquad (16)$$

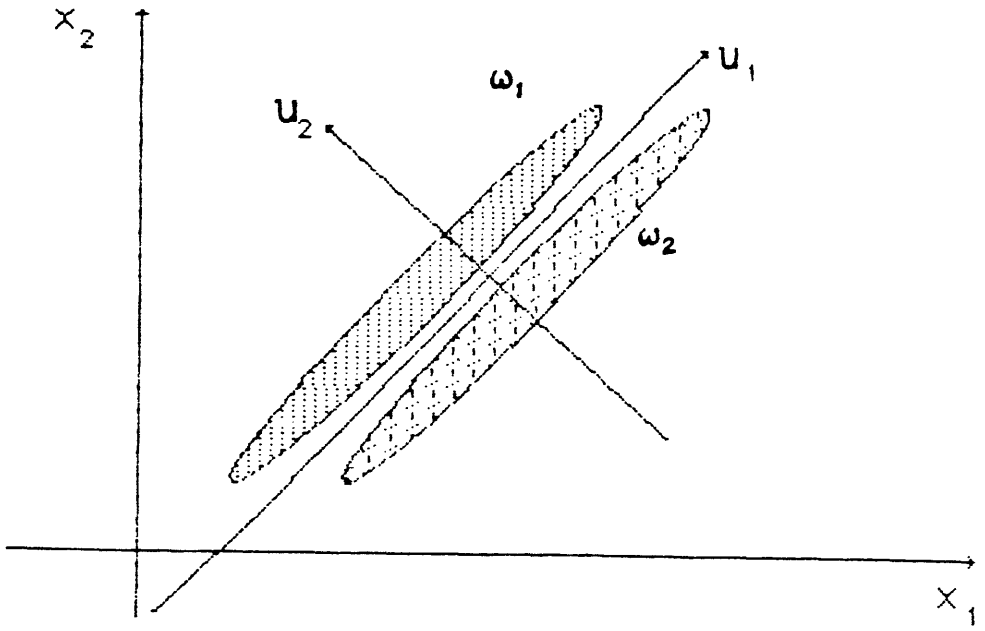


Figure 4.6 A classical example to show the deficency of the K-L transformation. The two ellipses each represent a class of samples and the data set is adequately represented in a two dimensional space with axis x_1 and x_2. The transformed space is represented by the new coordinate axes u_1 and u_2. The K-L transformation will select u_1 as the best axis whereas the K-Y transformation will select u_2.

where var(.) is the variance operator. Thus, the variance of the estimated scatter matrix is related to the dimensionality n' prior to transformation and the number N of (training) samples used to estimate it. Figure 4.7 is a plot of var($\hat{Q}$) versus N at three different values of n'. Thus, a large number of samples is required for a small value of var($\hat{Q}$). For a dimensionality of 20 (as in our lung sound analysis), figure 4.7 indicates that it may require between 100 to 200 samples (per class) to estimate the scatter matrix.

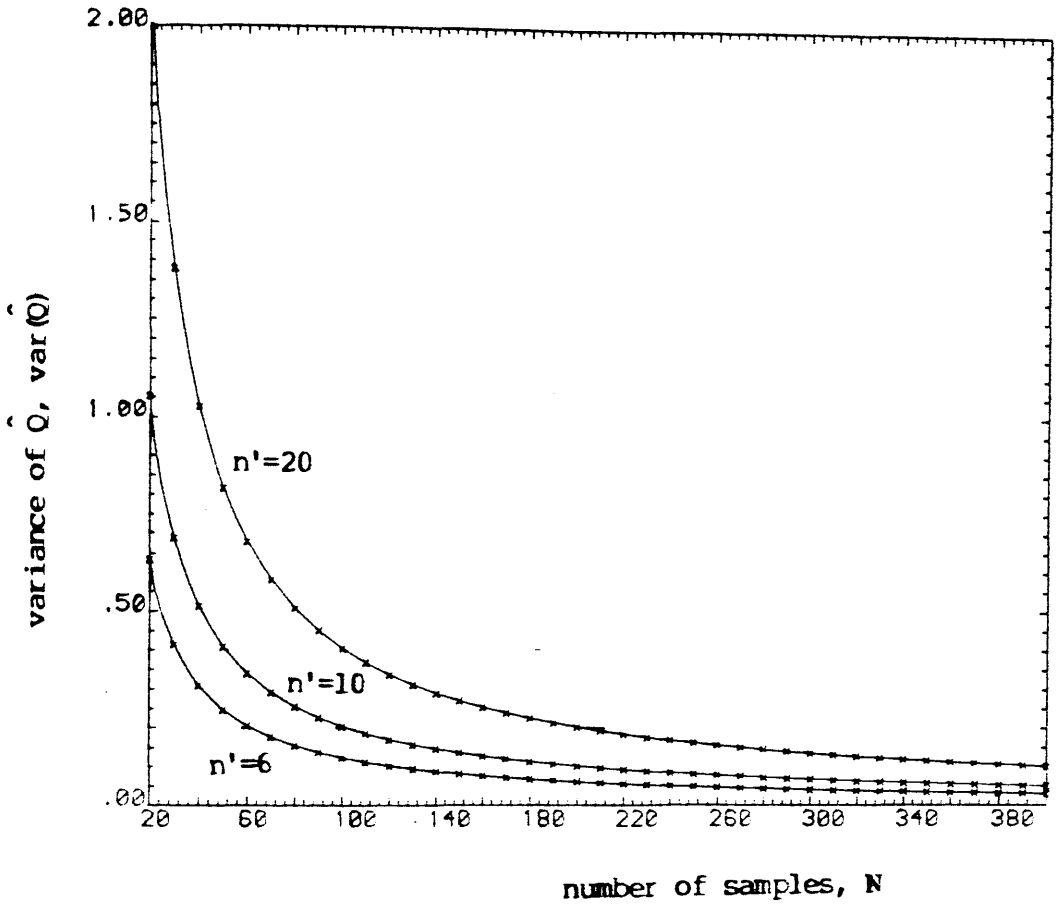


Figure 4.7 A graph of the variance of $\hat{Q}$, var($\hat{Q}$), versus the number of samples N at three different values of dimensionality n' of an estimated scatter matrix with a gaussian distributed data set.

References

[01] Babu, C.C. (1972). "On the application of divergence to feature selection in pattern recognition". **IEEE Trans. on Systems, Man, and Cybernetics**, SMC-2, 668-670.

[02] Calvert, T.W. and Young T.Y. (1969). "Randomly generated non-linear transformations for pattern recognition". **IEEE Trans. on Systems, Science and Cybernetics**, SSC-5, 266-273.

[03] Chien, Y.T. and Fu, K.S. (1967). "On the generalized Karhunen-Loeve expansion". **IEEE Trans. on Information Theory**, 13, 518-520.

[04] Devijver, P.A. and Kittler, J. (1982). **Pattern Recognition: A Statistical Approach**. Prentice-Hall International, London.

[05] Fisher, R.A. (1936). "The use of multiple measurements in taxonomic problems". **Annual of Eugenics**, 7, 179-188.

[06] Foley, D.H. and Sammon, J.W. (1975). "An optimal set of discriminant vectors". **IEEE Trans. on Computers**, C-24, 281-289.

[07] Fukunaga, K. and Koontz, W.L.G. (1970). "Application of the Karhunen-Loeve expansion to feature selection and ordering". **IEEE Trans. on Computers**, C-19, 311-318.

[08] Fukunaga, K. and Mantock, J.M. (1983). "Nonparametric discriminant analysis". **IEEE Trans. on Pattern Analysis and Machine Intelligence**, PAMI-5, 671-678.

[09] Gower, J.C. (1966). "Some distance properties of latent root and vector methods in multivariate analysis". **Biometrika**, 53, 325-338.

[10] Kalayeh, H.M. and Landgrebe, D.A. (1983). "Predicting the required number of training samples". **IEEE Trans. on Pattern Analysis and Machine Intelligence**, PAMI-5, 664-667.

[11] Kittler, J. (1975). "Mathematical models of feature selection in pattern recognition". **International Journal of Man-Machine Studies**, 7, 609-637.

[12] Kittler, J. and Young, P.C. (1973). "A new approach to feature selection based on the Karhunen-Loeve expansion". **Pattern Recognition**, 5, 335-352.

[13] Kruskal, J.B. (1971). "Comments on a non-linear mapping for data structure analysis". **IEEE Trans. on Computers**, C-20, 1014.

[14] Kulikowski, C.A. (1970). "Pattern recognition approach to medical diagnosis". **IEEE Trans. on Systems, Science and Cybernetics**, SSC-6, 173-178.

[15] Longstaff, I.D. (1985). "A powerful extension to Fishers linear discriminant function for pattern recognition". Presented in the British Pattern Recognition Association Third International Conference, St. Andrews, 25-27 September, 1985.

[16] Macfarlane, K.T. (1987). "Automatic classification of tubing defects by analysis of eddy current signals". Ph.D. Thesis, University of Glasgow, to be submitted.

[17] Macleod, J.E.S. (1982). "Pattern classification in the automatic inspection of tubes scanned by a rotating eddy current probe". **Signal Processing**, 5, 445-450.

[18] Sammon, J.W. (1969). "A non-linear mapping for data structure analysis". **IEEE Trans. on Computers**, **C-18**, 401-409.

[19] Sammon, J.W. (1970). "Interactive pattern analysis and classification". **IEEE Trans. on Computers**, **C-19**, 594-616.

[20] Short, R.D. and Fukunaga, K. (1981). "The optimal distance measure for nearest neighbour classification". **IEEE Trans. on Information Theory**, **IT-27**, 622-627.

[21] Urquhart, R.B. (1983). "Mapping techniques in pattern recognition". **The GEC Journal of Research (Incorporating the Marconi Review)**, **1**, 108-121.

[22] Wang, D.K. (1983). "Mixed measure error function for nonlinear mapping algorithms". **Electronics Letters**, **19**, 634-635.

[23] Young, T.Y. and Calvert, T.W. (1974). **Classification, Estimation and Pattern Recognition**. American Elsevier Publishing Company, Inc., New York.

Nearest Neighbour Classification

This work was supported by the Croucher Foundation, Hong Kong.

Chapter 5: Nearest Neighbour Classificaton

Summary

This chapter presents a survey on the various nearest neighbour classification algorithms. Some of the interesting properties of NN classification are also briefly discussed.

5.1. Introduction

In chapter 4, linear mapping techniques have been briefly reviewed. These various algorithms provide the users a powerful tool to visualize a higher dimensional data set (such as the 20-dimensional lung sound data). The interpretation of the lower dimensional data can then be achieved (a) by the users own subjective judgement or (b) by other objective means. In certain applications, subjective judgements by the users are often more desirable, e.g. a physician will always prefer to base his/her decision(s) upon his/her own experience. Nevertheless, if objective criteria are available, these may assist his/her decision. One of these objective means is nearest neighbour (NN) classification (Cover and Hart, 1967). The basic idea behind this technique is that data which fall close together in a space of any dimensionality are likely to belong to the same class.

This chapter attempts to provide a survey of various nearest neighbour classification algorithms. A more in-depth survey of the algorithms is given than is possible in most of the

textbooks in pattern recognition. However, emphasis will be on the algorithms themselves rather than on the theoretical aspects (such as the convergence) of the algorithms. This does not imply that the theoretical aspects are of lesser importance. In fact, an algorithm without a careful theoretical analysis will always be prone to different unexpected errors. Nevertheless, it is not the intention of this review to go through all the mathematical/statistical analysis of each of the algorithms. A brief summary of some of the interesting theoretical properties of NN classification will be given at the end of the review. Interested readers are recommended to refer back to the original papers and standard textbooks on pattern recognition for the full analysis.

The nearest neighbour algorithms are surveyed in the next section. Some of the interesting properties of NN classification are briefly discussed in section 5.3.

5.2. A review of nearest neighbour classification

5.2.1 Introduction

Nearest Neighbour (NN) classification was first proposed and developed by Fix and Hodges (1951,1952) and later by Cover and Hart (1967). The popularity of NN classifiers within the communities of pattern recognition and industry stems from the fact that (in comparison with, for example, iteratively-trained classifiers) they are very simple in both implementation

and use and (in comparison with parametric classifiers) they do not require prior knowledge of the underlying distributions of the data. Furthermore, it is intuitively appealing because a reasonable assumption is that samples very close together in feature space are likely to belong to the same class (or category or group) (Nilsson, 1965; Cover and Hart, 1967). Unfortunately, the disadvantage of this method is that it does require a large number of classified or labelled (and hopefully correct and independent identically distributed) prototype (or training) samples (or prototypes) to be available at the actual time of classification. This implies that a substantial amount of memory is required to store the prototypes and imposes a heavy penalty in the computation of the set of nearest neighbours. Furthermore, the probability of error for the NN rules is always greater than or equal to the Bayesian (minimum) error (Devijver and Kittler, 1982).

Before presenting the various NN classification algorithms, some notations are introduced. Each sample is represented as a vector $\mathbf{x}$ in an n-dimensional feature space. Let $S_N=\{(\mathbf{x}_1, \theta_1), (\mathbf{x}_2, \theta_2), \ldots, (\mathbf{x}_N, \theta_N)\}$ be the set of N classified prototypes, where θ_i (i=1,2,...,N) is a reference or "true" class label assigned by the designer to the prototype sample $\mathbf{x}_i$. Each label can be assigned to any of the c classes $\omega_1, \omega_2, \ldots, \omega_c$. Also it is assumed that S_N is defined over a suitable measurement space, such as the Euclidean space. Within the set S_N, let $N_1, N_2, \ldots, N_c$ be the number of prototypes belonging to classes $\omega_1, \omega_2, \ldots, \omega_c$ respectively, with

$$N = \sum_{p=1}^{c} N_p.$$

Let $S(k)=\{(\mathbf{x}(1), \theta(1)), (\mathbf{x}(2), \theta(2)), \ldots, (\mathbf{x}(k), \theta(k))\}$ be the set of k ($k \geq 1$) nearest neighbours (from S_N) of a test (or unknown) sample $\mathbf{x}$ such that $\mathbf{x}(j)$ ($j=1,2,\ldots,k$) is the j-th nearest neighbour to $\mathbf{x}$, where $\theta(j)$ is the class label of $\mathbf{x}(j)$. The class membership of the test sample $\mathbf{x}$ is denoted as θ. Finally, let r_q ($q=1,2,\ldots,c$) be the number of nearest neighbours from class ω_q within the set S(k), with

$$k = \sum_{q=1}^{c} r_q.$$

In the next section, a number of NN classification algorithms will be introduced. When the size of S_N is large, considerable time is needed to compute the set of nearest neighbour S(k). Two approaches have been proposed to reduce this computation. One of them is to reduce the number of prototypes by selecting a representative subset from S_N. The other alternative is to shorten the searching time for each nearest neighbour in the set S(k), which usually involves some reordering of the set S_N. A more comprehensive review of the two approaches has been presented in an internal report by Luk (1987).

5.2.2 Nearest neighbour classification algorithms

5.2.2.1 Introduction

The nearest neighbour (NN) classification algorithms proposed throughout the past three decades can be grouped into three categories, namely:

(a) those NN rules which use a voting system for making a decision;

(b) those NN rules which employ a distance-related measurement to make a decision; and

(c) those NN rules which are based on the non-parametric estimation of the class conditional probability density functions in the Bayesian classification rule (Devijver and Kittler, 1982).

In the following subsections, these three categories of NN algorithms will be introduced and some of their relationships with other NN algorithms will be described. (It must be emphasised that the third category is described only for the sake of completeness in this survey.)

5.2.2.2 Algorithms that use a voting system

5.2.2.2.1 k-nearest neighbour classification rule

The (traditional or classical) k-NN classification rule (Fix and Hodges, 1951,1952; Cover and Hart, 1967) (where $k \geq 1$) assigns a test sample $\mathbf{x}$ to that class ω_i (say) which receives

the majority of votes from its k nearest neighbours. If indecision has occured (for example due to ties between two or more classes), it can be resolved arbitrarily or by other means. Formally this rule can be defined by

$$\hat{\omega}(k) = \begin{cases} \omega_i & \text{if } r_i = \max_{j=1}^{c} r_j \\ \text{resolved arbitrarily or by other means if indecision has occured} \end{cases} \qquad (1)$$

where $\hat{\omega}(k)$ indicates that the decision depends on the value of k. Many interesting properties of this algorithm have been examined by various researchers. However, these will be deferred to later sections so that the reader can have a better overview of all the other proposed algorithms.

5.2.2.2.2 (k,ℓ)-nearest neighbour classification rule

Hellman (1970) and Tomek (1976) independently suggested the "k-NN rule with a reject option" (i.e. the (k,ℓ)-NN rule) for a two class problem. Generalised to the multiclass problem, the rule may be stated thus: a test sample **x** is assigned to class ω_i provided ω_i receives at least ℓ votes (where $\ell \geq \lceil k/c \rceil$). Otherwise, **x** is placed in the "rejected" class ω_0. Thus, the parameter ℓ serves as a rejection index. Ties can again be resolved arbitrarily or by other means, or samples with tied

votes can be placed in the reject class. Formally, the decision rule can be defined by

$$\hat{\omega}(k) = \begin{cases} \omega_i & \text{if } r_i \geq \ell \text{ and } r_i = \max_{j=1}^{c} r_j \\ \omega_0 & \text{otherwise.} \end{cases} \qquad (2)$$

If $\ell = \lceil k/c \rceil$, equations (2) becomes the k-NN rule (section 5.2.2.2.1).

5.2.2.2.3 (k,ℓ_i)-nearest neighbour classification rule

This rule was proposed by Devijver (1977) and is a generalization of the (k,ℓ)-NN rule (section 5.2.2.2.2). It assigns a test sample $\mathbf{x}$ to class ω_i if the number of majority votes from that class is at least ℓ_i (where $\ell_i \geq \lceil k/c \rceil$). The parameters $\{\ell_i, i=1,2,...c\}$ are specified by the user. Otherwise, $\mathbf{x}$ is placed in the "reject" class ω_0. (Ties can be dealt with as mentioned in section 5.2.2.2.1.) Formally, this rule can be defined by

$$\hat{\omega}(k) = \begin{cases} \omega_i & \text{if } r_i \geq \ell_i \text{ and } r_i = \max_{j=1}^{c} r_j \\ \omega_0 & \text{otherwise.} \end{cases} \qquad (3)$$

It is clear that if $\mathfrak{L}_i = \mathfrak{L}$ (i=1,2,...,c), equations (3) are identical to equations (2). Also, comparing equations (2) and (3) with equations (1) indicates the philosophy behind using a reject class: a rejection may in some way reduce the cost of a wrong decision. It is also clear from equations (2) and (3) that one method of resolving indecision is simply not to make any decision, i.e. to place the test sample into class ω_0.

5.2.2.3 Algorithms that use distance-related measurement

5.2.2.3.1 k-th nearest neighbour classification rule

This algorithm was proposed by Goldstein (1972) for a two class problem. Let $d_i(k)$ be the distance between the test sample $\mathbf{x}$ and its k-th nearest neighbour from class ω_i. In a form generalized to the multiclass problem, the algorithm can be stated as follows:

$$\hat{\omega}(k) = \begin{cases} \omega_i & \text{if } d_i(k) = \min\limits_{j=1}^{c} d_j(k) \\ \text{resolved arbitrarily or by other means if indecision has occured} & \end{cases} \qquad (4)$$

that is, the test sample $\mathbf{x}$ is assigned by equations (4) to the class ω_i having the smallest distance between $\mathbf{x}$ and its k-th NN from class ω_i. Note that in equations (4) (and the subsequent

NN algorithms) indecision could only occur if the measurement space is discrete or a certain maximum limit has been imposed on $d_i(k)$ (or other distance-related measurements). It has been shown (Patrick and Fischer, 1970; Goldstein, 1972) that if the measurement space is continuous, equations (4) are equivalent to a v-NN rule (where v=ck) provided the v-NN rule produces no ties.

5.2.2.3.2 k-means nearest neighbour classification rule

The k-means NN rule was proposed by Rabiner et al (1979). Let $\bar{d}_i(k)$ be the averaged distance of the k nearest neighbours from class ω_i of a test sample **x**, i.e.

$$\bar{d}_i(k) = (1/k) \sum_{q=1}^{k} d_i(q)$$

The rule then assigns **x** to the class with the smallest averaged distance. Formally, it is defined as

$$\hat{\omega}(k) = \begin{cases} \omega_i & \text{if } \bar{d}_i(k) = \min_{j=1}^{c} \bar{d}_j(k) \\ \text{resolved arbitrarily or by other means if indecision has occured.} \end{cases} \qquad (5)$$

5.2.2.3.3 Distance-weighted NN classification rule

Dudani (1976) has suggested that nearest neighbours

closest to the test sample **x** should be weighted most heavily. He has proposed the use of a weight which decreases with increasing sample-to-neighbour distance. Let d(j) be the distance between **x** and its j-th NN **x**(j). Dudani defines the weight w(j) for the j-th NN as

$$w(j) = \begin{cases} \dfrac{d(k) - d(j)}{d(k) - d(1)} & d(j) \neq d(1) \\ 1 & d(j) = d(1). \end{cases} \qquad (6)$$

The distance-weighted k-NN classification rule then assigns **x** to class ω_i with the largest total weight $\mathcal{V}_i$, i.e.,

$$\hat{\omega}(k) = \begin{cases} \omega_i & \text{if } \mathcal{V}_i = \max_{q=1}^{c} \mathcal{V}_q \\ \text{resolved arbitrarily or by other means if indecision has occured} \end{cases} \qquad (7)$$

where

$$\mathcal{V}_i = \sum_{j=1}^{k} w(j) * I\{\theta(j), \omega_i\}$$

and

$$I\{\theta(j),\omega_i\} = \begin{cases} 1 & \text{if } \theta(j) = \omega_i \\ 0 & \text{if } \theta(j) \neq \omega_i \end{cases}$$

(the function $I\{\theta(j),\omega_i\}$ is known as the indicator function).

Before continuing to the next distance-related NN classification algorithms, it is worth noting that the weight defined in equations (6) will not improve the performance of equations (7) (in terms of averaged probability of errors for a finite set of prototype samples) when compared with the k-NN classification rule defined in equations (1) (Bailey and Jain, 1978; Morin and Raeside, 1981; Macleod et al, 1987). Moreover, Bailey and Jain (1978) have also shown that the asymptotic performance (i.e., the performance when the number of prototypes approaches infinity) of the k-NN classification rule is at least as good as that of any weighted rule. The topics in this paragraph are discussed further in section 6.2, where some new results are presented.

5.2.2.3.4 Dasarathy's NN classification rule

Very often, information initially available to the designer of a pattern recognition system is insufficient to account for the future environment. In a medical problem, for example, a new lung disorder caused by an ideopathic agent may arise. In order to identify the possible existence of one or more such undefined classes, Dasarathy (1976) has proposed the

inclusion of a "near enough" parameter Δ_i in the nearest neighbour classification rule. This parameter can be defined as the maximum nearest neighbour distance determined over the set of all prototypes belonging to class ω_i, i.e.

$$\Delta_i = \max_{\mathbf{x}_j \in \omega_i} \{ \min_{\mathbf{x}_p \in \omega_i} d_i(\mathbf{x}_j, \mathbf{x}_p) \}$$

where $d_i(\mathbf{x}_j, \mathbf{x}_p)$ denotes the distance between two class ω_i prototypes $\mathbf{x}_j$ and $\mathbf{x}_p$. On including this "near enough" parameter, the k-th nearest neighbour classification rule (section 5.2.2.3.1) can be rewritten as

$$\hat{\omega}(k) = \begin{cases} \omega_i & \text{if } d_i(k) = \min_{j=1}^{c} d_j(k) \leq \Delta_i \\ \omega_0 & \text{otherwise.} \end{cases} \qquad (8)$$

Any test samples that fails the "near enough" criteria in equations (8) will be discarded into a "rejected" class ω_0 (as in section 5.2.2.2.2). This set of discarded test samples can then be subjected to further analysis by other techniques (such as those mapping algorithms that have already been discussed in chapter 4, or the graph-theorical algorithms developed by Urquhart, 1983). This permits evolution of new classes as and when deemed necessary by the designer(s) of the pattern recognition system.

5.2.2.4 Algorithms that are based on nonparametric estimation

5.2.2.4.1 Introduction

The above mentioned NN algorithms are constructed using some rather ad hoc criteria (or what Dasarathy and Sheela, 1977, refer to as "abstractions"), such as the different voting mechanisms employed in subsection 5.2.2.2. These algorithms are ad hoc because their derivations are not based on any rigorous statistical/mathematical theorems. In this subsection, the derivations of the algorithms are based on the Bayesian decision rule (Devijver and Kittler, 1982), and hence the algorithms have a better statistical/mathematical foundation. One disadvantage of the following algorithms (contrast section 5.2.1) is that the a priori probability P_i that a test sample **x** comes from class ω_i (where i=1,2,...c) is assumed to be known and the class conditional probability density function $p(\mathbf{x}|\omega_i)$ at the test samples **x** is assumed fixed and continuous but of unknown form. (Note that P_i is also assumed to be available in other classification algorithms.) Let L_{ij} be a loss (or cost) function (that is the cost of making a wrong decision when the test sample **x** is assigned to class ω_i when it is actually drawn from class ω_j). If $p(\mathbf{x}|\omega_i)$ is known, the Bayesian decision rule will assign **x** to class ω_i if

$$\sum_{j=1}^{c} L_{ij}p(\mathbf{x}|\omega_j)P_j = \min_{1\leq q\leq c} \sum_{j=1}^{c} L_{qj}p(\mathbf{x}|\omega_j)P_j \qquad (9)$$

Therefore, the aim of the following algorithms is to find an estimate $\hat{p}(\mathbf{x}|\omega_i)$ which is nonparametric and distribution-free (i.e. independent of the underlying probability density function) and which is based on the nearest-neighbour concept, of the class conditional probability density function $p(\mathbf{x}|\omega_i)$. The estimate is to be consistent i.e. $\hat{p}(\mathbf{x}|\omega_i)$ is to approach $p(\mathbf{x}|\omega_i)$ with probability 1.

5.2.2.4.2 Loftsgaarden and Quesenberry's classification rule

Loftsgaarden and Quesenberry (1965) have proposed a non-parametric method of estimating the class conditional probability density function which can be used for classification. As in Goldstein's (1972) k-th nearest neighbour classification rule, the distance $d_i(k)$ between the test sample $\mathbf{x}$ and the k-th nearest neighbour from class ω_i is calculated. The volume of the hypersphere with radius $d_i(k)$ and centre $\mathbf{x}$ is then given by

$$\phi_{d_i(k)} = \frac{2\pi^{n/2}\{d_i(k)\}^n}{n\Gamma(n/2)}$$

(where Γ is the gamma function with parameter n/2). These authors' estimator can then be written as

$$\hat{p}(x|\omega_i) = \frac{k-1}{N_i} \left\{ \frac{1}{\phi_{d_i(k)}} \right\}. \qquad (10)$$

Replacing the class conditional probability density function in equation (9) by the estimated function given in equation (10) yields the required classification rule.

5.2.2.4.2 Generalized k-nearest neighbour classification rule

Patrick and Fischer (1970) generalized the result of Loftsgaarden and Quesenberry (1965) by introducing the concept of a distribution-free tolerance region into equation (10). (For different methods of constructing a tolerance region, refer to the papers by Wilks (1941), Tukey (1947), Kemperman (1956), and Fraser and Guttman (1956).) Essentially, the N_i prototypes from class ω_i are processed to form tolerance regions (say using Tukey's construction technique) which partition the measurement space. To each of these tolerance regions, an index will be assigned. Let ξ_i be the index of the tolerance region for class ω_i which contains the test sample **x**. If **x** happens to be on the boundary between tolerance regions, the smaller index will be chosen. The class conditional probability density function can then be estimated as

$$\hat{p}(x|\omega_i) = \frac{\beta_{\xi_i}}{N_i + 1} \left\{ \frac{1}{\phi_{\xi_i}} \right\}. \qquad (11)$$

where β_{ξ_i} is the number of prototypes involved in constructing a tolerance region ξ_i and ϕ_{ξ_i} is the volume of that tolerance region. Again, the decision rule is obtained by replacing $p(x|\omega_i)$ in equation (9) by equation (11).

It is easy to observe (Patrick and Fischer, 1970) that if a spherical tolerance region is used in the above estimate (i.e., a set of all points inside the hypersphere centered at $\mathbf{x}$, which contains k - 1 class ω_i prototypes inside, one prototype sample on the surface (which is not in the tolerance region), and the rest of the N_i - k prototypes outside), equation (11) becomes the estimated class conditional probability density function used by Loftsgaarden and Quesenberry (1965).

On the other hand, it has been shown (Patrick and Fischer, 1970) for a two-class problem, the decision rule with $\hat{p}(x|\omega_i)$ given in equation (11) is equivalent to both the k-th-NN and k-NN classification rule if the following conditions are met. They are

(a) a spherical tolerance region is used in equation (11),

(b) $\beta_{\xi_1} = \beta_{\xi_2} = k$ in equation (11),

(c) a 0-1 loss function is used in equation (9), and

(d) $P_1/P_2 = (N_1 + 1)/(N_2 + 1)$ in equation (9).

5.2.2.5. Remarks: choice of NN algorithm

With the possible exception of the third category (section 5.2.2.4), all the above algorithms are easy to implement

and to use. The choice of algorithm is problem/data dependent. In certain problems, the simple k-NN classification rule may be better than the rest of the algorithms. On the other hand, some problems would prefer a rejection to a wrong decision, and in these circumstances the (k,ℓ)-NN or (k,ℓ_i)-NN rule may be preferable. Particular rules may be suitable for particular problems: Rabiner et al (1979), for example, claim that their rule is best suited for the speech recognition problem they were tackling. However, if the designer is uncertain about the sufficiency of his initial information, Dasarathy's NN classification rule may look more attractive. When more information about the data is available, such as the a priori probability, the application of Patrick and Fischer's algorithm may be more appropriate. On the whole, the best strategy seems to be to experiment with each algorithm and to find the one that best suits the problem as well as its associated environment.

5.3 Properties of nearest neighbour classification rules

5.3.1 Introduction

Three factors have to be considered in selecting and evaluating a classification rule (Penrod, 1976). They are

(a) the cost of implementation (development) and use,

(b) finite sample performance, and

(c) infinite sample performance.

Factor (a) has been addressed throughout the previous

sections and will not be considered further. In short, this factor depends on the actual problem. The finite sample performance, factor (b), of a rule with a particular data set which is currently available is even more important. Without the knowledge of the underlying probability density functions (class conditional or unconditional) of the prototype samples and/or the a prior probabilities, the smallest possible probability of error or minimal risk, i.e. the Bayes risk E^*, is unknown to a designer. Thus, even if the data set is quite large, there is no way to know how well the selected rule will perform. In fact, even though the rule should do reasonably well in the large sample case, its performance with the data set available may be unacceptably bad if either the inherent risk of the data set is high or simply the data set is not large enough. Since this is by itself a major research area, the discussion will be deferred to the next chapter where such a problem will be considered for a single class of nearest neighbour rules, namely weighted nearest neighbour rules (Dudani, 1976; Macleod et al, 1987).

In this section, a brief discussion is given of the theoretical results for nearest neighbour classification rules when the number of prototypes is infinite. These include convergence, asymptotic bounds, consistency and the rate of convergence.

5.3.2 Convergence of NN classification rules

Given an infinite prototype set, both the unconditional

probability density function (pdf) of the nearest neighbours and their respective a posteriori probabilities approach the unconditional pdf $p(\mathbf{x})$ of the test sample $\mathbf{x}$ and its a posteriori probability $p(\omega_i \mid \mathbf{x})$ (sometimes also denoted as $\eta_i(\mathbf{x})$ or simply η_i, where i= 1, 2, ...,c) respectively (Devijver and Kittler, 1982; Peterson, 1970). This fact is basic to theoretical studies on convergence of NN classification rules when the number of prototypes tends to infinity.

Let $\hat{e}_k(\mathbf{x})$ be the finite sample, conditional probability of error of a NN classificaton rule given the test sample $\mathbf{x}$ and the set S(k) of k nearest neighbours (where $S(k) \subset S_N$), i.e.

$$\hat{e}_k(\mathbf{x}) = \Pr\{ \theta \neq \hat{\omega} \mid \mathbf{x}, S(k) \}$$

$$= \sum_{i=1}^{c} \Pr\{ \theta = \omega_i, \hat{\omega} \neq \omega_i \mid \mathbf{x}, S(k) \} \qquad (12)$$

where θ and $\hat{\omega}$ are as defined in sections 5.2.1 and 5.2.2.2.1 respectively. Since all the observations (i.e. the prototype samples and the test samples) are assumed to be independent and identically distributed (section 5.2.1), equation (12) can be rewritten as

$$\hat{e}_k(\mathbf{x}) = \sum_{i=1}^{c} \Pr\{ \theta = \omega_i \mid \mathbf{x} \} \Pr\{ \hat{\omega} \neq \omega_i \mid S(k) \}$$

$$= \sum_{i=1}^{c} \eta_i(\mathbf{x}) \Pr\{ \hat{\omega} \neq \omega_i \mid S(k) \} \qquad (13)$$

where $\Pr\{ \hat{\omega} \neq \omega_i \mid S(k) \}$ depends on the a posteriori probabilities of the k nearest neighbours. The problem of the convergence of a nearest neighbour rule is to show that $\hat{e}_k(\mathbf{x})$ approaches a constant as the number of prototype samples N increases. Cover and Hart (1967) have shown that the expected value of $\hat{e}_k(\mathbf{x})$ converges to a constant E_k as N approaches infinity under the conditions that the class conditional and unconditional pdf's of the test sample $\mathbf{x}$ are continuous, i.e. they have shown that

$$\lim_{N \longrightarrow \infty} \mathbb{E} \{ \hat{e}_k(\mathbf{x}) \} \longrightarrow E_k \quad \text{in probability} \qquad (14)$$

where $\mathbb{E}$ is the expectation operator. The constant E_k is known as the average asymptotic error probability or simply the error rate of the k-NN classificaton rule. Wagner (1971) and Penrod (1976) have pointed out that the result in equation (14) is important if a designer has a large number of very large data sets. What a designer really would like to know is what happens to a particular NN rule when a prototype set approaches infinite size. Fortunately, Wagner (1971) and Fritz (1975) have shown that equation (14) also holds even without the expectation operator, i.e.

$$\lim_{N \longrightarrow \infty} \hat{e}_k(\mathbf{x}) \longrightarrow e_k(\mathbf{x}) \quad \text{in probability} \tag{15}$$

under the same conditions as in Cover and Hart, where the constant $e_k(\mathbf{x})$ is the conditional asymptotic probability of error of the classification rule. Stone (1977) and Devroye (1981) have further relaxed these conditions to (a) that $p(\mathbf{x})$ is distributed in a separable metric space (or non-atomic space) and (b) that $\eta_i(\mathbf{x})$ is almost everywhere continuous (i.e. it is decomposable into a continuous component plus a series of mass points).

5.3.3 Asymptotic bounds for NN classification rules

Cover (1968) in his classical study of nearest neighbour classification rules gave the following bounds for the constants E_k in the multiclass case as:

$$\text{for } k = 1,\ E^* \leq E_1 \leq 2E^*, \text{ and} \tag{16a}$$

$$\text{for } k > 1,\ E^* \leq E_k \leq E^* + (1/k)E^*. \tag{16b}$$

Gyorfi and Gyorfi (1978) tightened the bounds in expression (16b) to

$$\text{for } k > 1,\ E^* \leq E_k \leq E^* + (\sqrt{c/k})E_1. \tag{16c}$$

When only two classes are involved, Cover and Hart (1967) and Devijver (1979) further tightened the bounds for odd k in expressions (16) to

$$\text{for } k = 1,\ E^* \leq E_1 \leq 2E^*(1 - E^*) \tag{17a}$$

$$\text{for } k > 1 \text{ and } k \text{ odd},\ E^* \leq E_k \leq E^* + (\sqrt{1/k'\pi})E_1 \tag{17b}$$

where k' is the largest integer of k/2. The case when k is even does not enter into expressions (17) because it has been shown (Devijver and Kittler, 1982) that

$$E_{2j} = E_{2j-1} \tag{18}$$

where j=1, 2, ...

On the other hand, Devijver (1979) has also shown that the average asymptotic error probabilities $E_{k,\ell}$ and E_{k,ℓ_i} (where i=1, 2, ..., c) for the (k,ℓ)-NN and (k,ℓ_i)-NN classification rules are always less than or equal to the Bayes risk E^*, except when ℓ or ℓ_i is equal to k' (i.e. without rejection). In that case, it is similar to E_k, that is it is always greater than or equal to E^*. Thus,

$$E_{k,\ell} \text{ or } E_{k,\ell_i} \leq E^* \leq E_{k,k'} = E_k \tag{19}$$

with $\ell > k'$ and $\ell_i > k'$. Equation (19) also indicates that the use of a reject option in (k,ℓ)-NN and (k,ℓ_i)-NN classification rules may reduce the average probability of misclassification below that of the Bayes risk, a very desirable property for the two types of rules.

5.3.4 Consistency and rate of convergence of NN classification

Devroye and Wagner (1980) showed that if (a) $k \longrightarrow \infty$ and (2) $k/N \longrightarrow 0$ as $N \longrightarrow \infty$, then using the inequalities of Stone (1977),

$$E_k \longrightarrow E^* \quad \text{in probability} \quad \text{as} \quad N \longrightarrow \infty. \tag{20}$$

That is, the NN classification rule is Bayes risk consistent if conditions (a) and (b) are satisfied. With additional restrictions and by using a recursive algorithm, Krzyzak and Pawlak (1984) have also shown that

$$e_k(x) \longrightarrow e^*(x) \quad \text{in probability} \quad \text{as} \quad N \longrightarrow \infty \tag{21}$$

where $e^*(x)$ is the Bayes risk given the test sample $\mathbf{x}$ (or conditional Bayes risk). Again, equation (21) shows that the conditional asymptotic probability of error is conditional Bayes risk consistent.

If both conditions (a) and (b) are satisfied, Gyorfi (1981) has shown that the rate of convergence of such a NN classification rule is $N^{-1/(2 + n/a)}$ where $a > 0$. However, if these conditions are not satified, Gyorfi (1978) and Fritz (1975) have shown that the k-NN classification rules ($k \geq 1$) have an exponential rate of convergence which depends on the number of prototype samples N, the dimensionality n and the underlying distributions of the prototype samples.

References

[01] Bailey, T. and Jain, A.K. (1978). "A note on distance-weighted k-nearest neighbour rules". **IEEE Trans. on Systems, Man, and Cybernetics**, **SMC-8**, 311-313.

[02] Cover, T.M. (1968). "Estimation by the nearest neighbour rule". **IEEE Trans. on Information Theory**, **IT-14**, 50-55.

[03] Cover, T.M. and Hart, P.E. (1967). "Nearest neighbour pattern classification". **IEEE Trans. on Information Theory**, **IT-13**, 21-27.

[04] Dasarathy, B.V. (1976). "Is your nearest neighbour near enough a neighbour". **Proceedings of the first International Conference on Information Sciences and Systems**, Patras, Greece, August, 1976, 114-117.

[05] Dasarathy, B.V. and Sheela, B.V. (1977). "Visiting nearest neighbours: a survey of nearest neighbour classification techniques". **Proceedings of the International Conference on Cybernetics and Society**, May, 1977, 630-636.

[06] Devijver, P.A. (1977). "Reconnaissance des formes par la méthode des plus proches voisins". Doctoral Dissertation, University of Paris VI.

[07] Devijver, P.A. (1979). "New error bounds with the nearest neighbour rule". **IEEE Trans. on Information Theory**, **IT-25**, 749-753.

[08] Devijver, P.A. and Kittler, J. (1982). **Pattern Recognition: A Statistical Approach.** Prentice-Hall International, London.

[09] Devroye, L.P. and Wagner, T.J. (1980). "Distribution-free consistency results in nonparametric discrimination and regression function estimation". **The Annals of Statistics**, **8**, 231-239.

[10] Devroye, L.P. (1981). "On the inequality of Cover and Hart in nearest neighbour discrimination". **IEEE Trans. on Pattern Analysis and Machine Intelligence**, **PAMI-3**, 75-78.

[11] Dudani, S.A. (1976). "The distance-weighted k-nearest neighbour rule". **IEEE Trans. on Systems, Man, and Cybernetics**, **SMC-6**, 325-327.

[12] Fix, E. and Hodges, J.L. (1951). "Discriminatory analysis, nonparametric discrimination". Project 21-49-004, Report No. 4. USAF School of Aviation Medicine, Randolph Field, Texas, February 1951.

[13] Fix, E. and Hodges, J.L. (1952). "Discriminatory analysis, small sample performance". Project 21-49-004, Report No. 11. USAF School of Aviation Medicine, Randolph Field, Texas, August 1952.

[14] Fraser, D.A.S. and Guttman, I. (1956). "Tolerance regions". **Annals of Mathematical Statistics**, **27**, 162-179.

[15] Fritz, J. (1975). "Distribution-free exponential error bound for nearest neighbour pattern classification". **IEEE Trans. on Information Theory**, **IT-21**, 552-557.

[16] Goldstein, M. (1972). "k_n-nearest neighbour classification". **IEEE Trans. on Information Theory**, **IT-18**, 627-630.

[17] Gyorfi, L. (1978). "On the rate of convergence of nearest neighbour rules". **IEEE Trans. on Information Theory**, **IT-24**, 509-512.

[18] Gyorfi, L. (1981). "The rate of convergence of k_n-NN regression estimates and classification rules". **IEEE Trans. on Information Theory**, **IT-27**, 362-364.

[19] Gyorfi, L. and Gyorfi, Z. (1978). "An upper bound on the asymptotic error probability of the k-nearest neighbour rule for multiple classes". **IEEE Trans. on Information Theory**, **IT-24**, 512-514.

[20] Hellman, M.E. (1970). "The nearest neighbour classification rule with a reject option". **IEEE Trans. on Systems, Science and Cybernetics**, **SSC-6**, 179-185.

[21] Kemperman, J.H.B. (1956). "Generalized Tolerance Limits". **Annals of Mathematical Statistics**, **27**, 180-186.

[22] Krzyzak, A. and Pawlak, M. (1984). "Distribution-free consistency of a nonparametric kernel regression estimate and classification". **IEEE Trans. on Information Theory**, **IT-30**, 78-81.

[23] Loftsgaarden, D.O. and Quesenberry, C.P. (1965). "A nonparametric estimate of a multivariate density function". **Annals of Mathematical Statistics**, **36**, 1049-1051.

[24] Luk, A. (1987). "Nearest neighbour classification". Internal Report, University of Glasgow.

[25] Macleod, J.E.S., Luk, A. and Titterington, D.M. (1987). "A re-examination of the distance weighted k-NN classification rule". **IEEE Tran. on Systems, Man, and Cybernetics**, to be published.

[26] Morin, R.L. and Raeside, D.E. (1981). "A reappraisal of distance-weighted k-nearest neighbour classification for pattern recognition with missing data". **IEEE Trans. on Systems, Man, and Cybernetics**, **SMC-11**, 241-243.

[27] Nilsson, N.J. (1965). **Learning Machines**. McGraw-Hill, New York.

[28] Patrick, E.A. and Fischer, F.P. II. (1970). "A generalized k-nearest neighbour rule". **Information and Control**, **16**, 128-152.

[29] Penrod, C.S. (1976). "Nonparametric estimation with local rules". Ph.D. dissertation, University of Texas at Austin.

[30] Peterson, D.W. (1970). "Some convergence properties of a nearest neighbour rule". **IEEE Trans. on Information Theory**, **IT-16**, 26-31.

[31] Rabiner, L.R., Levinson, S.E., Rosenberg, A.E. and Wilpon, J.G. (1979). "Speaker-independent recognition of isolated words using clustering techniques". **IEEE Trans. on Acoustics, Speech, and Signal Processing**, **ASSP-27**, 336-349.

[32] Stone, C.J. (1977). "Consistent nonparametric regression". **The Annals of Statistics**, **5**, 595-645.

[33] Tomek, I. (1976). "A generalization of the k-NN rule". **IEEE Trans. on Systems, Man, and Cybernetics**, **SMC-6**, 121-126.

[34] Tukey, J.W. (1947). "Nonparametric estimation II. Statistically equivalent blocks and tolerance regions - the continuous case". **Annals of Mathematical Statistics**, **18**, 529-539.

[35] Urquhart, R.E. (1983). "Some new techniques for pattern recognition research and lung sound signal analysis". Ph.D. thesis, University of Glasgow.

[36] Wagner, T.J. (1971). "Convergence of the nearest neighbour rule". **IEEE Trans. on Information Theory**, **IT-17**, 566-571.

[37] Wilks, S.S. (1941). "Determination of samples sizes for setting tolerance limits". **Annals of Mathematical Statistics**, **12**, 91-96.

Some new results in nearest neighbour classification

(The author would like to express his thanks to Professor D.M. Titterington, to Dr. J.E.S. Macleod, and to the referees of two of the author's publications cited herein, for their helpful advice and comments during the preparation of this chapter; and to Dr. D.J. Hand for helpful discussion during the presentation of a paper in the British Pattern Recognition Association Third International Conference, St. Andrews, 25-27 September, 1985.)

This work was supported by the Croucher Foundation, Hong Kong.

Some new results in nearest neighbour classification

Summary

This chapter presents some of the recent results on nearest neighbour classification. In particular, the finite performance of the weighted k-NN rule is compared with that of the traditional k-NN rule. Macleod et al (1987) have shown that a weighted rule may in some cases achieve a lower finite, conditional probability of error than the unweighted rule when the size of the prototype set is finite. This conclusion was confirmed by solving analytically a particular simple problem and by experimental results using a generalised form of Dudani's weighting function.

An alternative nearest neighbour classification scheme is also introduced. Using a finite set of gaussian data, its finite behaviour is examined. Modifications to reduce rejections are suggested.

6.1 Introduction

This chapter is divided into two parts. The first part addresses the question left over from chapter 5, namely, the finite sample performance (i.e. the performance when the size of the prototype set is finite) of a nearest neighbour classifier (section 5.4.1). As this is a very difficult problem, only one type of NN classification algorithm, viz the weighted k-NN rule

(section 5.2.2.3.3), will be examined in the next section (section 6.2). As mentioned in section 5.2.2.3.3, Bailey and Jain (1978) have proved that the asymptotic conditional probability of error $e_k(\mathbf{x})$ of the unweighted k-NN rule (section 5.2.2.2.1) for a test sample $\mathbf{x}$ is lower than that of any weighted k-NN rule. In section 6.2, this conclusion is reconsidered for the case when the size of the prototype set is finite. In particular, equations for the finite, conditional probability of error $\hat{e}_k(\mathbf{x})$ of $\mathbf{x}$ are developed. It is then argued that a weighted rule may in some cases achieve a lower $\hat{e}_k(\mathbf{x})$ than the unweighted rule. This conclusion is confirmed by analytically solving a particular simple problem and, as an illustration, experimental results obtained using a generalised form of Dudani's weighting function (1976) are also presented.

The second part (section 6.3) of this chapter introduces a new type of NN algorithm. Essentially, the nearest neighbours in S(k) are examined sequentially in order of increasing distance from the unknown sample $\mathbf{x}$ until a specified majority m of votes in favour of some pattern class over its nearest rival occurs. Again, experiments with a finite number of normally distributed prototype samples are described. As with the $(k,\mathbf{l})$ or $(k,\mathbf{l}_i)$ NN classification rules, rejection increases as m increases. Therefore, modifications to reduce the probability of rejection are also suggested in section 6.3. Experiments are given to illustrate the results of these modifications.

6.2 Finite-sample performance of weighted k-NN classification rules

6.2.1. Introduction

In many pattern classification problems, a set of classified prototype samples (not necessarily completely correct) and an additional set of test samples are available. Many classical statistical pattern recognition techniques can be applied. One of these techniques is the k-nearest neighbour (NN) classification rule introduced in section 5.2.2.2.1. Another intuitively appealing idea, due to Dudani (1976), is that a prototype sample closest to an unclassified test sample should be weighted most heavily. Dudani proposed the use of a weight which increases as the distance between the test sample and its nearest neighbours decreases.

It has however been shown by Bailey and Jain (1978) that the conditional asymptotic probability of error $e_k(\mathbf{x})$ of the traditional unweighted k-NN rule (i.e. its performance assuming an infinite set of prototype samples) is better than that of any weighted k-NN rule. This conclusion is not disputed. In the same paper, Bailey and Jain also present the results of an experiment in which a k-NN rule gives a lower finite, average frequency of misclassification than a distance-weighted k-NN rule using Dudani's weighting function. Similar results were obtained by Morin and Raeside (1981). These results would tend to imply that the above conclusion for the conditional asymptotic probability

of error may also apply when the number of prototype samples is finite. However from experimental results in three recent papers, one can gather evidence suggesting that it does not apply. Brown and Koplowitz (1979) used a NN rule weighted according to the numbers of samples in the respective pattern classes and obtained better performance than from the unweighted rule on a finite prototype set. Keller et al (1985) proposed a fuzzy k-NN rule which can be considered as another weighted rule (weighting in this case being based on fuzzy logic); for a finite number of prototype samples, these workers' rule also performed better than the unweighted rule. Most interestingly of all from the present viewpoint, Fukunaga and Flick (1985) (in a paper on NN methods of Bayes risk estimation) used both distance-weighted and unweighted distance measures and obtained lower classification error rates when using the weighted measures.

The first aim of section 6.2 is to show that the following hypothesis is not generally applicable:

Hypothesis 1 That the error rate of the unweighted k-NN rule is lower than that of any weighted k-NN rule even when the number of prototype samples is finite.

In section 6.2.2 the basic differences between the cases of finite and infinite prototype sets are discussed, expressions for the conditional probability of error of a test sample are developed, and it is argued intuitively that under certain conditions a weighted rule may achieve a lower $\hat{e}_k(\mathbf{x})$ than an

unweighted rule. In section 6.2.3, a particular example (2-NN, 1 dimension, 2 classes, 2 prototype samples per class, particular class conditional probability density functions) is solved. It is shown analytically, for this particular case, that a suitably weighted NN rule gives a lower finite, expected (or average) probability of error (i.e. $E\{ \hat{e}_2(\mathbf{x}) \}$ with the expectation taking over all possible values of $\mathbf{x}$) than the corresponding unweighted rule for any prototype set (subject to the above restrictions) generated from the specified pdf's. This example may be regarded as confirming analytically what was shown or suggested experimentally for the problems studied in the above three cited references. It may also be regarded as a counter-example to Hypothesis 1.

The second aim of this section is to investigate Hypothesis 1 experimentally. It is shown in this section that the higher average frequency of misclassification observed by Bailey and Jain and by Morin and Raeside in the case of distance-weighted k-NN classification is due not to any inherent general property of weighted k-NN rules but rather to the particular weighting function used, that of Dudani (1976). In section 6.2.4 a generalised version of Dudani's rule is proposed and experimental results are presented to show that in some cases a lower finite, average frequency of misclassification can be achieved when using the weighted measure.

6.2.2. Classification error of nearest neighbour rules when the number of prototype samples is finite

For the sake of simplicity, a two-class problem is considered. Formally, let S_N be a set of N classified prototype samples $\{ (\mathbf{x}_1, \theta_1), (\mathbf{x}_2, \theta_2), \ldots, (\mathbf{x}_N, \theta_N) \}$ where θ_i is the label of the prototype sample $\mathbf{x}_i$ and may be assigned to either class ω_1 or class ω_2. Let $\mathbf{x}$ be the unclassified test sample and $(\mathbf{x}(j), \theta(j))$ be the j-th nearest neighbour (j=1, 2, ..., k) to $\mathbf{x}$.

The conditional asymptotic probability of error $e_k(\mathbf{x})$ of the k-NN rule for a test sample $\mathbf{x}$ can be expressed in terms of the a posteriori probabilities of the two classes ω_1 and ω_2. For r=1,2,...,

$$e_{2r-1}(\mathbf{x}) = \eta_1(\mathbf{x}) \sum_{j=0}^{r-1} \binom{2r-1}{j} \eta_1^j(\mathbf{x}) \eta_2^{2r-j-1}(\mathbf{x})$$

$$+ \eta_2(\mathbf{x}) \sum_{j=0}^{r-1} \binom{2r-1}{j} \eta_1^{2r-j-1}(\mathbf{x}) \eta_2^j(\mathbf{x}) \qquad (1a)$$

$$e_{2r}(\mathbf{x}) = \eta_1(\mathbf{x}) \sum_{j=0}^{r-1} \binom{2r}{j} \eta_1^j(\mathbf{x}) \eta_2^{2r-j}(\mathbf{x})$$

$$+ \eta_2(\mathbf{x}) \sum_{j=0}^{r-1} \binom{2r}{j} \eta_1^{2r-j}(\mathbf{x}) \eta_2^j(\mathbf{x}) + g(s_{2r}) \qquad (1b)$$

where $\eta_i(\mathbf{x})$, i=1,2 is the a posteriori probability of class ω_i,

s_{2r} is the set of 2r nearest neighbours, and

g is an error function due to the resolution of ties (which may or may not depend on s_{2r}. For generality of notation, g is denoted by $g(s_{2r})$ in the above equation).

(c.f. Devijver and Kittler, 1982, page 77).

When the number of prototype samples is finite, the above equations do not hold in general because the a posteriori probability of ω_i given the j-th nearest neighbour $\mathbf{x}(j)$ for any j=1,2,...,k will not in general equal the a posteriori probability of ω_i given the test sample $\mathbf{x}$. Thus, following Fukunaga and Flick (1985), the a posteriori probability of ω_i given $\mathbf{x}(j)$ has to be expressed as

$$\eta_1(\mathbf{x}) + \eta_2(\mathbf{x}) = 1 \tag{2a}$$

$$\eta_1(\mathbf{x}(j)) = \eta_1(\mathbf{x}) \pm h_j \tag{2b}$$

$$\eta_2(\mathbf{x}(j)) = \eta_2(\mathbf{x}) \mp h_j \tag{2c}$$

where h_j is the amount of "distortion" (or deviation) between the a posteriori probabilities of $\mathbf{x}(j)$ and $\mathbf{x}$ (hereafter, h_j will be referred to as finite distortion). In the following development, for clarity, the positive sign in equation (2b) is arbitrarily chosen. With the help of equations (2), the finite, conditional probability of error of the test sample $\mathbf{x}$ for a finite prototype set can be rewritten in the following way. For r=1,2,...,

$$\hat{e}_{2r-1}(\mathbf{x}) = e_{2r-1}(\mathbf{x}) + \sum_{q=1}^{2r-1} A_r^{(q)} \overline{h_{2r-1}^{(q)}} \qquad (3a)$$

$$\hat{e}_{2r}(\mathbf{x}) = e_{2r}(\mathbf{x}) + \sum_{q=1}^{2r} B_r^{(q)} \overline{h_{2r}^{(q)}} + \{\hat{g}(s_{2r}) - g(s_{2r})\}. \qquad (3b)$$

In equations (3), $\overline{h_k^{(q)}}$ denotes the average of all q-fold products involving different factors drawn from $\{h_p : p=1,2,\ldots,k\}$. For instance,

$$\overline{h_k^{(1)}} = k^{-1} \sum_{p=1}^{k} h_p,$$

$$\overline{h_k^{(2)}} = \binom{k}{2}^{-1} \sum_{p=1}^{k-1} \sum_{t=p+1}^{k} h_p h_t,$$

...,

$$\overline{h_k^{(k)}} = \prod_{p=1}^{k} h_p.$$

Also in equations (3), $\{A_r^{(q)}\}$ and $\{B_r^{(q)}\}$ are functions of $\eta_1(\mathbf{x})$ and $\eta_2(\mathbf{x})$, independent of the $\{h_p\}$. The fact that, in general, $B_r^{(1)} = 0$, is noted by Fukunaga and Flick (1985).

To go further, the tie breaking mechanism has to be specified in detail. (Note that, in Fukunaga and Flick's results ties were "rejected" from consideration.) Suppose, for instance, that ties are broken at random. Then, for r=1,2,...,

$$g(s_{2r}) = (1/2)\binom{2r}{r} \eta_1^r(\mathbf{x})\,\eta_2^r(\mathbf{x}).$$

In this case, it turns out that

$$e_{2r}(\mathbf{x}) = e_{2r-1}(\mathbf{x}). \tag{4}$$

The result in equation (4) is established by the same argument as in Devijver and Kittler (1982), p.101 for the expected values, averaged over **x**.

It does not follow, however, that $\hat{e}_{2r}(\mathbf{x}) = \hat{e}_{2r-1}(\mathbf{x})$, although interesting relationships do appear. The following conjecture was suggested by Titterington (Macleod et al, 1987):

$$\hat{e}_{2r-1}(\mathbf{x}) = e_{2r-1}(\mathbf{x}) + \sum_{q=1}^{2r-1} A_r^{(q)} \overline{h_{2r-1}^{(q)}} \tag{5a}$$

$$\hat{e}_{2r}(\mathbf{x}) = e_{2r-1}(\mathbf{x}) + \sum_{q=1}^{2r-1} A_r^{(q)} \overline{h_{2r}^{(q)}} \tag{5b}$$

As yet, there is no general proof of this result,

although it is easy to check that the coefficient of $\overline{h_{2r}^{(2r)}}$ in equation (5b) is indeed zero. It is less easy but possible to confirm that the important first-order terms in equations (5) do match (Macleod et al, 1987). Titterington (Macleod et al, 1987) has pointed out the following interesting statistical conclusions from equations (5):

(a) If "$\mathbb{E}$" denotes expectation over the distortions $\{h_p\}$ and if the prototype set is very large, then

$$\mathbb{E}\,\hat{e}_{2r}(\mathbf{x}) = \mathbb{E}\,\hat{e}_{2r-1}(\mathbf{x}).$$

(b) If ties are broken at random and not simply rejected as Fukunaga and Flick (1985), then $\hat{e}_{2r}(\mathbf{x})$ does contain first-order terms in the distortion.

(c) $\hat{e}_{2r}(\mathbf{x})$ and $\hat{e}_{2r-1}(\mathbf{x})$ differ in moments of higher order than first.

From the above equations, it is clear that these finite distortions $\{h_p: p=1, 2, \ldots, k\}$ will in general affect the finite, conditional probability of error of the k-NN rule whenever a finite set of prototype samples is used. If a weighting function is carefully constructed, the a posteriori probability of ω_i given the j-th nearest neighbour $\mathbf{x}(j)$ may be altered so that the finite distortion h_j will be reduced. This is equivalent to the replacement of h_j in equations (2b) and (2c) by a new variable h'_j, where $h'_j \leq h_j$. On substituting this set of variables $\{h'_p: p=1, 2, \ldots, k\}$ into equations (5) it is not difficult to see, that for a finite number of prototype samples, the finite, conditional probability of error of the sample $\mathbf{x}$ may be less with a weighted rule than with the unweighted rule.

Section 6.2.3 we examine $\hat{e}_1(\mathbf{x})$ and $\hat{e}_2(\mathbf{x})$ in more detail for a particular example. As we shall see, the 1-NN rule is often equivalent to a weighted 2-NN rule, so that the comparison between $\hat{e}_1(\mathbf{x})$ and $\hat{e}_2(\mathbf{x})$ has implications as a comparison between unweighted and weighted 2-NN rules. From the above formulae, the differential effect in distortion is dictated by the difference in the properties of h_1 and $\overline{h_2^{(1)}} = (1/2)(h_1 + h_2)$.

6.2.3. A simple example

It is possible to develop expressions for the error probabilities for a finite set of prototypes in terms of integrals of pdf's. However evaluation of these integrals will be feasible only if highly simplified models for the pdf's are assumed. In this section an analytical calculation of the finite, expected error probabilities for one particular, highly simplified example is presented. Here, the expectation operation will be taken over all the possible values of a test sample **x**.

Recalling equation (13) from section 5.2.4.2 and the fact that when ties are resolved (i.e. when no rejections are allowed), probability of error = 1 - probability of correct classification, the finite, conditional probability of error can be written as

$$\hat{e}_k(\mathbf{x}) = 1 - \sum_{i=1}^{C} \eta_i(\mathbf{x}) \Pr\{ \hat{\omega} = \omega_i \mid S(k) \}. \qquad (6)$$

The expectation of equation (6) gives the finite, expected error probability and is given by

$$E\{\hat{e}_k(x)\} = 1 - \int \sum_{i=1}^{c} \eta_i(x) \Pr\{\hat{\omega} = \omega_i \mid S(k)\} p(x) dx \quad (7)$$

where $p(x)$ is the mixture (or unconditional) pdf of x and c is the number of classes. Using Bayes rule equation (7) can also be written as

$$E\{\hat{e}_k(x)\} = 1 - \int \sum_{i=1}^{c} f_i(x) \Pr\{\hat{\omega} = \omega_i \mid S(k)\} \pi_i dx \quad (8)$$

where $f_i(x)$ is the class-conditional pdf of x and π_i is the a priori probability of class ω_i. It is possible to specify the a posteriori probabilities and the mixture pdf of x and then to calculate a solution for equation (7). However $p(x)$ is related to the class-conditional pdf's and the a priori probabilities, and it is easier to specify $f_i(x)$ and π_i. Thus, it is easier to compute a solution for $E\{\hat{e}_k(x)\}$ using equation (8) than with equation (7). Furthermore, the second term on the right-hand side of equation (8) (and equation 7) is the finite, expected probability of correct classification. Thus, the problem reduces to finding the analytical solutions for the finite, expected probability of correct classification for the weighted and unweighted k-NN rules.

It is worth remarking that, even in the example below, the calculation (see below) of the finite expected probability of correct classification from the assumed probability density functions is very complicated and tedious.

The example

Again a two-class one-dimensional problem with a finite number of prototype samples is considered: specifically, the prototype set S_N is assumed to have four elements, two from each class. The a priori class probabilities for the test sample are Π_1 and Π_2 with $\Pi_1 + \Pi_2 = 1$, and particular class-conditional probability density functions (namely those defined in equations 9 and 10, below) are assumed. A weighted 2-NN rule is compared with the corresponding unweighted one, 2 being the least number of nearest neighbours which allows the use of a weighted rule. For this example, it will be shown that there exists a weighted 2-NN rule which will achieve a higher finite, expected probability of correct classification than the (unweighted) k-NN rule. Since this conclusion applies for any set of prototype samples (two per class) that could be generated from the particular class-conditional pdf's assumed, we have in this way shown analytically by counter-example that Hypothesis 1 is not generally applicable.

To further simplify the notation, let a and b be the first and second nearest neighbours respectively (i.e. a=x(1) and b=x(2)) and let w_1 and w_2 be the corresponding weights for the

first and the second nearest neighbour, respectively, such that $w_1 > w_2$. Further, let $f_1(x)$ and $f_2(x)$ be the class-conditional probability density functions for class ω_1 and class ω_2 respectively and let them take the following simple forms (see figure 6.1)

$$f_1(x) = \begin{cases} (3/4)(1 - x^2) & |x| < 1 \\ 0 & \text{otherwise} \end{cases} \qquad (9)$$

and

$$f_2(x) = \begin{cases} (3/4)(1 - (x-1)^2) & |x-1| < 1 \\ 0 & \text{otherwise.} \end{cases} \qquad (10)$$

Let $F_1(x)$ and $F_2(x)$ be the corresponding class-conditional cumulative distribution functions. From table 6.1 it is obvious that the weighted 2-NN rule provides the same decision as the unweighted 1-NN rule (note that, again, ties are resolved randomly). The problem can be broken down into two subproblems: the first one is to evaluate the finite, expected probability P_{1NN} of correct classification for the 1-NN rule, which in this particular example is also the equivalent finite, expected probability P_{W2NN} of correct classification for the weighted 2-NN rule, and the second one is to evaluate the finite, expected probability P_{2NN} of correct classification of the unweighted 2-NN rule. It is required to show, therefore, that $P_{1NN} > P_{2NN}$.

Note that, although the test sample x is allowed to

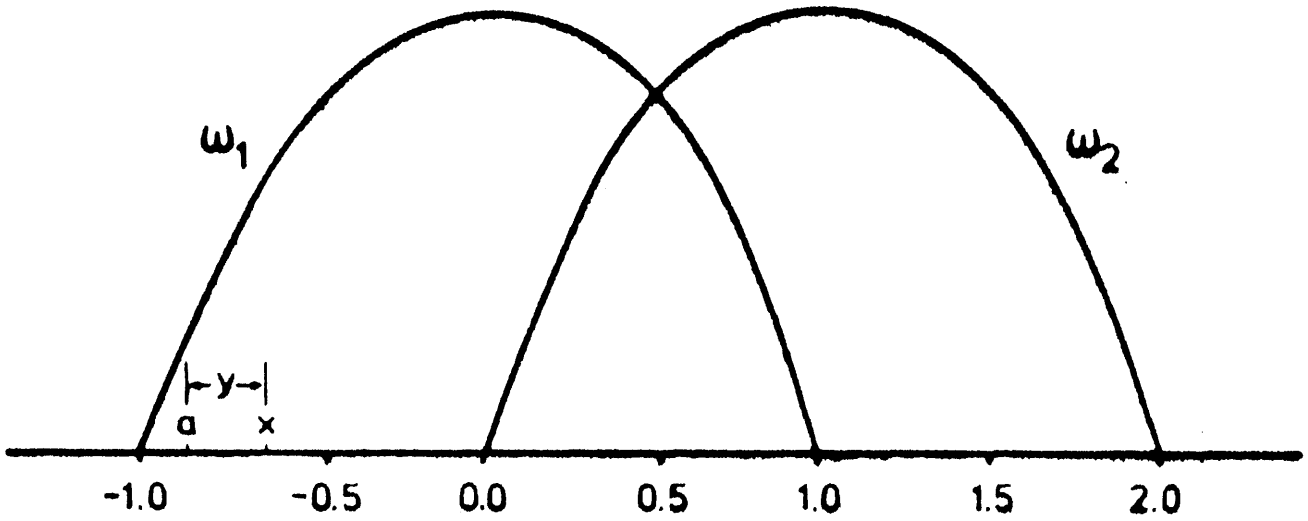


Figure 6.1 Probability density functions $f_1(x)$ and $f_2(x)$ of the two-class one-dimensional problem used in section 6.2.3.

Table 6.1 Decision table for the weighted 2-NN rule, unweighted 2-NN rule and unweighted 1-NN rule

	Class membership of nearest neighbour		Decision of the		
	a = x(1)	b = x(2)	Unweighted 1-NN rule	Weighted 2-NN rule	Unweighted 2-NN rule
Case A	1	1	1	1	1
Case B	1	2	1	1	*
Case C	2	1	2	2	*
Case D	2	2	2	2	2

Note * denotes ties condition has occurred. Ties were resolved randomly.

come from either class ω_1 or class ω_2, the structure of the prototype set (namely four samples, two from each class) will be taken as fixed.

First subproblem: evaluating P_{1NN}

It is obvious from table 6.1 that P_{1NN} is equal to the average, weighted with respect to the a priori class probabilities π_1 and π_2 for the test sample x, of the probability P_{A1}, that, given that the test sample comes from class ω_1, then its nearest neighbour a is also from class ω_1 and the corresponding probability P_{D1} associated with class ω_2. The former probability can be evaluated as follows.

Consider all possible choices x for the position of the new observation, X (figure 6.1). The probability that X lies within a particular small interval (x, x+dx) is

$$P(x \leq X \leq x+dx) \approx f_1(x)dx. \qquad (11)$$

Let y be the absolute value of the distance between x and its nearest neighbour a. If the neighbour is to belong to class ω_1, one of the two prototype samples which belong to class ω_1 must be within a small range dy near either x+y or x-y. The probability of this is

$$2\{f_1(x+y)dy + f_1(x-y)dy\}. \qquad (12)$$

The factor of 2 is present because the nearest neighbour could be either of the 2 prototype samples which belongs to class ω_1. The other class ω_1 sample is outside the interval (x-y,x+y), as are the two class ω_2 samples, which gives the following factors towards the probability P_{A1}:

$$\{1 - F_1(x+y) + F_1(x-y)\}\{1 - F_2(x+y) + F_2(x-y)\}^2. \qquad (13)$$

In principle, P_{A1} can be obtained by multiplying expressions (11), (12) and (13) together and integrating over x and y. However, the problem is complicated by the fact that $f_j(x) = 0$ (j=1,2) outside finite intervals. It is therefore necessary to split the range of x into four subranges, i.e. $-1 \leq x \leq -1/2$, $-1/2 \leq x \leq 0$, $0 \leq x \leq 1/2$ and $1/2 \leq x \leq 1$ (see figure 6.1). When x is in the range $-1 \leq x \leq -1/2$, then y is in the range $0 < y < 1-x$. On splitting this range of y up into three subintervals, namely $0 < y < 1+x$, $1+x < y < -x$ and $-x < y < 1-x$, one sees that

(i) at $y = 1+x$, $x-y = -1$, so $F_1(x-y)$ is zero,

(ii) until $y = -x$, $1 - F_2(x+y) + F_2(x-y) = 1$, and

(iii) for $y > 1+x$, the nearest neighbour has to be at x+y.

Bearing in mind these and other considerations, one can write the contribution of the range $-1 \leq x \leq -1/2$ to the probability P_{A1} as

$$2\int_{-1}^{-1/2}\int_{0}^{1+x} f_1(x)\,[f_1(x+y)+f_1(x-y)]\,[1-F_1(x+y)+F_1(x-y)]\,dydx$$
$$+\ 2\int_{-1}^{-1/2}\int_{1+x}^{-x} f_1(x)\,f_1(x+y)\,[1-F_1(x+y)]\,dydx$$

$$+ 2\int_{-1}^{-\frac{1}{2}}\int_{-x}^{1-x} f_1(x) f_1(x+y)[1-F_1(x+y)][1-F_2(x+y)]^2 dydx.$$

By similar arguments, the contributions from the other three subranges of x can be written as

$$2\int_{-\frac{1}{2}}^{0}\int_{0}^{-x} f_1(x)[f_1(x+y)+f_1(x-y)][1-F_1(x+y)+F_1(x-y)]dydx$$

$$+ 2\int_{-\frac{1}{2}}^{0}\int_{-x}^{1+x} f_1(x)[f_1(x+y)+f_1(x-y)][1-F_1(x+y)+F_1(x-y)][1-F_2(x+y)]^2 dydx$$

$$+ 2\int_{-\frac{1}{2}}^{0}\int_{1+x}^{1-x} f_1(x) f_1(x+y)[1-F_1(x+y)][1-F_2(x+y)]^2 dydx$$

$$+ 2\int_{0}^{\frac{1}{2}}\int_{0}^{x} f_1(x)[f_1(x+y)+f_1(x-y)][1-F_1(x+y)+F_1(x-y)][1-F_2(x+y)+F_2(x-y)]^2 dydx$$

$$+ 2\int_{0}^{\frac{1}{2}}\int_{x}^{1-x} f_1(x)[f_1(x+y)+f_1(x-y)][1-F_1(x+y)+F_1(x-y)][1-F_2(x+y)]^2 dydx$$

$$+ 2\int_{0}^{\frac{1}{2}}\int_{1-x}^{1+x} f_1(x) f_1(x-y) F_1(x-y)[1-F_2(x+y)]^2 dydx$$

$$+ 2\int_{\frac{1}{2}}^{1}\int_{0}^{1-x} f_1(x)[f_1(x+y)+f_1(x-v)][1-F_1(x+v)+F_1(x-y)][1-F_2(x+y)+F_2(x-y)]^2 dydx$$

$$+ 2\int_{\frac{1}{2}}^{1}\int_{1-x}^{x} f_1(x) f_1(x-y) F_1(x-y)[1-F_2(x+y)+F_2(x-y)]^2 dydx$$

$$+ 2\int_{\frac{1}{2}}^{1}\int_{x}^{1+x} f_1(x) f_1(x-y) F_1(x-y)[1-F_2(x+y)]^2 dydx.$$

Computation of the above twelve integrals gives $P_{A1} = 78.05\%$.

The probability P_{D1} of obtaining both x and a from class ω_2 can be found by similar methods and, by symmetry, the same numerical solution can be obtained. Thus the finite, expected probability of correct classification for the 1-NN and the weighted 2-NN rules is

$$P_{1NN} = P_{W2NN} = \Pi_1 P_{A1} + \Pi_2 P_{D1} = 78.05\%.$$

Second subproblem: evaluating P_{2NN}

P_{2NN} is equal to the average of the four cases listed in table 6.1: the probability P_{A2} that, given that the test sample x comes from class ω_1, its two nearest neighbours a and b are also from class ω_1; the corresponding probability P_{D2} associated with class ω_2; the probability P_{B2} that, given that x comes from class ω_1, its nearest neighbour a comes from class ω_1, whilst its second nearest neighbour b comes from class ω_2; and the probability P_{C2} which has the same description as P_{B2} except that the classifications of a and b are reversed.

P_{A2} can be evaluated using the same method as in the first problem. Let z be the absolute value of the difference between the distances of the test sample from its two nearest neighbours a and b. Expressions (11) and (12) still hold. Expression (13) has to be modified into the following two expressions of probability

$$\{f_1(x+y+z) + f_1(x-y-z)\}dz \tag{14}$$

and

$$\{1 - F_2(x+y+z) + F_2(x-y-z)\}^2. \tag{15}$$

Again, the ranges of x, a and b have to be split into subintervals. Thus an expression for P_{A2} can be found: it involves twenty-four integrals (see appendix 6.A). Evaluation of these integrals gives $P_{A2} = 53.64\%$.

Similarly, by symmetry, $P_{D2} = 53.64\%$. P_{B2} and P_{C2} correspond to situations where ties can occur. Suppose that, in this case, the decision is made for class ω_1 or class ω_2 at random. Then

$$\begin{aligned} P_{2NN} &= \pi_1 P_{A2} + \pi_2 P_{D2} + (1/2)(\pi_1+\pi_2)(P_{B2}+P_{C2}) \\ &= P_{A2} + (1/2)(P_{B2}+P_{C2}) \\ &< P_{A2} + (1/2)(1-P_{A2}) \\ &= (53.64 + 23.68)\% \\ &= 77.32\%. \end{aligned}$$

With reference to the above inequality, note that if

$$P_{E2} = 1 - P_{A2} - P_{B2} - P_{C2}$$

then P_{E2} denotes the probability of having a test sample from class ω_1 for which the two nearest neighbours in the prototype data are both from class ω_2.

Comparing the solution of the first subproblem with that of the second subproblem, it is obvious that

$$P_{W2NN} \ (= P_{1NN}) > P_{2NN}.$$

Hence, for this simple example, it has been shown for any set of prototype samples (two per class) generated from the assumed class-conditional pdf's that there exists a weighted k-NN rule which will perform better in terms of correct classification than

the traditional (unweighted) k-NN rule when the number of prototype samples is finite.

6.2.4. A generalisation of Dudani's weighting function

For $j < k$, let d_j be the distance (which can be defined by any suitable distance metric) between the j-th nearest neighbour and an unclassified sample **x**. Dudani (1976) defines the weight w_j for the j-th nearest neighbour as

$$w_j = \begin{cases} \dfrac{d_k - d_j}{d_k - d_1} & d_k \neq d_1 \\ 1 & d_k = d_1 \end{cases} \qquad (16)$$

(see section 5.2.2.3.3). The distance-weighted k-NN classification rule assigns a label θ (for the unclassified test sample **x**) to that class with the largest total weight.

The weighting function in equations (16) effectively removes the k-th nearest neighbour from participating in the k-NN classification. This is because if $d_j = d_k$ in equations (16), then $w_k = 0$. Let E_k be the average asymptotic error probability of the k-NN rule. It has been shown (Devijver and Kittler, 1982) that $E_k \leq E_{k-1}$. Therefore, from the above arguments and those in section 6.2.2, one would expect that the distance-weighted k-NN rule would not classify better than the traditional k-NN rule

with ties resolved by some means.

Equations (16) are therefore modified to

$$w_j = \begin{cases} \dfrac{(d_s - d_j) + \alpha (d_s - d_1)}{(1 + \alpha)(d_s - d_1)} & d_s \neq d_1 \\ 1 & d_s = d_1 \end{cases} \qquad (17)$$

where

d_s is the distance between the test sample $\mathbf{x}$ and its s-th nearest neighbour $\mathbf{x}(s)$ (s=k, k+1, ...), and

α is a positive constant.

It is obvious that when $d_s = d_k$ and $\alpha = 0$, equations (17) become the original weighting function proposed by Dudani. Since there are many possible versions of equations (17), only the following three special cases will be experimented with in this section.

(Case 1) when $\alpha = 0$ and $\mathbf{x}(s)$ depends on the value of k, i.e. s = pk where p=2,3,.... The first two values of p (i.e. p = 2 and p = 3) were investigated;

(Case 2) when $\alpha = 0$ and $s \neq k$ and s does not depend on the value of k. Two such examples have been experimented with, i.e. s = N and s = (N/c), where again N = total number of prototype samples, and c = number of classes;

(Case 3) when s = k and α varies. Again, two such examples

were investigated, namely, with α = 1.0 and α = 2.0.

The experiment

The new experiment that has been conducted was similar to that described by Dudani (1976) but used the weighting functions of equations (17). The sample set was a three-class bivariate data set generated by a random number generator using NAG routines on a GEC 4180 computer. 3000 test samples were generated. The experiment was performed six times on this same set of test samples, each time with an independently generated set of 150 prototype samples, 50 from each class. Figures 6.2 to 6.7 show the six sets of 150 prototype samples used for the experiment. For each of the above three special cases, the sample-based finite, average frequency of misclassification $\bar{P}_e$ or simply averaged probability of error (as estimated by a simple error counting method) over the six runs is plotted against the number of neighbours k used for classification. The results are shown in figures 6.8 to 6.10. For comparison purposes, the corresponding averaged probability of error estimated by the same method (a) for the k-NN rule with c-class ties broken randomly by simulating a toss of a c-faced fair die, and (b) for Dudani's distance-weighted k-NN rule, are also shown in the figures.

Although the above three special cases are selected in a rather ad hoc manner, figures 6.8 to 6.10 show that in some cases Dudani's rule with the modified distance weighting has

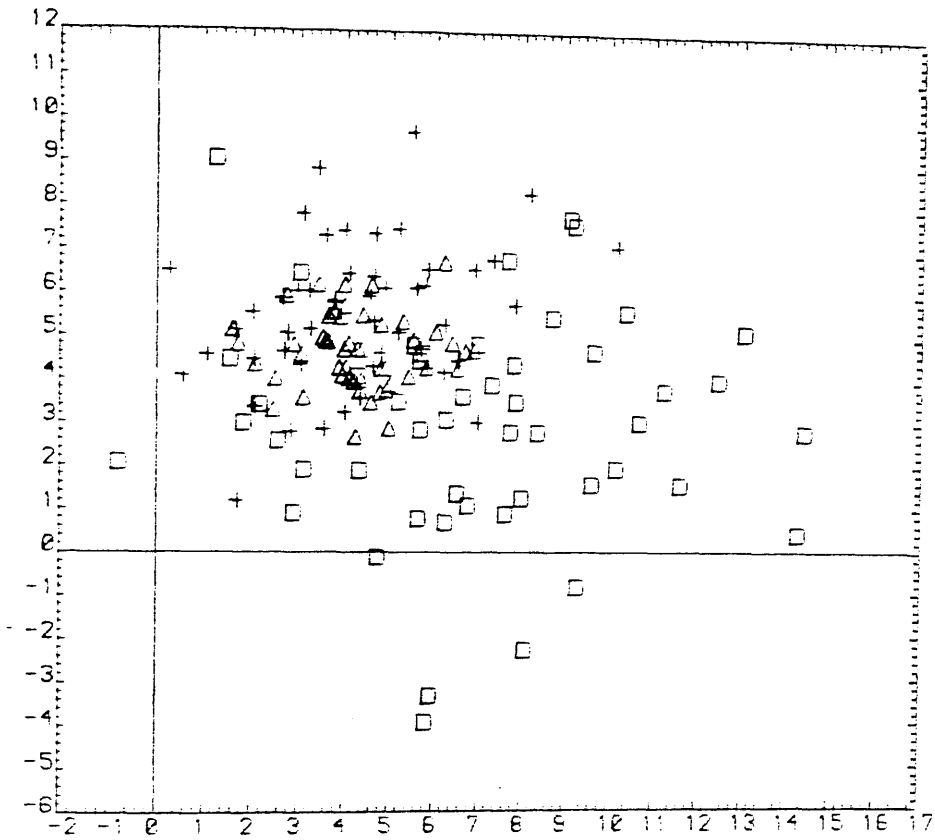


Figure 6.2 The first 3-class 2-dimensional Gaussian prototype set used in the experiment in section 6.2.4 of the text. (The distributions of the three classes of prototype samples are described in Dudani's paper, 1976.) (△) denotes first class prototypes, (+) denotes second class prototypes and (□) denotes third class prototypes.

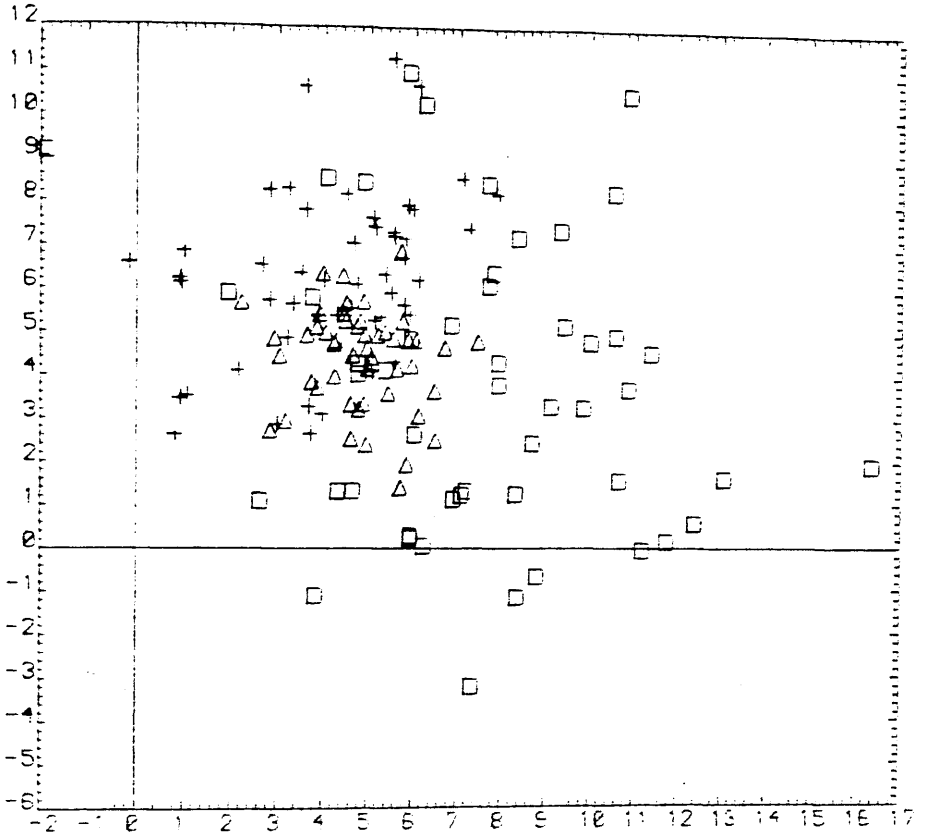


Figure 6.3 The second 3-class 2-dimensional Gaussian prototype set used in the experiment in section 6.2.4 of the text. Symbols as in figure 6.2.

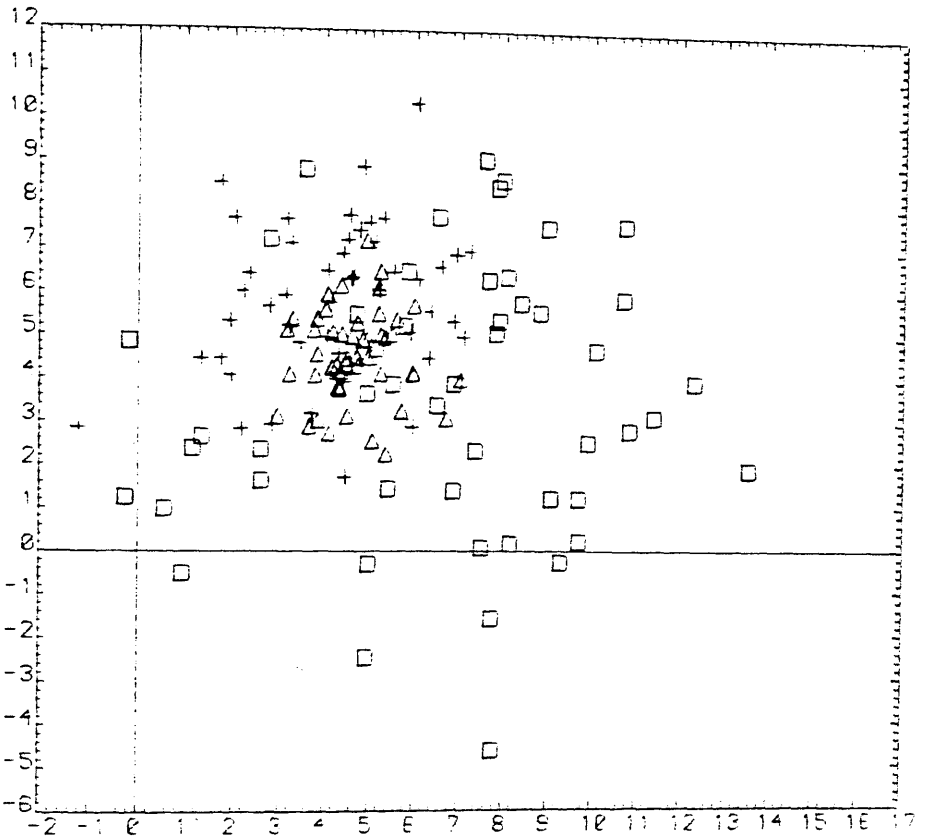


Figure 6.4 The third 3-class 2-dimensional Gaussian prototype set used in the experiment in section 6.2.4 of the text. Symbols as in figure 6.2.

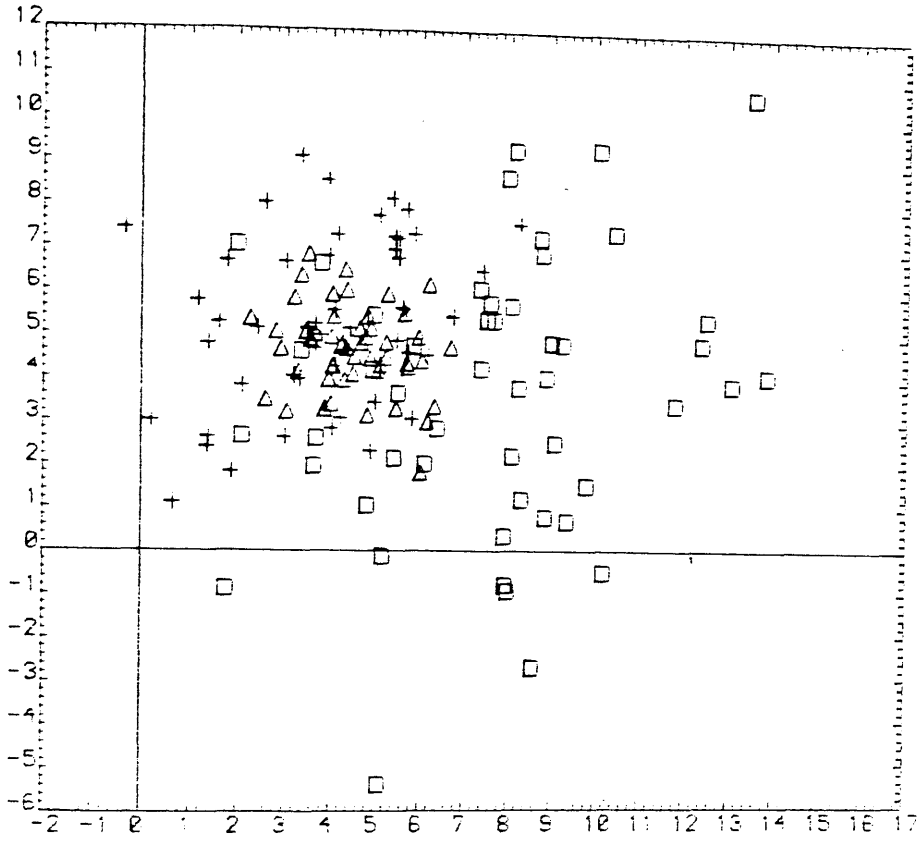


Figure 6.5 The fourth 3-class 2-dimensional Gaussian prototype set used in the experiment in section 6.2.4 of the text. Symbols as in figure 6.2.

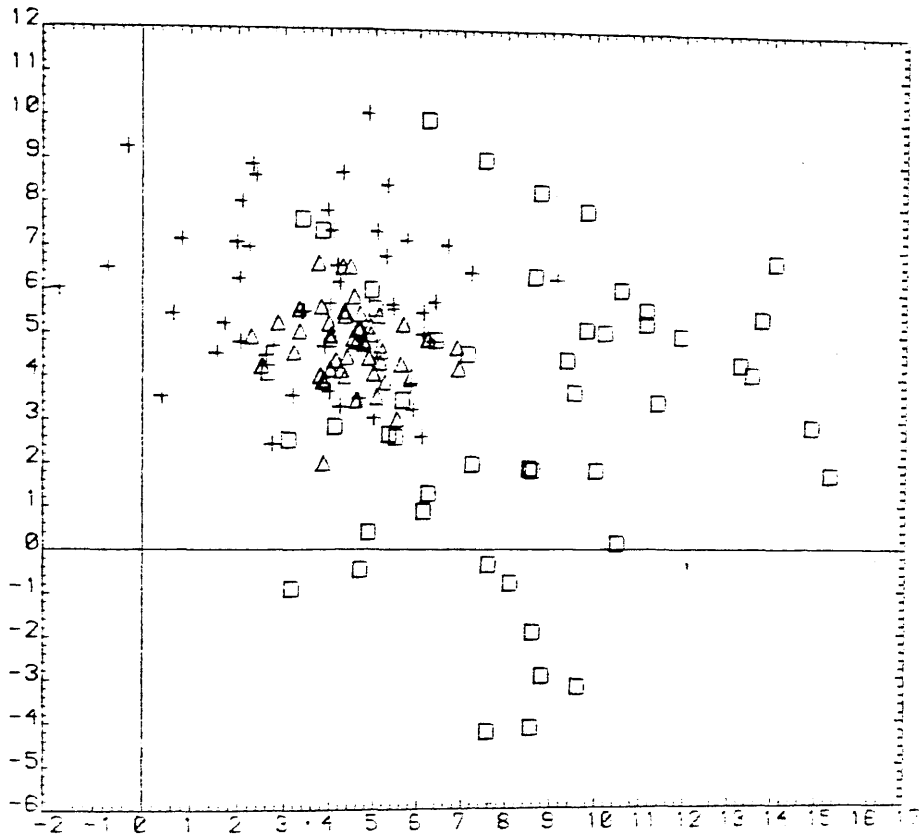


Figure 6.6 The fifth 3-class 2-dimensional Gaussian prototype set used in the experiment in section 6.2.4 of the text. Symbols as in figure 6.2.

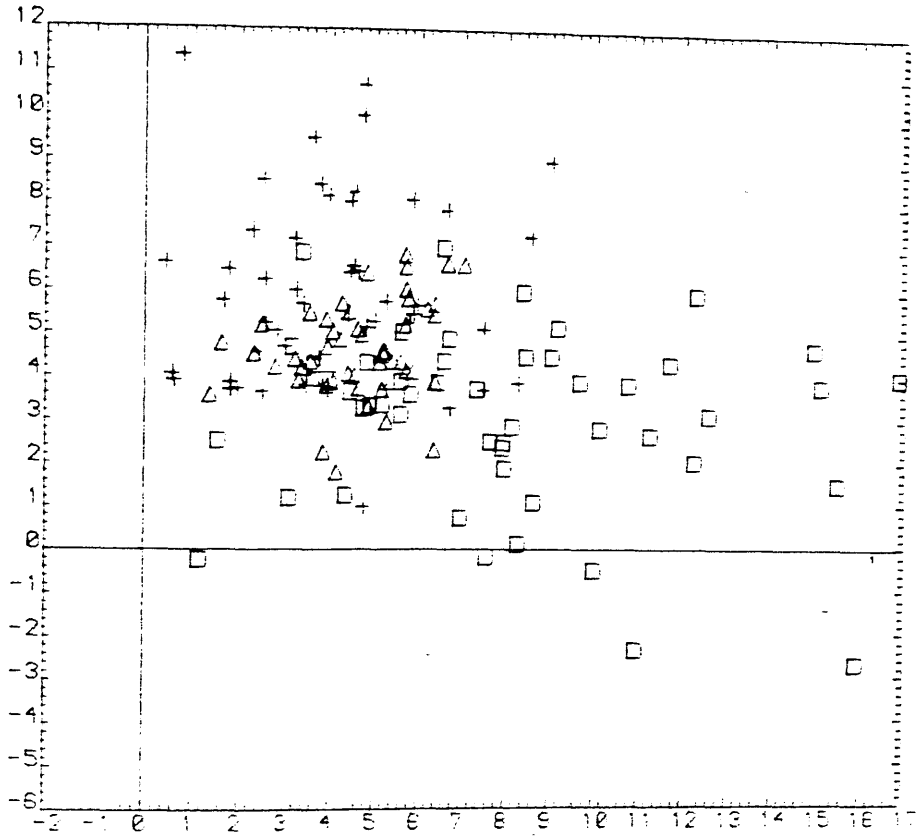


Figure 6.7 The sixth 3-class 2-dimensional Gaussian prototype set used in the experiment in section 6.2.4 of the text. Symbols as in figure 6.2.

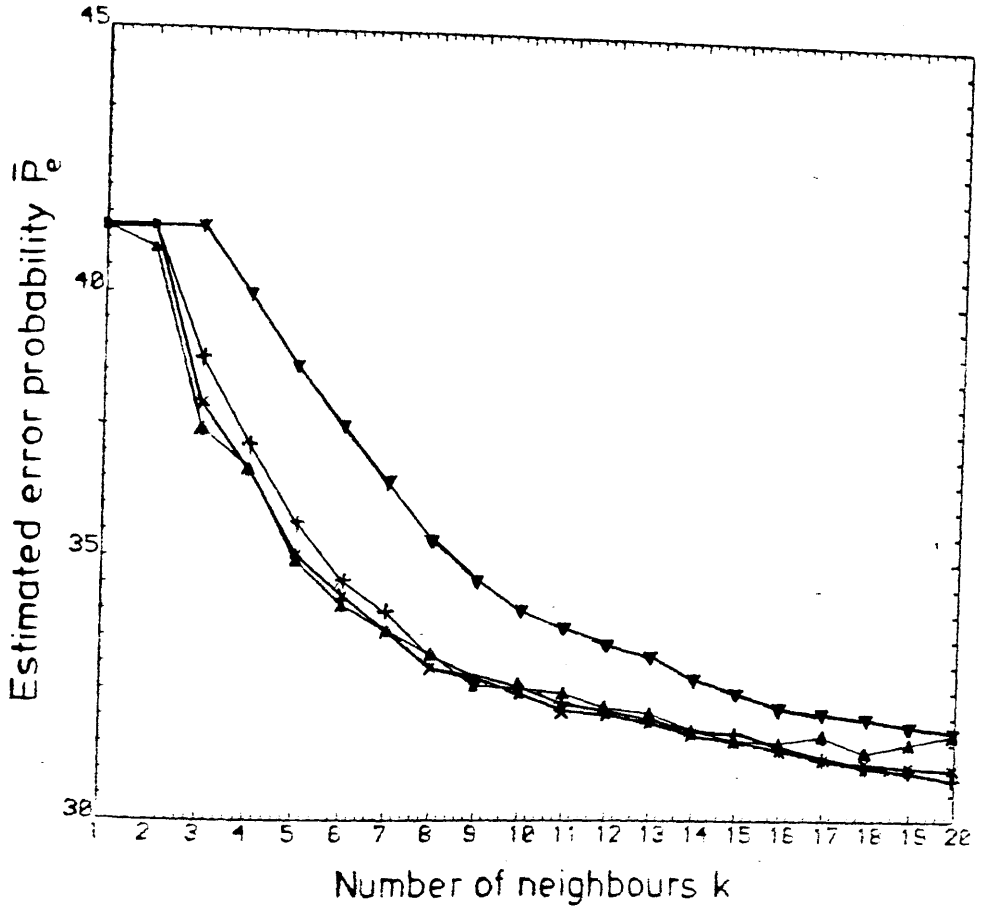


Figure 6.8 Finite, average probability of error $\bar{P}_\epsilon$ averaged over six runs versus number of neighbours k for a distance-weighted k-NN rule using the weighting function of equations (17) with α = 0 and (i) s = 2k (+) and (ii) s = 3k (×). For comparison, $\bar{P}_\epsilon$ against k is also plotted for (i) k-NN rule (▲) and (ii) Dudani's distance-weighted k-NN rule (▼).

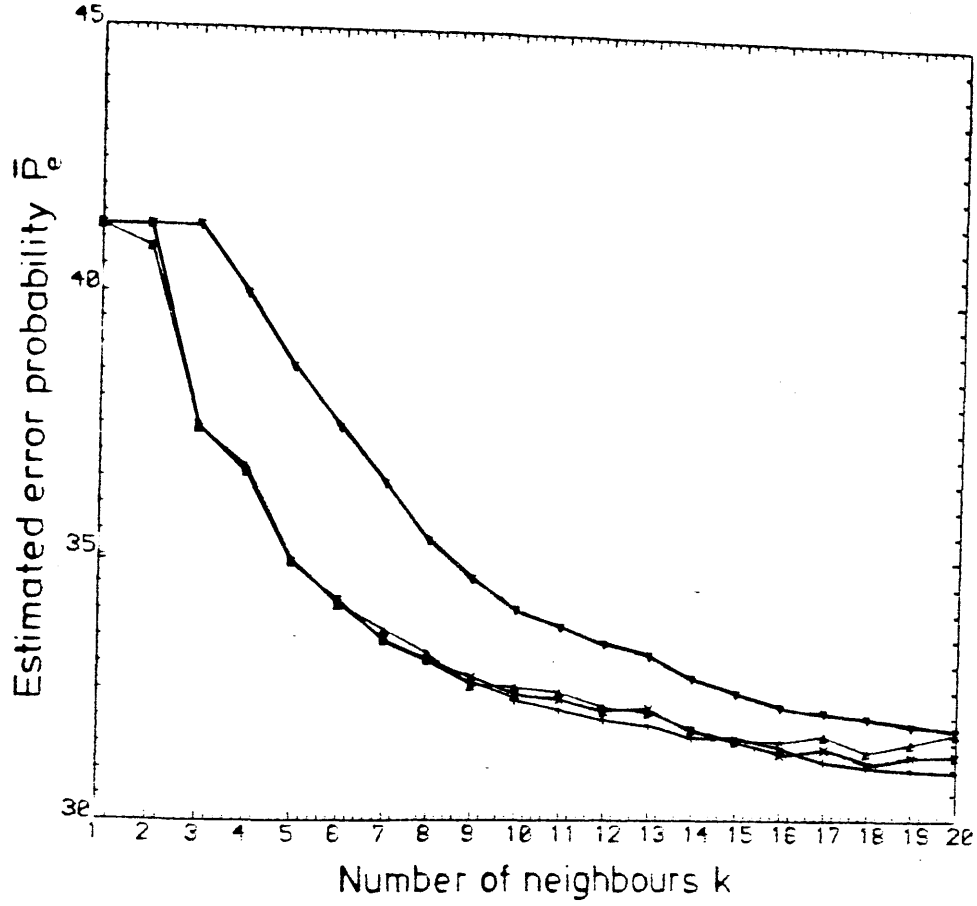


Figure 6.9 Finite, average probability of error $\bar{P}_e$ averaged over six runs versus number of neighbours k for a distance-weighted k-NN rule using the weighting function of equations (17) with α = 0 and (i) s = N (±) and (ii) s = N/c (⊗). Other symbols are as figure 6.8.

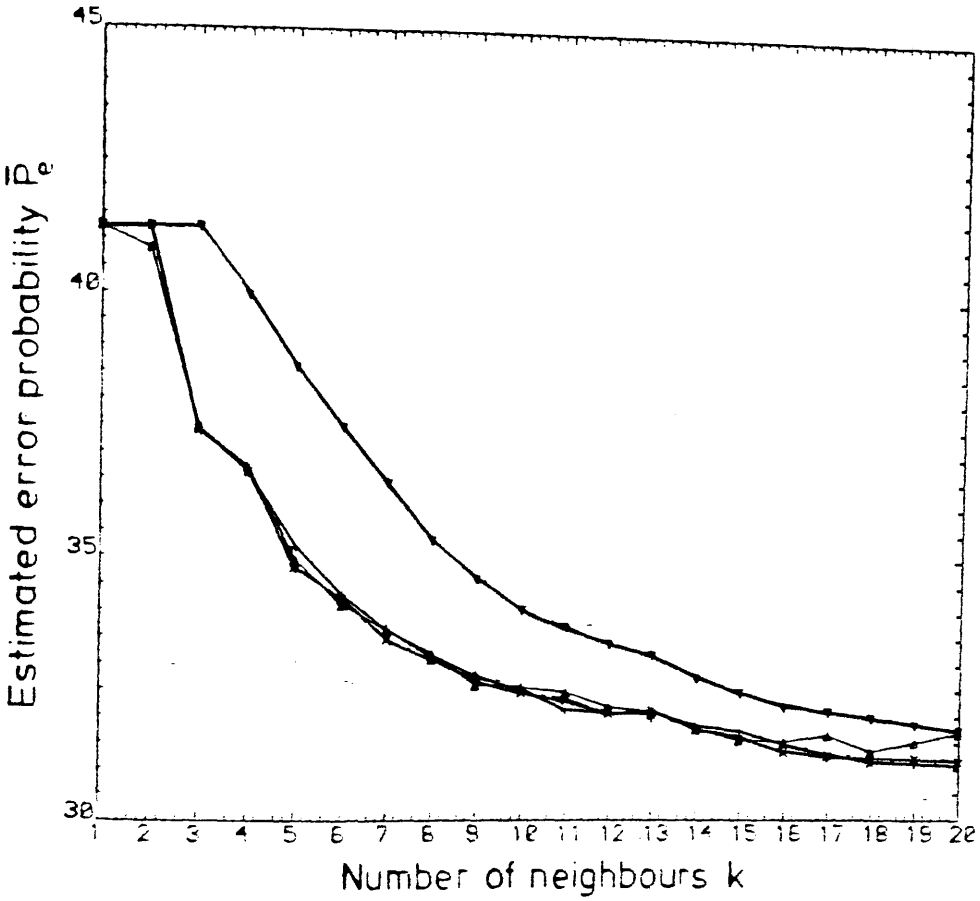


Figure 6.10 Finite, average probability of error $\bar{P}_e$ averaged over six runs versus number of neighbours k for a distance-weighted k-NN rule using the weighting function of equations (17) with s = k and (i) α = 1.0 (±) and (ii) α = 2.0 (×). Other symbols are as figure 6.8.

performed better than the unweighted k-NN rule, for example when $k \geq 10$ in figures 6.8 to 6.10. The worst performance of all is from Dudani's unmodified rule. It must be emphasized that although only the average plots were shown, the same observation (i.e. in some cases Dudani's rule with modified distance weighting has performed better than the unweighted k-NN rule) was noted in each run.

Figure 6.11 illustrates the effect of varying α (with $s = k$ in equations 17) from 0.2 to 4.0 for three values of k (namely, $k = 3$, 5 and 16) on one of the prototype sets (the sixth prototype set, figure 6.7) used in the above experiment. It can be seen that as α increases from 0.2 to 4.0, the (finite sample) estimated probability of error $\hat{P}_e$ decreases and then eventually levels off at different values of α for different values of k. Again, in some cases (namely $k = 5$ and 16) the weighted k-NN rule gives a lower frequency of misclassification than the unweighted k-NN rule.

These results are further experimental confirmation, in addition to three references cited in section 6.2.1, that Hypothesis 1 is not generally applicable. Indeed, in the experiment the weighted rule (modified Dudani) has sometimes outperformed the unweighted rule even with a ratio (number of prototype samples per class) / (dimensionality) as large as 25 - i.e. with a prototype set which would be considered effectively "infinite" for many purposes in classifier design. Thus Hypothesis 1 may not apply even for "fairly large" finite sample

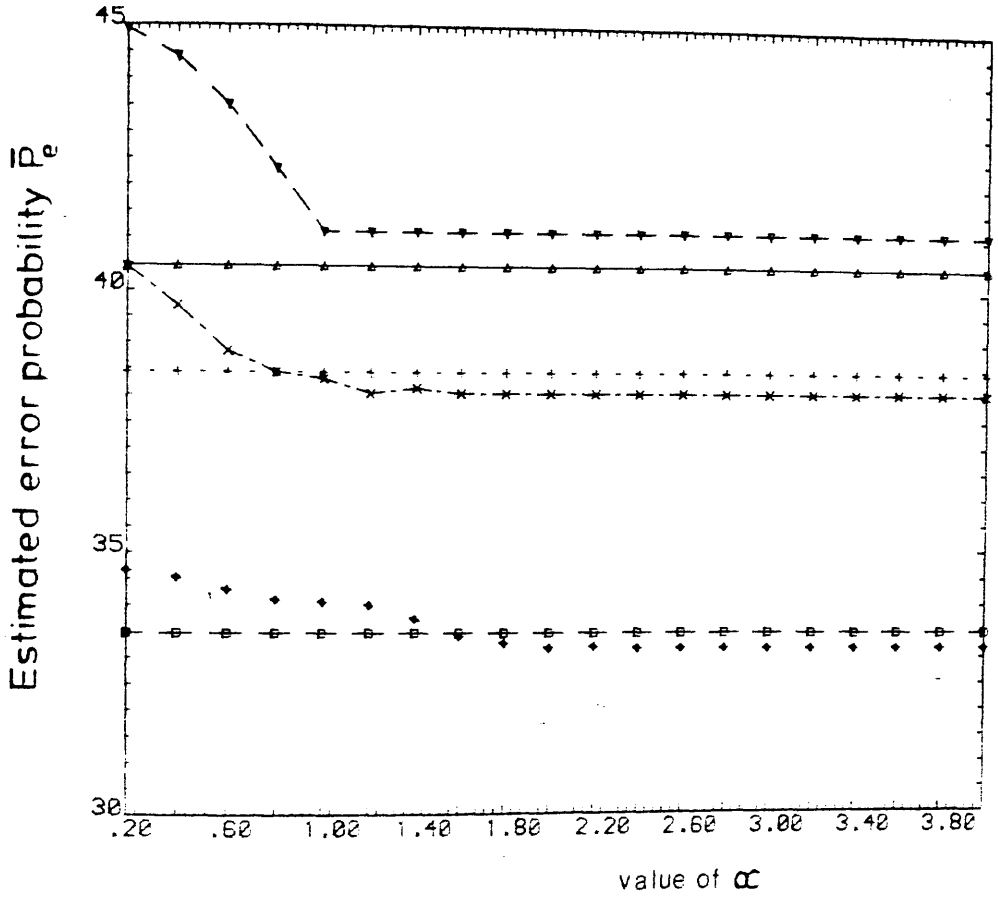


Figure 6.11 Finite, estimated probability of error $\hat{P}_e$ versus value of α for a distance-weighted k-NN rule using the weighting function of equations (17) at three different values of k, namely k = 3 (∇), k = 5 ($\times$) and k = 16 ($\diamond$). For comparison, $\hat{P}_e$ for the k-NN rule at these values is plotted as three straight lines, i.e. k = 3 ($\triangle$), k = 5 ($+$) and k = 16 ($\square$).

sets.

6.2.5. Discussion

Although in section 6.2.3 only a 2-class one-dimensional problem has been considered, the arguments can in principle be extended to higher dimensions. The extension to the multiclass problem will be more difficult and is beyond the scope of this thesis. This example can be regarded as illustrating how the equations in section 6.2.2 apply to a particular case. It can also be regarded as a counter-example to the hypothesis that the conclusion of Bailey and Jain for the conditional, asymptotic probability of error also applies when the prototype set is finite.

The following conclusions were arrived at in discussion with Professor Titterington. The asymptotic comparisons between two types of weighted k-NN rule have not been investigated. However, this is achieved by Bailey and Jain (1978) for the case where one of the weightings is special, namely, the unweighted rule. A natural approach would be to follow the argument of Bailey and Jain as far as their equation (9), but where their "m" and "t" now refer to nontrivial weightings. However, the next stage of the argument in their paper then fails. It seems that even asymptotic comparisons will require at least partial specification of characteristics such as the class-conditional densities f_1 and f_2 and, consequently, it will be necessary to undertake numerical work on the lines of that in section 6.2.3.

6.3 An alternative nearest neighbour classification scheme

6.3.1. Introduction

This section introduces a nearest neighbour classification scheme (Luk and Macleod, 1986) which is somewhat different from those surveyed in section 5.2.1. The basic idea will be described in the next section. As in the (k,ℓ)-NN and (k,ℓ_i)-NN rules, rejection arises naturally from the basic idea when the size of the prototype set is finite. Methods to reduce the probability of rejection are suggested in sections 6.3.4 and 6.3.6. Experiments on gaussian data are conducted to demonstrate the results of these modifications.

6.3.2. Proposed classification scheme

Let the set of N ordered nearest neighbours of a test sample $\mathbf{x}$ be defined over a suitable measurement, such as the Euclidean metric. This effectively maps a higher dimensional space to a one dimensional space. The proposed classification scheme is to examine the nearest neighbours one at a time in order of increasing distance from $\mathbf{x}$ until the number of votes ℓ_i for class ω_i (i=1,2,...,c, c = number of classes) has exceeded the number of votes ℓ_j for the nearest "rival" class ω_j ($j \neq i$, j=1,2,...,c) by a majority m, where $m \geq 1$. At this point the algorithm stops and $\mathbf{x}$ is assigned to class ω_i. The proposed rule is the same as the 1-NN rule when m = 1. For m > 1, it will continue to search the ordered set of nearest neighbours until a

decision $\hat{\omega}(m)$, which depends on m, is made. It differs from the (k,ℓ_i)-NN rule (section 5.2.2.2.3) in that (for $m > 1$) both k and ℓ_i are variable. Formally, for an infinite set of samples, the rule can be defined as

$$\hat{\omega}(m) = \omega_i \quad \text{if} \quad (\ell_i - \ell_j) \geq m, \forall j \neq i, \; i,j=1,2,\ldots,c, \; m \geq 1 \quad (18$$

The convergence property of the rule, like that of the k-NN rule (section 5.2.4), is guaranteed if $h \rightarrow \infty$ and $h/N \rightarrow 0$ as $N \rightarrow \infty$, where

$$h = \sum_{p=1}^{c} \ell_p .$$

(Basically, it is necessary to show that none of the h nearest neighbours fall outside a hypersphere of radius ξ as the number of prototype samples approaches infinity. Then the a posteriori probability of the nearest neighbours will approach the a posteriori probability of the test sample **x**. It then follows that the finite sample, conditional probability of error $\hat{e}_m(\mathbf{x})$ of the proposed rule will converge to a constant, say E_m which is the average asymptotic error probability of the proposed rule. The proof is similar to Devroye, 1981, and a simpler version can be found in Devijver and Kitter, 1982, page72-74).

In practical situations the prototype set S_N is finite and to ensure a small value of h/N (i.e. to avoid the "curse of finite sample size") only a small subset can be searched, say k_{max} ($k_{max} < N$), of the N classified prototypes (Devijver and

Kittler, 1982, page 108). Obviously, we can "reject" the test sample **x** if the algorithm has searched through k_{max} nearest neighbours without finding the requisite clear majority of votes in favour of any class. Thus equation (18) can be rewritten as

$$\hat{\omega}(m,k_{max}) = \begin{cases} \omega_i & \text{if } (\ell_i - \ell_j) \geq m, \ \forall\ j \neq i, \\ & i,j=1,2,\ldots,c, \ m \geq 1, \\ & \sum_{p=1}^{c} \ell_p \leq k_{max} \qquad (19a) \\ \omega_0 & \text{otherwise} \qquad (19b) \end{cases}$$

where ω_0 is the class of rejected samples and $\hat{\omega}(m,k_{max})$ indicates that the decision depends on both m and k_{max}. It is apparent that the value m should be small in order to prevent excessive rejection in equations (19). Since equations (19) rather than equation (18) would have to be used in all practical situations, the name "(m,k_{max})-nearest neighbour classification rule" is suggested for the proposed method. A simple experiment to investigate the effect(s) on finite probability of error (or frequency of misclassification) P_e and finite probability of rejection P_r at different values of m and k_{max} is now described. (For simplicity, the word "finite" in both P_e and P_r will be dropped in subsequent sections: it will be obvious from the content when "finite" is implied.)

6.3.3. First experiment

The same three classes of 2-dimensional Gaussian data as used in the experiment in section 6.2.4 is employed. Again, the data set comprised 150 prototype samples (50 samples from each class) and 3000 test samples. The test samples were classified

(a) by the (m,k_{max})-NN rule with various values of m and k_{max};

(b) by the k-NN rule with various values of k. for comparison;

(c) also for comparison, by the (k,ℓ)-NN rule with various values of k and with $\ell = \lceil k/c \rceil + 1 = s$ (say) for each k value; and

(d) again by the (k,ℓ)-NN rule with various values of ℓ for k = 15 (referred to as the "$(15,\ell)$-NN rule").

The error probability and the rejection probability were estimated by simple counting of the errors (rejections), i.e. the estimates were respectively $\hat{P}_e = N_{ea} / (N - N_r)$ and $\hat{P}_r = N_r / N$, where N = total number of test samples, N_r = number of rejected samples, and N_{ea} = number of errors for accepted samples (Note that $\hat{P}_e$ is the estimate of the error probability for accepted samples). Figures 6.12 and 6.13 show respectively $\hat{P}_e$ and $\hat{P}_r$ against m at different values of k_{max}. $\hat{P}_e$ and $\hat{P}_r$ for the k-NN rule, the (k,s)-NN rule and the $(15,\ell)$-NN rule are also shown in figures 6.12 and 6.13; for the k-NN rule as used here, all rejections are due to ties for maximum number of votes.

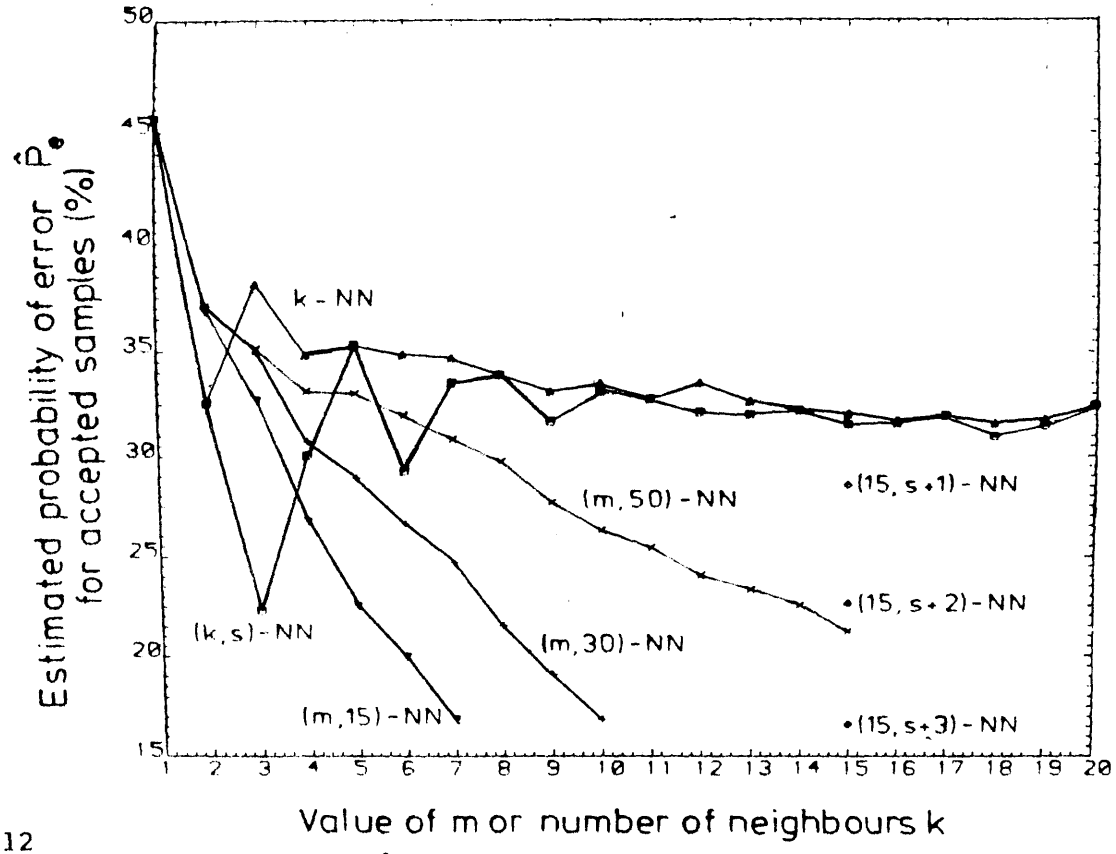


Figure 6.12 (i) Estimated probability of error $\hat{P}_e$ for accepted samples versus required majority m at different values of k_{max}; (ii) $\hat{P}_e$ versus k for traditional k-NN rule with ties taken as rejections; (iii) $\hat{P}_e$ versus k for the (k,s)-NN rule (with $s = \lceil k/c \rceil + 1$); (iv) $\hat{P}_e$ at several values of ℓ for the $(15,\ell)$-NN rule.

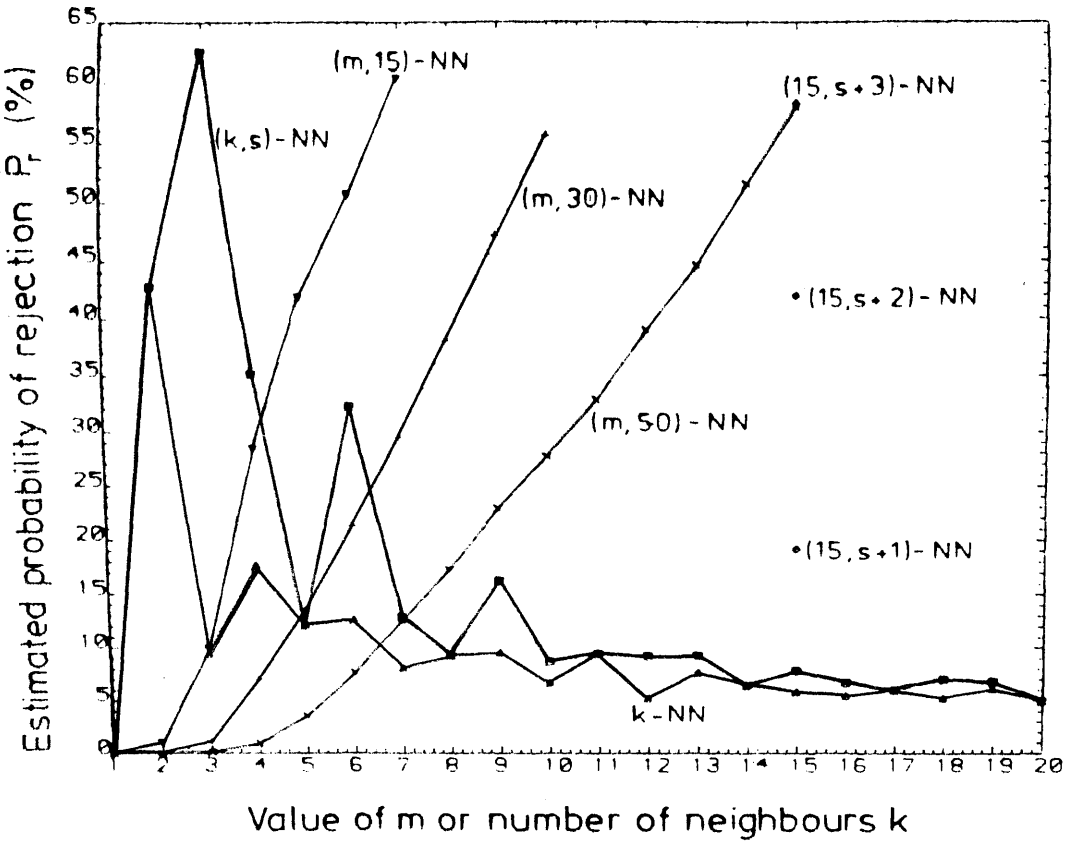


Figure 6.13 (i) Estimated probability of rejection $\hat{P}_r$ versus required majority m at different values of k_{max}; (ii) $\hat{P}_r$ versus k for traditional k-NN rule with ties taken as rejections; (iii) $\hat{P}_r$ versus k for the (k,s)-NN rule (with $s = \lceil k/c \rceil + 1$); (iv) $\hat{P}_r$ at several values of ℓ for the $(15,\ell)$-NN rule.

6.3.4. Remarks on the first experiment

From the above experiment it is clear that with the proposed method, for any value of k_{max}, $\hat{P}_e$ decreases as m increases and $\hat{P}_r$ increases rapidly with m. A comparison with the k-NN and (k,ℓ)-NN rules is desirable. Figure 6.12 shows that for the data set used, by a suitable choice of m and k_{max}, the proposed scheme can be made to have a smaller $\hat{P}_e$ than is achievable with the k-NN rule with any value of k studied. Usually, this improvement is at the cost of a larger $\hat{P}_r$, although figure 6.13 shows that $\hat{P}_r$ for the proposed method can be kept (by suitable choice of m and k_{max}) to levels that would probably be acceptable (provided m is not too large). Over a small range of values of m and k_{max}, in this experiment, $\hat{P}_r$ as well as $\hat{P}_e$ was actually smaller than with the k-NN rule (e.g. k_{max}=50, m=5).

Thus in applications where a large reject rate is tolerable, the proposed rule may have an advantage over the k-NN rule in allowing use of the error-rejection tradeoff. However since rules of the (k,ℓ)-NN family have a similar advantage, it is desirable to compare the (m,k_{max})-NN rule with these rules. Specifically, comparisons between the (15,ℓ)-NN and (m,15)-NN rules, and between the (m,k_{max})-NN and (k,s)-NN rules, are instructive. Neither rule can be claimed significantly better than the other under all circumstances, and which one is to be preferred in a given application would depend on the application. Assuming similar statistical properties to those of the data set studied here, for example, figures 6.12 and 6.13 show that the

(m,k_{max})-NN rule would allow more flexible use of the error-rejection tradeoff in designing a classifier (see the large vertical spacing of the points for the (15,1)-NN rules in the two figures). In this context it should be noted that although the (m,k_{max})-NN rule will tend to be computationally slightly more costly than the k-NN and (k,1)-NN rules, the increase in cost will be fairly small because of the overhead incurred in calculating and sorting the interpoint distances. No particular effort was made to investigate computation times systematically or to optimise the algorithms used, but it may be of some interest that a typical increase in CPU time was around 5% when compared with the k-NN and (k,1)-NN rules. However, see Luk (1987) for alternative methods of reducing the time for searching the nearest neighbours.

Figure 6.14 is a plot of the estimate $\hat{P}_{et}$ of the total probability $P_{et} = P_r + P_e(1 - P_r)$ when both rejections and misclassifications are regarded as errors. For each value of k_{max} studied, $\hat{P}_{et}$ for the proposed method shows a minimum with respect to m, the value of $\hat{P}_{et}$ at this minimum decreasing with increasing k_{max}. The increase of $\hat{P}_{et}$ from this minimum value as m decreases is due to misclassifications and the increases as m increases is due to rejections. $\hat{P}_{et}$ for the k-NN rule is represented by the solid line in figure 6.14. $\hat{P}_{et}$ for the (k,s)-NN rule is not shown, because for all values of k this rule gave at least as large a $\hat{P}_{et}$ value as did the k-NN rule. Viewed from this standpoint there is little difference in the best performance available from the (m,k_{max})-NN and k-NN rule: however

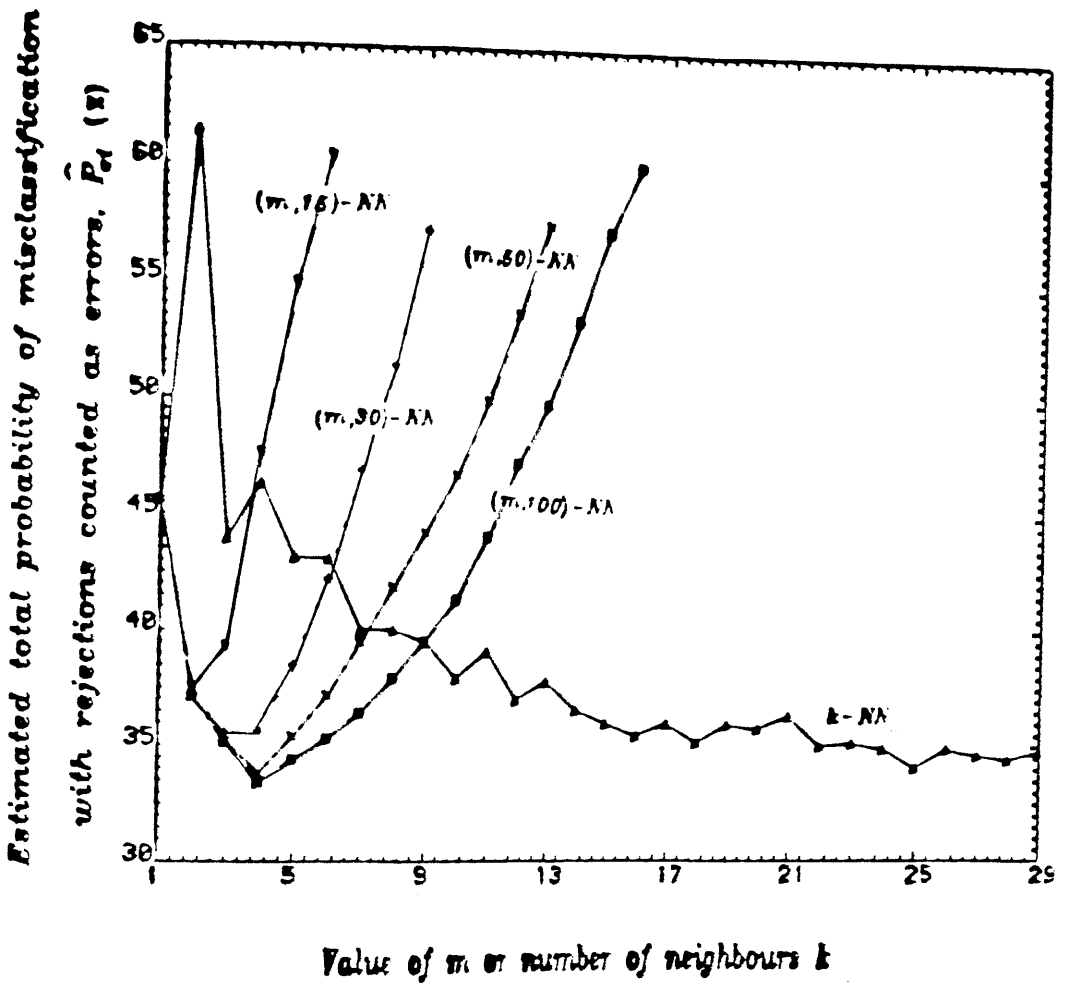


Figure 6.14 (i) Estimated probability of error $\hat{P}_{et}$ (when both rejections and misclassifications are counted as errors) versus required majority m at different values of k_{max}; (ii) $\hat{P}_{et}$ versus k for traditional k-NN rule with ties taken as rejections.

the main advantage of the proposed rule is not in this overall performance but in the use it allows to be made of the error-rejection tradeoff.

A possible method of reducing the often high rejection rate of the (m,k_{max})-NN classification rule would be to invoke the k-NN rule as a second stage to give a final decision $\hat{\omega}(m,k_{max},k)$ whenever the (m,k_{max})-NN rule gives a rejection. Equations (19) for the finite sample set case would then be rewritten as

$$\hat{\omega}(m,k_{max},k) = \begin{cases} \omega_i \quad \text{if} \quad (\ell_i - \ell_j) \geq m, \ \forall\ j \neq i, \\ \qquad\qquad i,j=1,2,\ldots,c, \ m \geq 1, \\ \qquad\qquad \sum_{p=1}^{c} \ell_p \leq k_{max} & \text{(20a)} \\ \hat{\omega}(k) \quad \text{otherwise} & \text{(20b)} \end{cases}$$

where $\hat{\omega}(k)$ denotes the decision made by the k-NN rule (rejections still outstanding could be resolved by other means or left unresolved, according to the application). If our objective is simply to reduce the rejection in equations (19), equation (20b) can be implemented as a k-NN rule with a reject option (i.e. a (k,ℓ)-NN rule with $\ell = k' = \lceil k/c \rceil$, section 5.2.2.2.2). Again, the rejection in the k-NN rule is due to ties between two or more classes. However since the k-NN rule is now being used on samples that are in some intuitive sense difficult to

classify, it is not obvious that the error probability associated with the use of equation (20b) would be better than random. This is especially important at large values of m, where the (m,k_{max})-NN rule has a huge reject rate. A second experiment was therefore performed to investigate the effect(s) of using such a second stage on the probability of error and probability of rejection at different values of m and k_{max} under a <u>fixed</u> value of k (which is not necessary the optimal value for a particular prototype set).

6.3.5. Second Experiment

The first experiment (section 6.3.3) was repeated with k fixed as 15 in equation (20b). Those rejections (all due to ties) still outstanding after equation (20b) was applied were left as rejections (constrast section 6.3.7, below). The probability of error and the probability of rejection were estimated using similar techniques to those mentioned in section 6.3.3 and were defined as

$\hat{P}'_e$ = estimate of probability of error for samples <u>accepted</u> by the (m,k_{max})-NN rule or (failing which) by the 15-NN rule (with a reject option)

$$= \frac{N_{eb}}{N - N'_r}$$

and

$\hat{P}'_r$ = estimate of probability of rejection (i.e. samples rejected by equation 20a and still rejected by the 15-NN rule)

$$= \frac{N'_r}{N}$$

where N'_r = number of samples rejected by equation (20a) and the 15-NN rule (with reject option), and

N_{eb} = number of samples accepted by equation (20a) or the 15-NN rule, but misclassified.

Figure 6.15 and 6.16 show respectively $\hat{P}'_e$ and $\hat{P}'_r$ against m at different values of k_{max}. $\hat{P}_e$ and $\hat{P}_r$ for the traditional k-NN rule (with k varied from 1 to 20) are also shown (solid lines in figure 6.15 and 6.16).

6.3.6 Remarks on the second experiment

It can be seen from the second experiment that for both values of k_{max}, $\hat{P}'_r$ at first increases and $\hat{P}'_e$ decreases with increasing m. Both $\hat{P}'_r$ and $\hat{P}'_e$ gradually level off at larger values of m, depending on the values of k and k_{max} used. It is also necessary to point out that the value of k (= 15) used in equation (20b) is suboptimal for this set of data (figure 6.15). A more careful examination of figure 6.16 will reveal that for this data set $\hat{P}'_r$ is significantly less than $\hat{P}_r$ of the k-NN rule over most of the range of values of m studied. Furthermore, at

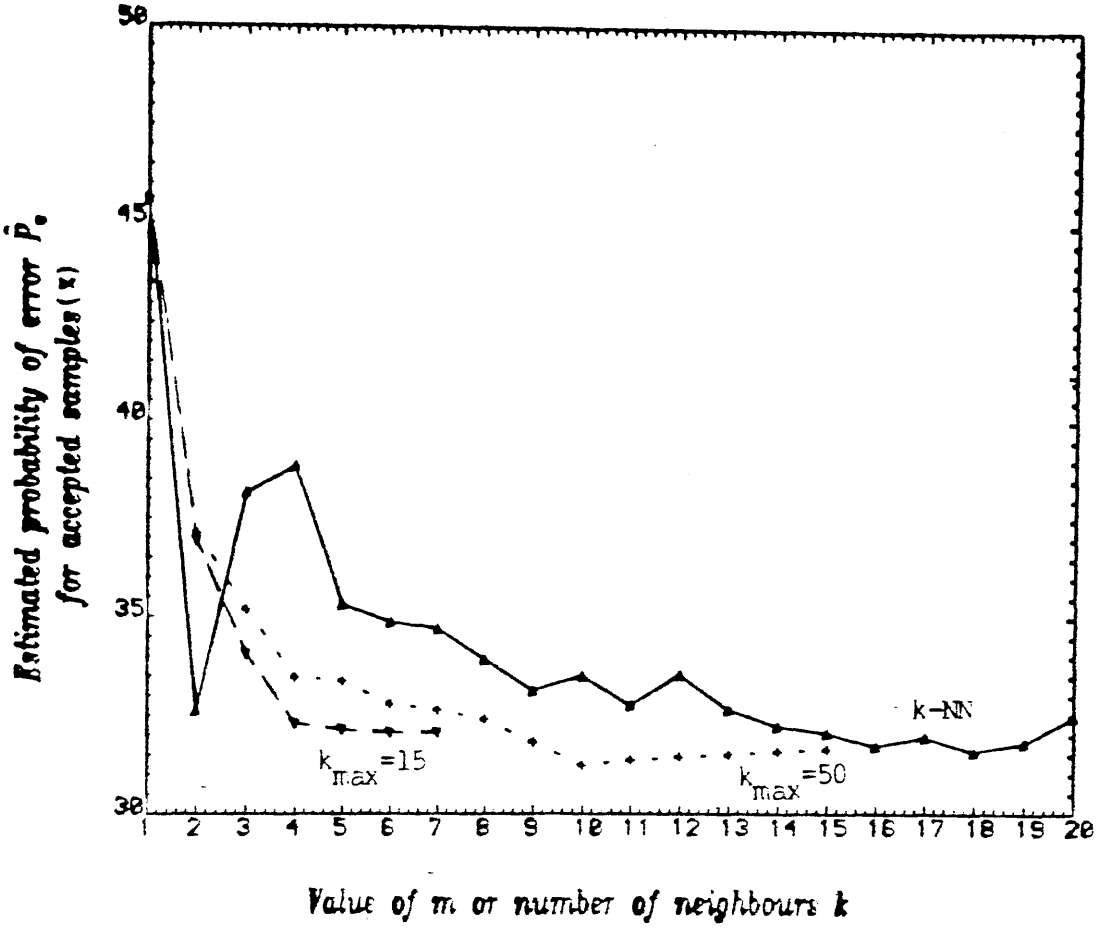


Figure 6.15 (i) Estimated probability of error $\hat{P}'_e$ for accepted samples versus required majority m at different values of k_{max} (k = 15 throughout); (ii) $\hat{P}_e$ versus k for traditional k-NN rule with ties taken as rejections.

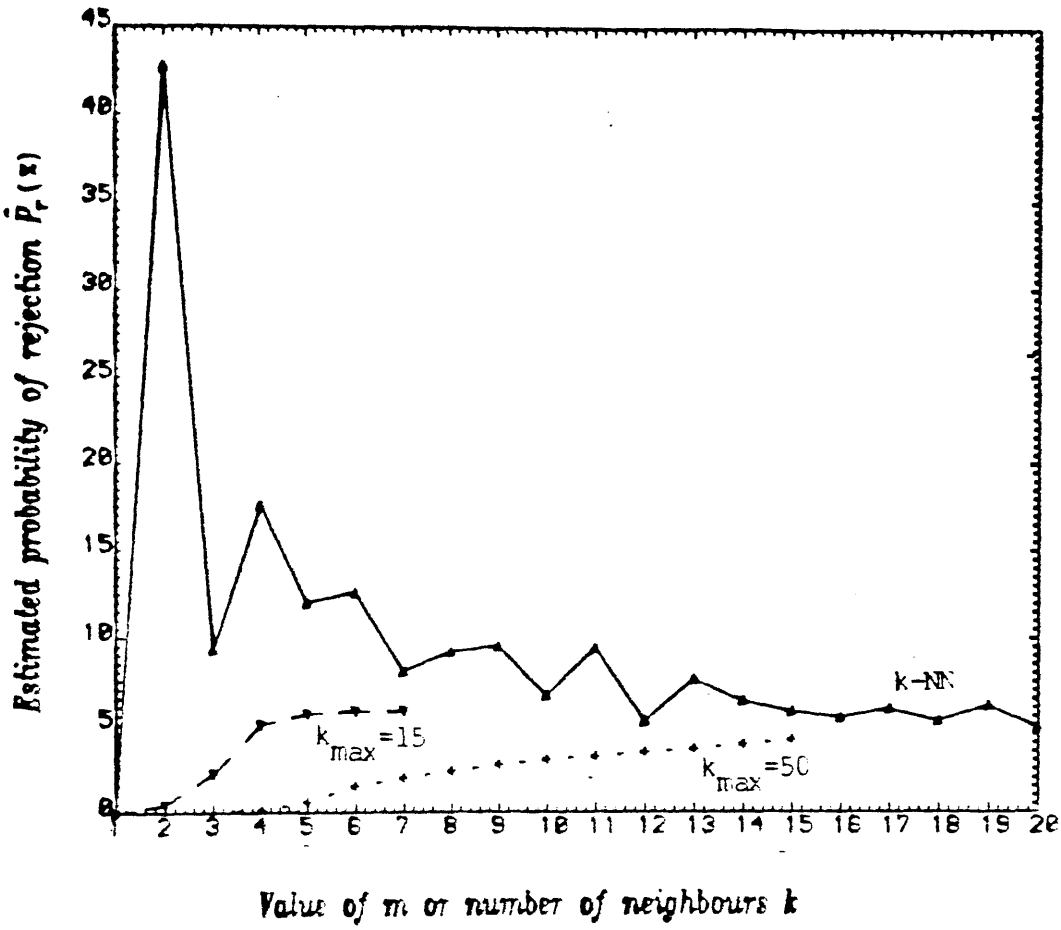


Figure 6.16 (i) Estimated probability of rejection $\hat{P}_r$ versus required majority m at different values of k_{max} (k = 15 throughout); (ii) $\hat{P}_r$ versus k for traditional k-NN rule with ties taken as rejections.

some of these values of m and k_{max}, the proposed method has achieved a smaller value of $\hat{P}_e'$ than the smallest value (which is when k = 18) for the k-NN rule (although the actual difference is less than 0.4% and so it is not statistically very meaningful for the number of test samples used in the experiment). Nevertheless, this suggests that the performance (in term of the probability of error) of the proposed scheme depends heavily on the selection of the values for the parameters m, k and k_{max} in equations (20).

Having illustrated by the above experiment that the combined effect of the (m,k_{max})-NN rule and k-NN rule with a reject option could achieve a slightly smaller $\hat{P}_e'$ when the size of the prototype set S_N is finite, it is also interesting to examine the case when no rejection is permitted. In that case, the second stage of equations (20) could be implemented as a k-NN rule with all the outstanding rejections resolved arbitrarily. Again the probability of error would be of interest because this probability is associated with the errors contributed by the k-NN rule and the arbitration process for the outstanding rejections. Obviously, for this data set, if the error probability with equation (20b) has the value around (c - 1)/c = 2/3 expected when the decision between classes is made randomly (Devijver and Kittler, 1982, page 28), the use of the second stage will be no better than a simple random assignment of class membership for the samples rejected by equation (20a). A third experiment was therefore performed with the aim of studying error probabilities when equations (20a) and (20b) are used with outstanding

rejections resolved arbitrarily.

6.3.7. Third Experiment

The same data set as in section 6.3.3 was classified using equations (20a) and (20b). In equation (20b), k was again taken as 15. In contrast to the second experiment (section 6.3.5) rejections still left outstanding by equation (20b) were resolved arbitrarily by using a random number generator to simulate a c-faced fair die. Since the random number generator was non-repeatable, the test was conducted 10 times and the results were averaged. The results are shown in figures 6.17 and 6.18 in the form of plots against m (for different values of k_{max}) of the mean values $\bar{P}_t$ and $\bar{P}_{15}$ (over 10 tests) of the error probabilities $\hat{P}_t$ and P_{15}, defined as follows:

$\hat{P}_t$ = estimate of probability of error for all samples

$$= \frac{N_{et}}{N}$$

P_{15} = estimate of probability of error associated with the use of the 15-NN classification rule

$$= \frac{N_{et} - N_{ea}}{N_r}$$

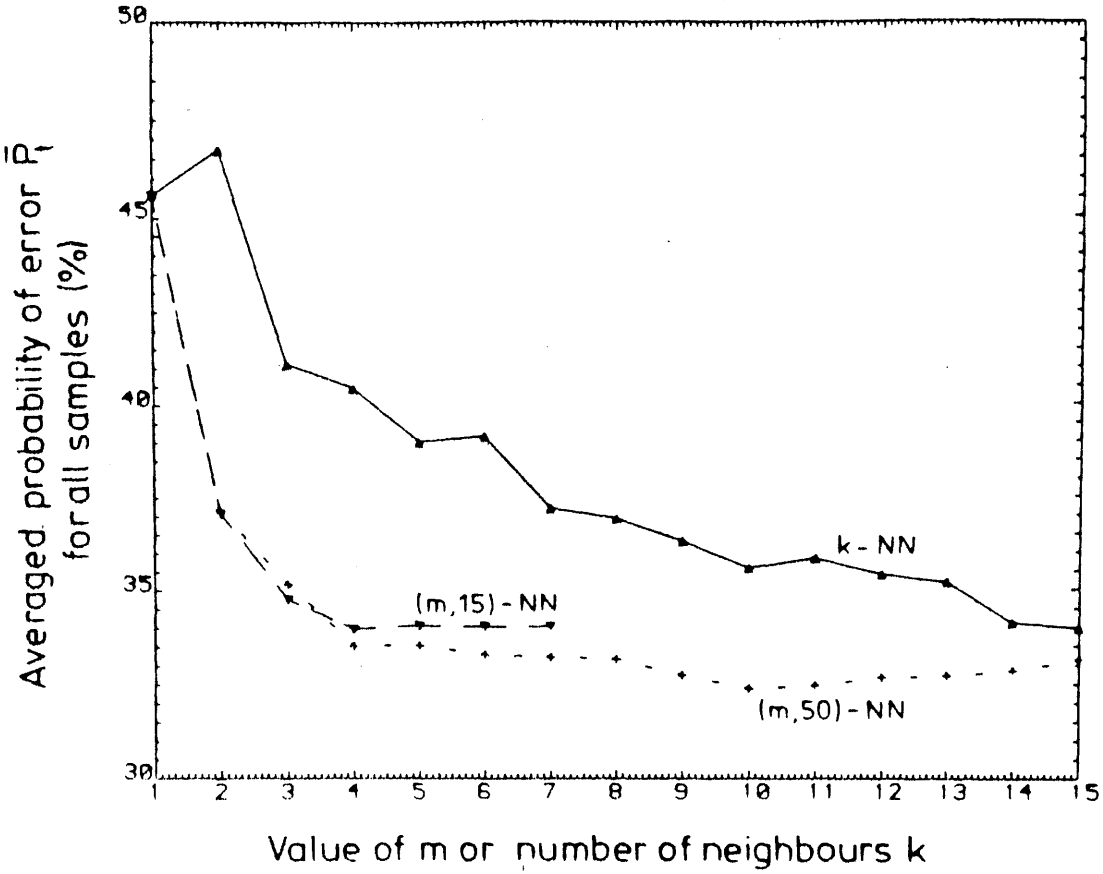


Figure 6.17 (i) Averaged probability of error $\bar{P}_t$ for all samples versus required majority m at different values of k_{max}; (ii) $\bar{P}_t$ versus k for traditional k-NN rule.

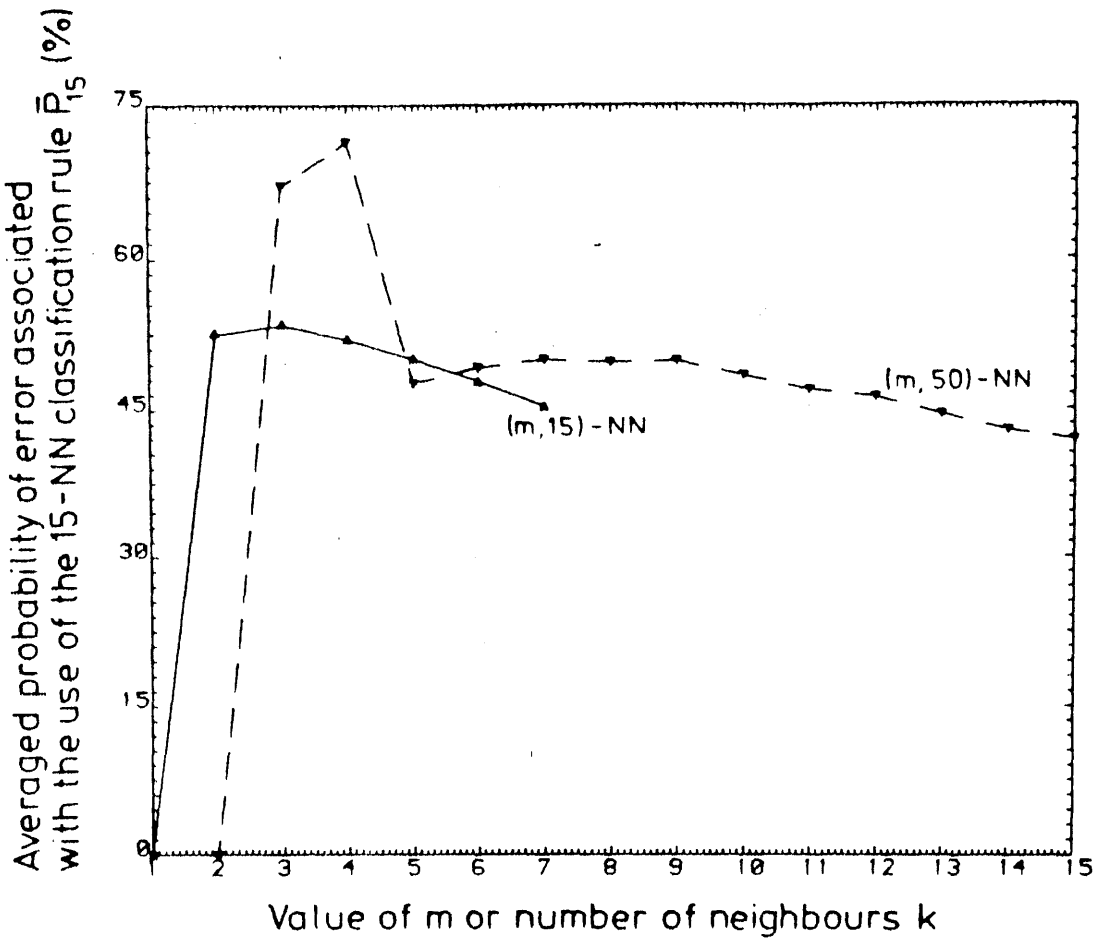


Figure 6.18 Averaged probability of error associated with the use of the 15-NN classification rule $\bar{P}_{15}$ versus required majority m at two values of k_{max}.

$$= \begin{cases} \dfrac{\hat{P}_t - \hat{P}_e(1 - \hat{P}_r)}{\hat{P}_r} & , N_r \neq 0 \\ 0 & , N_r = 0 \end{cases}$$

where now N_{et} = total number of errors

N_{ea} = number of errors for samples accepted by the (m,k_{max})-NN classification rule

N_r = number of samples rejected by the (m,k_{max})-NN classification rule

Note that the definition of N_r is not essentially changed: all rejections in both experiments are due to the (m,k_{max})-NN rule. As before, N is the total number of test samples. In figure 6.17, $\bar{P}_t$ for the traditional k-NN rule with rejections resolved arbitrarily as above is plotted versus k for comparison (solid line).

6.3.8. Remarks on the third experiment

When equations (20a) and (20b) are used with $k_{max} = 50$ and with outstanding rejections resolved by the above arbitrary procedure, figure 6.17 shows that by a suitable choice of the parameter m it is possible with this data set for the (m,k_{max},k)-NN rule to achieve a smaller $\bar{P}_t$ than the k-NN rule (for which the appropriate value for comparison is that with the same value of k as used in equation 20b viz k = 15; note however that this value

is suboptimal). The improvement is small but can be significant: the best improvement in this experiment was at m = 10, where $\bar{P}_t$ was approximately 32.5% as against 34% with the 15-NN rule (the 95% confidence limits in figure 6.17 are at $\pm$ 0.5%). This conclusion applies only if enough samples are permitted to be searched: using k_{max} = 15 yields no improvement on the traditional k-NN rule (figure 6.17).

When m is large and the rejection by equation (20a) is consequently huge (figure 6.13), the value of $\bar{P}_t$ is largely controlled by k. Hence a minimum in a plot of total error rate versus m is to be expected: such a minimum is observed in figure 6.17 at m = 10 for k_{max} = 50.

In figure 6.18, for k_{max} = 50, $\bar{P}_{15}$ at the larger values of m decreases to about 42% as compared with the value 2/3 expected. Thus the use of the 15-NN rule on this data set to classify samples rejected by the (m,k_{max})-NN rule has given an error rate which is better than random on these difficult-to-classify samples, in spite of the arbitrary resolution of rejections still left outstanding by the 15-NN rule.

The shape of the plot for k_{max} = 50 in figure 6.18 for smaller values of m can be interpreted as follows: (a) for m = 1 and m = 2 there are no rejections by equation (20a) and P_{15} is by definition zero; (b) for intermediate values there are relatively few rejections by equation (20a) (e.g. for m=3 only 3 of the 3000 test samples were rejected - see figure 6.18) and P_{15} is no

longer a statistically valid estimate.

6.3.9. Discussion

The behaviour of the proposed nearest neighbour classification rule, the (m,k_{max})-NN rule, has been investigated in three experiments on a Gaussian data set comprising 150 prototype samples and 3000 test samples. For this data set, the first experiment has shown that the proposed scheme can give a lower error probability than the k-NN classification rule but tends to give a higher rejection probability. The rejection probability can become very high in some circumstances but can be kept down by suitable choice of the parameters m and k_{max} to levels that are probably acceptable. The first experiment has also indicated that the proposed method can be used to complement the (k,ℓ)-NN rule because of the difference in the rejection probability. It is suggested that the proposed rule, in spite of its probably larger computational cost, could be useful in applications where a low error probability is essential and a somewhat large rejection rate is tolerable, and where the (k,ℓ)-NN rule cannot achieve the necessary tolerable rejection rate.

The second and third experiment have shown that lower finite probability of error can in some cases be achieved when the k-NN rule is used to classify the samples rejected by the (m,k_{max})-NN rule than when the k-NN rule is used alone. The reduction may not be significant, but it is accompanied by a significant reduction in rejection probability (where

applicable).

Several other questions have been left open in this section. Firstly, what are the asymptotic performance and the rate of convergence of equation (18) compared with the k-NN and (k,1)-NN classification rules? Secondly, what is the actual relationship between the misclassification rate of the proposed scheme and the Bayes risk (or error)? Thirdly, how is one to determine an optimal choice of the parameters m and k_{max}? These difficult questions will be left for future research.

Appendix 6.A: Expression used to evaluate P_{A2}

Following the arguments given in section 6.2.3, P_{A2} is given by the following expression.

$$
\begin{aligned}
&2\int_{-1}^{-1/2}\int_{0}^{1+x}\int_{0}^{1+x-y} f_1(x)\,[f_1(x+y)+f_1(x-y)]\,[f_1(x+y+z)+f_1(x-y-z)]\,dz\,dy\,dx \\
&+\ 2\int_{-1}^{-1/2}\int_{0}^{1+x}\int_{1+x-y}^{-x-y} f_1(x)\,[f_1(x+y)+f_1(x-y)]\,f_1(x+y+z)\,dz\,dy\,dx \\
&+\ 2\int_{-1}^{-1/2}\int_{0}^{1+x}\int_{-x-y}^{1-x-y} f_1(x)\,[f_1(x+y)+f_1(x-y)]\,f_1(x+y+z) \\
&\qquad [1-F_2(x+y+z)]^2\,dz\,dy\,dx \\
&+\ 2\int_{-1}^{-1/2}\int_{1+x}^{-x}\int_{0}^{-x-y} f_1(x)\,f_1(x+y)\,f_1(x+y+z)\,dz\,dy\,dx \\
&+\ 2\int_{-1}^{-1/2}\int_{1+x}^{-x}\int_{-x-y}^{1-x-y} f_1(x)\,f_1(x+y)\,f_1(x+y+z)\,[1-F_2(x+y+z)]^2\,dz\,dy\,dx \\
&+\ 2\int_{-1}^{-1/2}\int_{-x}^{1-x}\int_{0}^{1-x-y} f_1(x)\,f_1(x+y)\,f_1(x+y+z)\,[1-F_2(x+y+z)]^2\,dz\,dy\,dx \\
&+\ 2\int_{-1/2}^{0}\int_{0}^{-x}\int_{0}^{-x-y} f_1(x)\,[f_1(x+y)+f_1(x-y)]\,[f_1(x+y+z)+f_1(x-y-z)]\,dz\,dy\,dx \\
&+\ 2\int_{-1/2}^{0}\int_{0}^{-x}\int_{-x-y}^{1+x-y} f_1(x)\,[f_1(x+y)+f_1(x-y)]\,[f_1(x+y+z)+f_1(x-y-z)] \\
&\qquad [1-F_2(x+y+z)]^2\,dz\,dy\,dx \\
&+\ 2\int_{-1/2}^{0}\int_{0}^{-x}\int_{1+x-y}^{1-x-y} f_1(x)\,[f_1(x+y)+f_1(x-y)]\,f_1(x+y+z) \\
&\qquad [1-F_2(x+y+z)]^2\,dz\,dy\,dx \\
&+\ 2\int_{-1/2}^{0}\int_{-x}^{1+x}\int_{0}^{1+x-y} f_1(x)\,[f_1(x+y)+f_1(x-y)]\,[f_1(x+y+z)+f_1(x-y-z)] \\
&\qquad [1-F_2(x+y+z)]^2\,dz\,dy\,dx \\
&+\ 2\int_{-1/2}^{0}\int_{-x}^{1+x}\int_{1+x-y}^{1-x-y} f_1(x)\,[f_1(x+y)+f_1(x-y)]\,f_1(x+y+z) \\
&\qquad [1-F_2(x+y+z)]^2\,dz\,dy\,dx \\
&+\ 2\int_{-1/2}^{0}\int_{1+x}^{1-x}\int_{0}^{1-x-y} f_1(x)\,f_1(x+y)\,f_1(x+y+z)\,[1-F_2(x+y+z)]^2\,dz\,dy\,dx \\
&+\ 2\int_{0}^{1/2}\int_{0}^{x}\int_{0}^{x-y} f_1(x)\,[f_1(x+y)+f_1(x-y)]\,[f_1(x+y+z)+f_1(x-y-z)] \\
&\qquad [1-F_2(x+y+z)+F_2(x-y-z)]^2\,dz\,dy\,dx \\
&+\ 2\int_{0}^{1/2}\int_{0}^{x}\int_{x-y}^{1-x-y} f_1(x)\,[f_1(x+y)+f_1(x-y)]\,[f_1(x+y+z)+f_1(x-y-z)] \\
&\qquad [1-F_2(x+y+z)]^2\,dz\,dy\,dx \\
&+\ 2\int_{0}^{1/2}\int_{0}^{x}\int_{1-x-y}^{1+x-y} f_1(x)\,[f_1(x+y)+f_1(x-y)]\,f_1(x-y-z) \\
&\qquad [1-F_2(x+y+z)]^2\,dz\,dy\,dx
\end{aligned}
$$

$$+ 2\int_0^{1/2}\int_x^{1-x}\int_0^{1-x-y} f_1(x)[f_1(x+y)+f_1(x-y)][f_1(x+y+z)+f_1(x-y-z)]$$
$$[1-F_2(x+y+z)]^2 dz dy dx$$
$$+ 2\int_0^{1/2}\int_x^{1-x}\int_{1-x-y}^{1+x-y} f_1(x)[f_1(x+y)+f_1(x-y)]f_1(x-y-z)$$
$$[1-F_2(x+y+z)]^2 dz dy dx$$
$$+ 2\int_0^{1/2}\int_{1-x}^{1+x}\int_0^{1+x-y} f_1(x)f_1(x-y)f_1(x-y-z)[1-F_2(x+y+z)]^2 dz dy dx$$
$$+ 2\int_{1/2}^{1}\int_0^{1-x}\int_0^{1-x-y} f_1(x)[f_1(x+y)+f_1(x-y)][f_1(x+y+z)+f_1(x-y-z)]$$
$$[1-F_2(x+y+z)+F_2(x-y-z)]^2 dz dy dx$$
$$+ 2\int_{1/2}^{1}\int_0^{1-x}\int_{1-x-y}^{x-y} f_1(x)[f_1(x+y)+f_1(x-y)]f_1(x-y-z)$$
$$[1-F_2(x+y+z)+F_2(x-y-z)]^2 dz dy dx$$
$$+ 2\int_{1/2}^{1}\int_0^{1-x}\int_{x-y}^{1+x-y} f_1(x)[f_1(x+y)+f_1(x-y)]f_1(x-y-z)$$
$$[1-F_2(x+y+z)]^2 dz dy dx$$
$$+ 2\int_{1/2}^{1}\int_{1-x}^{x}\int_0^{x-y} f_1(x)f_1(x-y)f_1(x-y-z)[1-F_2(x+y+z)+F_2(x-y-z)]^2 dz dy dx$$
$$+ 2\int_{1/2}^{1}\int_{1-x}^{x}\int_{x-y}^{1+x-y} f_1(x)f_1(x-y)f_1(x-y-z)[1-F_2(x+y+z)]^2 dz dy dx$$
$$+ 2\int_{1/2}^{1}\int_{x}^{1+x}\int_0^{1+x-y} f_1(x)f_1(x-y)f_1(x-y-z)[1-F_2(x+y+z)]^2 dz dy dx$$

References

[01] Bailey, T. and Jain, A.K. (1978). "A note on distance-weighted k-nearest neighbour rules". **IEEE Trans. on Systems, Man, and Cybernetics**, SMC-8, 311-313.

[02] Brown, T.A. and Koplowitz, J. (1979). "The weighted nearest neighbour rule for class dependent sample size". **IEEE Trans. on Information Theory**, IT-25, 617-619.

[03] Cover, T.M. and Hart, P.E. (1967). "Nearest neighbour pattern classification". **IEEE Trans. on Information Theory**, IT-13, 21-27.

[04] Devijver, P.A. and Kittler, J. (1982). **Pattern Recognition: A Statistical Approach**. Prentice-Hall International, London.

[05] Devroye, L.P. (1981). "On the inequality of Cover and Hart in nearest neighbour discrimination". **IEEE Trans. on Pattern Analysis and Machine Intelligence**, PAMI-3, 75-78.

[06] Dudani, S.A. (1976). "The distance-weighted k-nearest neighbour rule". **IEEE Trans. on Systems, Man, and Cybernetics**, SMC-6, 325-327.

[07] Fukunaga, K. and Flick, T.E. (1985). "The 2-NN rule for more accurate NN risk estimation". **IEEE Trans. on Pattern Analysis and Machine Intelligence**, PAMI-7, 107-112.

[08] Keller, J.M., Gray, M.R. and Givens, J.A. Jr. (1985). "A fuzzy k-nearest neighbour algorithm". **IEEE Trans. on Systems, Man, and Cybernetics**, SMC-15, 580-585.

[09] Luk, A. and Macleod, J.E.S. (1986). "An alternative nearest neighbour classification scheme". **Pattern Recognition Letters**, 4, 375-381.

[10] Luk, A. (1987). "Nearest neighbour classification". Internal Report, University of Glasgow.

[11] Macleod, J.E.S., Luk, A. and Titterington, D.M. (1987). "A re-examination of the distance weighted k-NN classification rule". **IEEE Trans. on Systems, Man, and Cybernetics**, accepted for publication.

[12] Morin, R.L. and Raeside, D.E. (1981). "A reappraisal of distance-weighted k-nearest neighbour classification for pattern recognition with missing data". **IEEE Trans. on Systems, Man, and Cybernetics**, **SMC-11**, 241-243.

Lung sound Analysis: a possible non-invasive system

This work was supported by the Croucher Foundation, Hong Kong.

Chapter 7: Lung sound analysis

(a possible non-invasive examination system)

Summary

This chapter describes in detail each of the implementable modules developed at this stage, and brings together all results relevant to the proposed non-invasive examination system for patients exposed to asbestos dusts. Problems associated with the implementation of the respective modules are discussed.

7.1 Introduction

It has been mentioned in section 2.5.2 that one of the most important abnormal physical signs of asbestosis in a patient is the existence of persistent, bilateral, basal late-inspiratory fine crackles. However, auscultation using a binaural stethoscope is a rather subjective exercise. In fact, it has been shown (Ertel, et al., 1966b) that each type of stethoscope has its own frequency response: some types will attentuate the higher frequencies while others will attentuate the lower frequencies. Worse still Ertel, et al. (1966a) have demonstrated that age and some ear disorders, like presbycusis, will affect hearing through a stethoscope. In addition there is always some variations in response for different physicians. It is not surprising that auscultation (with a stethoscope) was slowly replaced in importance as advances were made in other diagnostic

tools such as radiography and more recently nuclear magnetic resonance imaging.

However, all this does not necessarily imply that lung sound itself is not useful. It merely indicates that a suitable transducer should be used to acquire the lung sound signals and that objective techniques are required to analyse and intepret the acquired signals. As pointed out in section 2.5.2, for a non-invasive system for routine examination of patients exposed to asbestos dusts (possibly deployed by para-medical personnel), lung sound seems to be a useful candidate.

The overall structure for the proposed non-invasive system has been introduced in section 1.3 (see also figure 1.1). Each individual module, with the exception of the unsupervised learning module, will be described in more detail in this chapter. As noted earlier in chapter 1, one of aims of the unsupervised learning module is to detect any new information in a given environment when the classification module give rises to excessive errors; and as such, it has been deliberately left out at this stage of the development because there is an insuffient number of asbestosis patients. In chapter 8 a brief discussion of this undeveloped module will be given in relation to possible future research work.

The layout of this chapter is similar to the organisation of the system itself. The theory relevent to this part of the thesis have been introduced in chapters 2, 3, 4 and

5, and hence only experimental results are presented in this chapter. A very brief summary of the conditions of the subjects used in this study is given in section 7.2. The data acquisition module and the prepocessing and features generating module are respectively discussed in sections 7.3 and 7.4. Experimental results from the mapping module are given in section 7.5. In section 7.6, the classification module will be described. This chapter ends with a few concluding remarks in section 7.7.

7.2 Conditions of the subjects

Three groups (or classes) of male subjects are used in this study: 5 patients with asbestosis; 5 patients who are known to have been exposed to asbestos fibres but who have not (yet) developed any known asbestos-related ailment (referred to as exposed subjects) ; and 5 healthy non-smoking persons (normal subjects). All asbestosis patients have established abnormal physical signs and abnormalities in both their chest x-ray film and their pulmonary functional tests (section 2.5.2). Also worth noting is that during the course of this project one of the exposed patients died from a cardiac disorder.

7.3 Data Acquisition: the equipment and the procedure

Sounds heard at a chest wall have been transmitted through a number of media: the lung parenchyma, the pleural cavity and the thoracic cavity (section 2.2). Ideally, the acoustic impedance of a transducer should match that of the chest

so that the transmitted sounds will be a maximum (Hueter and Bolt, 1955). Unfortunately, the impedance is not a constant at different locations on the chest nor is it the same for different individuals. Further, the elastic properties of the lung parenchyma and other connective tissues may not necessarily follow the simple linear Hooke's law (which states that restoring force is proportional to displacement) for a vibrating system. For instance, the thermodynamic properties of a gas may change considerably when the pressure becomes very high just before the opening of an obstructed or restricted small airway. The problem of turbulence and vorticity at the larger airways may add additional complications to the computation of the impedance. Consequently, there is still no known satisfactory model or method to determine the impedance of the chest wall.

Howie (1981) and Tierney (1983) have done some preliminary experiments to evaluate the acoustic impedance of the chest. Their experiment centred on the assumption that table-jelly has similar acoustic properties to human lung. However, their studies have failed to take into account the impedance of the thoracic cavity and the surrounding connective tissues. Another more accurate estimation method is to determine the impedance of the chest wall of a cadaver, but such an experiment would raise ethical and legal problems as well as technical ones.

For these reasons and also for compatiblilty with the previous recordings, a transducer which was designed by McGhee (1978), based on the work of Guard (1976), and was found to

produce "reasonably usable output" (Urquhart, 1983) was used for this study. It is an enclosure system with a General Radio 0.5-inch electret microphone (type 1962-9602) and a matched preamplifier surrounded by a thick tube of aluminium. This tube acts as a mechanical mass element which attentuates ambient sound (acoustic interference and noise) reaching the microphone. The head of this enclosure is fitted with a rigid Tufnol diaphragm 0.64mm thick to crudely match the impedance between the microphone and the chest wall. This improves the ability of the microphone to detect sound from the chest and, at the same time, reduces the interference recorded (Urquhart, 1983).

The microphone and the preamplifier have a -2dB bandwidth from 5Hz to 20kHz (manufacturer's data) although this is modified considerably by the aluminium tubing. Figure 7.1 shows the frequency response of the GR microphone and enclosure in air (solid curve). The response of a standard Bruel and Kjaer microphone (type 4134) is also shown for comparison (dotted curve). Both microphones were tested under the same conditions, and both were calibrated against another standard Bruel and Kjaer microphone (type 4165). It can be seen that this enclosure system has some rather irregular responses: (a) decreasing attenuation from 20Hz to 2000Hz; (b) amplification within the range of 3kHz to 6kHz; and (c) sharp attenuation at higher frequencies.

The lung sound signals picked up by this enclosed microphone system were recorded by a variable speed 4-channel

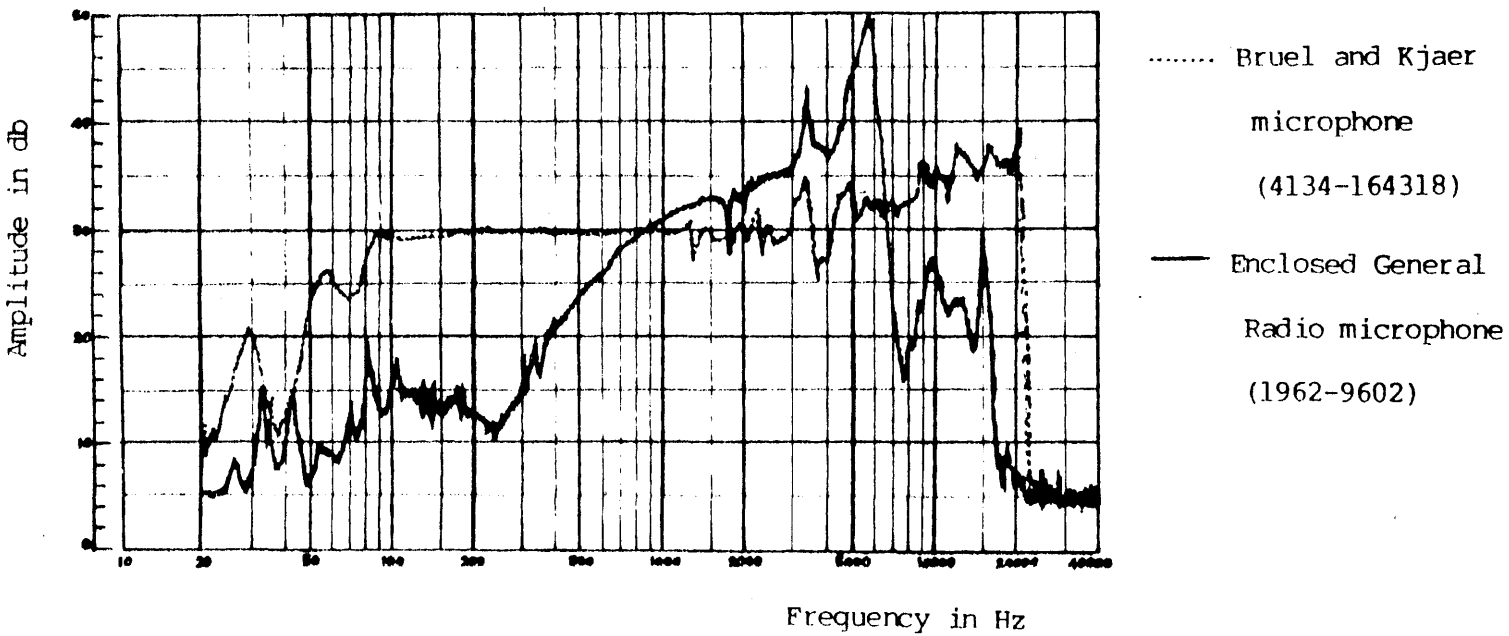


Figure 7.1 Frequency response in air for the enclosed General Radio microphone system and a Bruel and Kjaer microphone.

recorder (Racal Store-4D) with the speed set at 30 inches per second. Again the choice of the recording speed is mainly for compatibility with the previous recordings done by Urquhart (1983) so that the recordings may be reusable in this study. At this speed it permits all frequencies between 0 to 10kHz to be recorded (manufacturer's data).

All recording sessions were performed in one of the consultation rooms of the Glasgow Royal Infirmary under the supervision of an experience physician who has specialized in respiratory disorders. Each subject was asked to remove all his upper garments, to sit on a stool and to breath slighly harder than normal through his mouth. (Again, this was largely determined by the physician.) Auscultation with a stethoscope was then performed at the lower (or basal) posterior part of the chest near the ninth or tenth intercostal space prior to the actual recording. After the physician had selected the site of recording, the enclosed microphone system was held firmly with both hands and gently pressed against that location. Throughout the recording, the sound was continuously monitored by a Gould 15MHz oscilloscope to make sure that the recorded level was acceptable. Around 10 breath cycles were recorded. The whole recording was replayed (and displayed on the same oscilloscope) at a slightly slower speed and if there was no apparent serious flaw (such as sudden movement of the hand which can cause a lengthy and large deflection in the recorded signal), the recording was accepted for further analysis; else the whole procedure was repeated until the recording was acceptable.

The transitions between inspiration and expiration were recorded on a separate channel in the Racal Store-4D recorder. At the beginning of this study, the transition was identified using a thermistor mounted on a probe that was attached to a headset worn by the subject. However because of difficulties with the probe and for hygienic reasons, the breath transition was finally identified manually by using a simple push button, with inspiration indicated by pressing the button and expiration by releasing it.

After a recording session, the tape was transferred back to the University of Glasgow for further analysis. It was discovered that in order to digitize the lung sound, the recorder speed has to be reduced by a factor of 16 i.e. to 1.875 inches per second in order to overcome software overheads in the data logging programs (Campbell, 1983). The recorded signals were then digitized (or sampled) by a 12-bit Micro Consultants analog-to-digital converter with anti-aliasing filters under the control of a small 8-bit Z80-based micro-computer (model SBC-100). Since only inspirations are of interest in this study (section 7.1), the whole digitization process was manually controlled with the help of the breath transition signals and an oscilloscope. A schematic diagram of the data acquisition module is shown in figure 7.2. The digitized inspiration signals were permanently stored on floppy diskettes and transferred to the GEC 4180 minicomputer for further analysis by using a dedicated file-transferring SBC-200 microcomputer. Unfortunately, the GEC 4180 would not accept binary data without modifying it. Therefore,

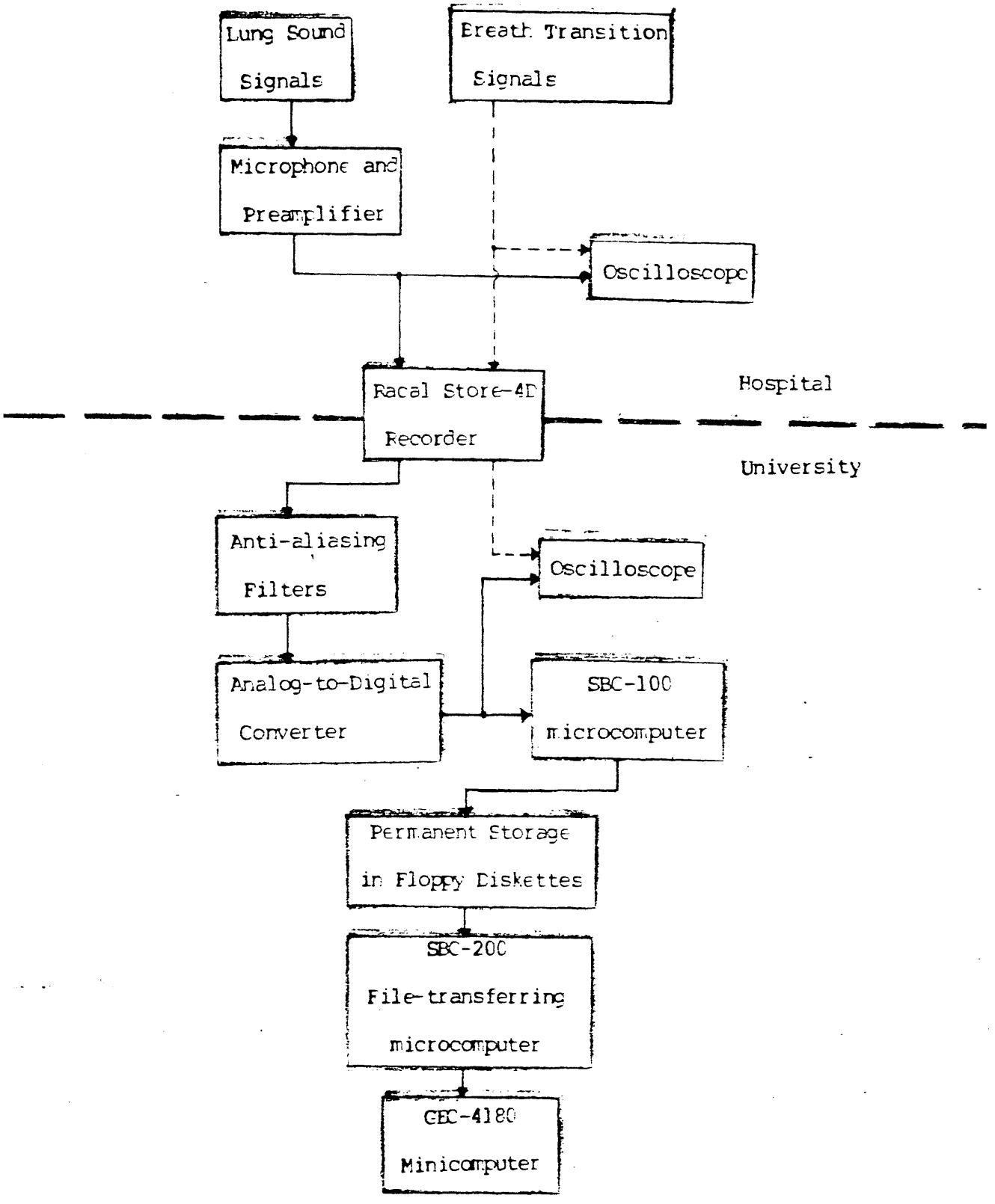


Figure 7.2 Schematic diagram of the data acquisition module.

the whole transferring procedure has to include a binary-to-hexadecimal conversion at the SBC-200 end and a hexadecimal-to-binary conversion (for storage efficiency) when the data finally arrived at the GEC 4180.

The whole process was extremely time consuming. The instrumentation tape recorder has to be transported back and forth by van between the Glasgow Royal Infirmary (for recording) and the University (for playback and data logging). The data had to be logged on the SBC-100 micro-computer under manual supervision as described above, the floppy disc moved to the SBC 200 for data transfer, the file translated from binary to hexadecimal, the data transferred down-line to the GEC 4180 (with further delays in the event of a fault on the SBC-200, the GEC 4180, or the line), and finally the data translated back to binary. Moreover in view of the large size of the data files, only about three could be held on GEC 4180 disc storage simultaneously, and therefore a new experiment involving the whole data set necessitated repetition of all the above steps from the data logging onwards. These cumbersome procedures retarded progress on the project and had the effect that of the 50 subjects from whom recordings were obtained, the data from only 15 could be analysed. The author understands that a more efficient data acquisition system is now under development.

7.4 Preprocessing and Feature Generation

7.4.1 Preprocessing: spectrum estimation

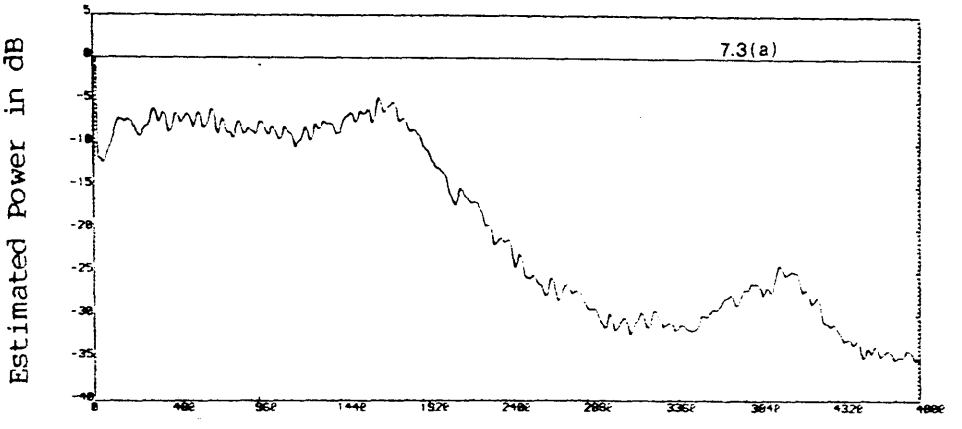
The length of the digitized inspiration signals varies between 10k and 28k (of measurements) because the duration of each inspiration varies (a) within an individual and (b) between different subjects. Thus, it would be useful if these variable length data records could somehow be transformed into records of fixed length. To do so, it may be necessary to transform these data from one domain to another domain. The inherent cyclic nature of respiration makes frequency domain analysis very attractive. In this study, a datum with a variable number of measurements in the time domain is transfomed into a datum with fixed number of measurements in the frequency domain by weighted overlapped segment averaging (WOSA) - section 3.3.3. More simple estimation techniques which are based directly on the fast Fourier transform (FFT) have been used by a number of researchers (Murphy and Sorensen, 1973; Mori et al., 1978; Gavriely et al., 1981; Urquhart et al., 1981).

It is worth pointing out here that another spectrum estimation technique, the maximum entropy method (MEM), was briefly investigated by Urquhart (1983). However, his preliminary study indicated that it is very difficult to select the right model order which is essential in MEM (Akaike, 1974; Kay and Marple, 1981). Considering the variablity of each breath cycle and the difficulty in finding the right model order, it is

not surprising that relatively very few publications in lung sound analysis are based on MEM or related techniques.

As pointed out in section 3.3.3 there are two methods to implement WOSA, namely the algorithm due to Carter and Nuttall (1980) and that due to Yuen (1983). Both of these algorithms are implemented in this study. In both methods, a 2048-point FFT is used for each segment and the quadratic (or lag) window is a 5-point Daniell window (Yuen and Fraser, 1979). In the algorithm proposed by Carter and Nuttall, each segment is 2k in size, with a 50% overlapping between each segment. The degree of overlapping is dictated by the window used with each segment (Nuttall, 1981), and in this study is a minimum 4-sample Blackman-Harris window. The averaged periodogram generated by either method will contain 1024 points and have a resolution of about 4.69Hz. Figure 7.3 shows the averaged periodogram produced by each algorithm for one of the asbestosis patients. It can be seen that there is not much difference between the two approaches. This may be attributed to the number of the measurements taken for each inspiration. In the study by Carter and Nuttall (1980), the slight difference between the two approaches is given for data of comparatively short size and large dynamic range. For this reason, the approach by Carter and Nuttall was adopted for subsequent transformation.

Figure 7.4 illustrates a periodogram from one of the exposed patients and another one from one of the normal subjects. It can be seen that the asbestosis patient has a periodogram



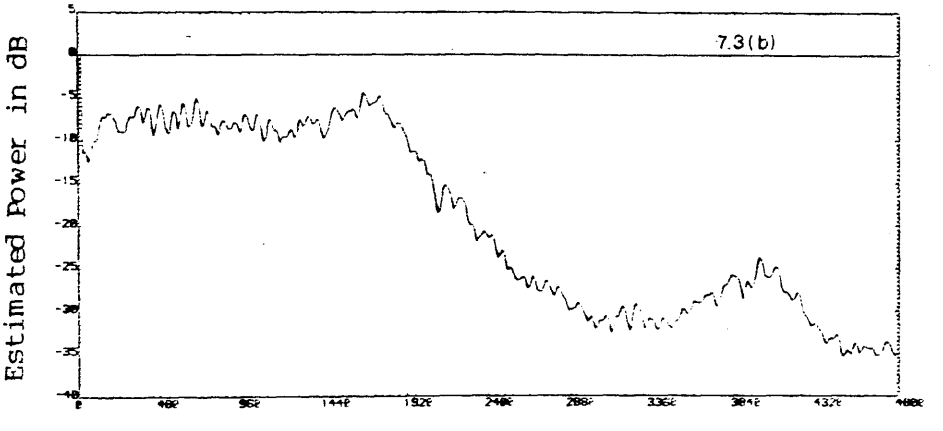
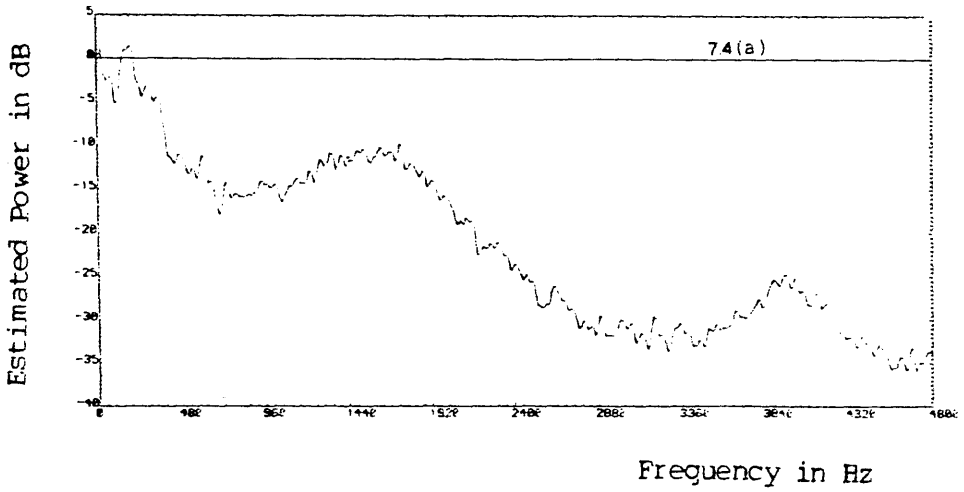


Figure 7.3 Averaged Periodograms for an asbestosis patient using (a) the algorithm due to Carter and Nuttall and (b) Yuen's algorithm.



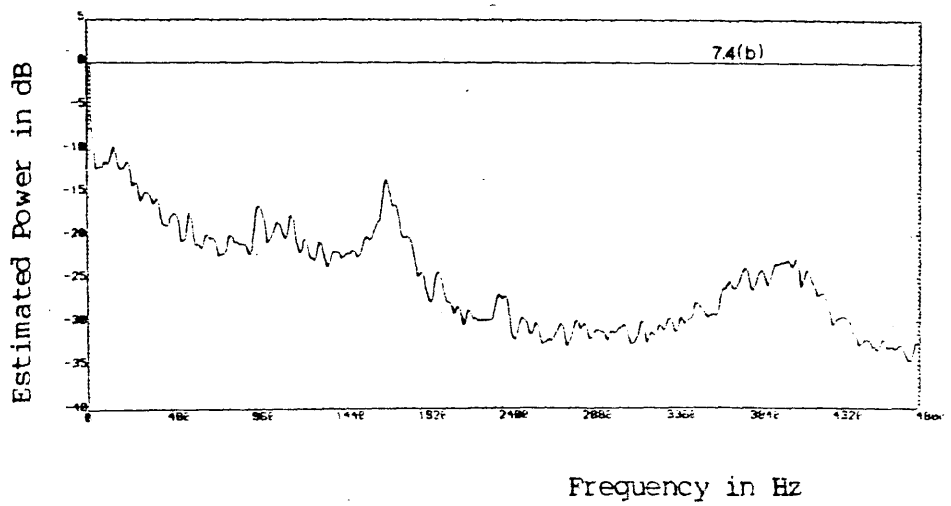


Figure 7.4 Averaged Periodogram formed by the algorithm due to Carter and Nuttall for (a) an exposed patient and (b) a normal subject.

between 20Hz to 2000Hz than either the exposed or the normal subject. Also worth notice is the small peak at around 3800Hz for each of the three classes of subjects studied. This may be due to the irregular response of the microphone rather than related to the lung sound.

7.4.2 Feature generation

One of the most difficult parts of the design of a pattern recognition system is the construction of the set of features. Unlike the identification of the set of pattern classes which is usually easy (for example in character recognition, the set of characters is the set of classes), the choice of features depends very much on the intuition of the designer. (Of course, once the set of features has been decided on, there are many methods, for either selecting or extracting the more useful ones.) In this study, once the inspiratory signals have been digitized and transformed into the frequency domain, the set of averaged periodograms can be considered as a set of patterns. Following Urquhart (1983), the averaged periodogram is divided into a number of frequency intervals, and then a feature is generated from each frequency interval. In this way the dimensionality is drastically reduced from 1024 to a much smaller value.

Also in their previous study, Urquhart et al. (1981) have concentrated their analysis on frequencies between 2Hz and 400Hz. In particular they have indicated that frequencies below

50Hz were of medical significance. Therefore, 6 frequency intervals, and hence 6 features, were constructed between 2Hz and 50Hz. Kraman (1983) suggested that sounds in that low frequncy range probably originated from the chest muscle rather than from the lung itself. Nevertheless, using only these 6 features, Urquhart (1983) was able to distinguish beteen normal subjects and a number of respiratory disorders, such as pulmonary oedema and asbestosis. Thus whether or not Kraman's suggestion is correct, Urquhart's results have established that lung sounds in the low frequency range contain information that is useful in distinguishing between lung conditions, whatever the mechanism by which the sounds are generated.

In this study, the frequency between 4Hz and 2000Hz is analysed because it has been shown (Benedetto et al., 1983) that crackles in fibrosing diseases, like asbestosis and cryptogenic (or idiopathic) fibrosing alveolitis, usually have some significant high frequency components. Therefore, the range between 0 and 2000Hz is divided equally into 20 intervals, each with a width of 200Hz. Hence a 20 dimensional feature vector can be generated from each inspiration. For each frequency interval, a feature is formed by averaging the estimated power (or the components of the averaged periodogram) within that interval. Furthermore, as Sayers (1975) has suggested, normalization is useful for biological signals because it is usually the shape of the spectrum rather than its actual magnitude that is significant. Unfortunately, the most obvious normalization procedure, that is dividing the average estimated power in each

interval by the total power within 0 to 2000Hz, did not produce the best result when some of the data were projected onto a two dimensional space using the Kittler and Young transformation (figure 7.5). After some trial and error experimentation, it was found that, at least for this particular data set, the best projection under the same transformation could be obtained by dividing the log of the average estimated power in each frequency interval by the log of the largest average estimated power (figure 7.6). (This has the effect of raising each estimated power to an index which is proportional to the inverse of the log of the largest averaged estimated power.) Thus, this normalization method is used in subsequent analysis.

7.5 Mapping

The above procedures obtain a 20-dimensional feature vector for each inspiration. The purpose of the next stage of the non-invasive system is to provide a user some visual aids. One method of doing this is to project the 20 dimensional vectors onto a 2 dimensional space through the application of one of the linear mapping algorithms discussed in chapter 4. The idea is to determine a mapping (or transformation) matrix U which optimizes a certain criterion function and, at the same time, has a small variance in the estimated scatter matrix (section 4.2.6). Assuming that somehow the mapping matrix U has been determined with sufficiently small variance, the user can then visually interpret the projected data. For example, it may be possible, with the help of the visual aid, to answer such questions as

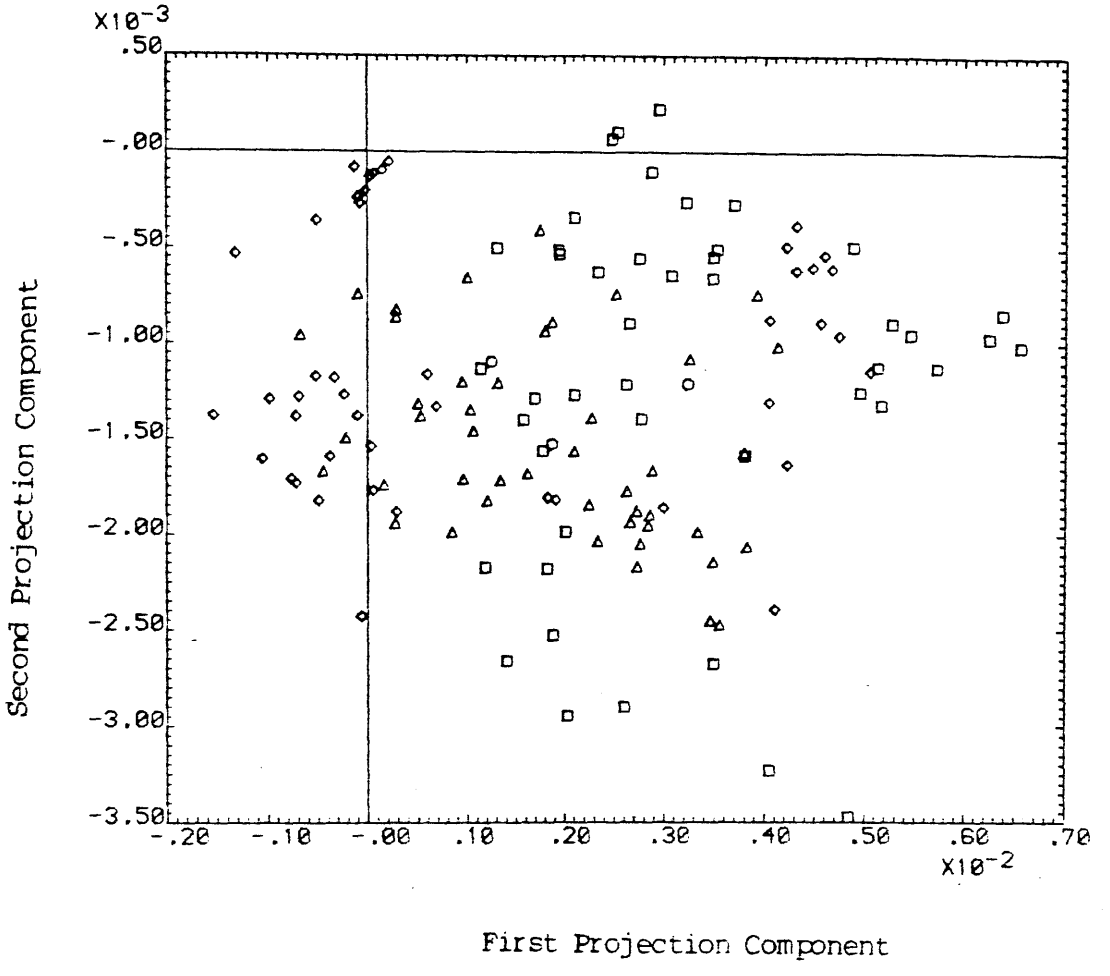


Figure 7.5 A two-dimensional projection of 135 20-dimensional lung sound feature vectors using the K-Y transformation. Each feature vector is normalized with respect to the total estimated power between 0 to 2000Hz. (△) denotes a single inspiration for a normal subject, (□) denotes that of an exposed patient and (◇) denotes that of an asbestosis patient.

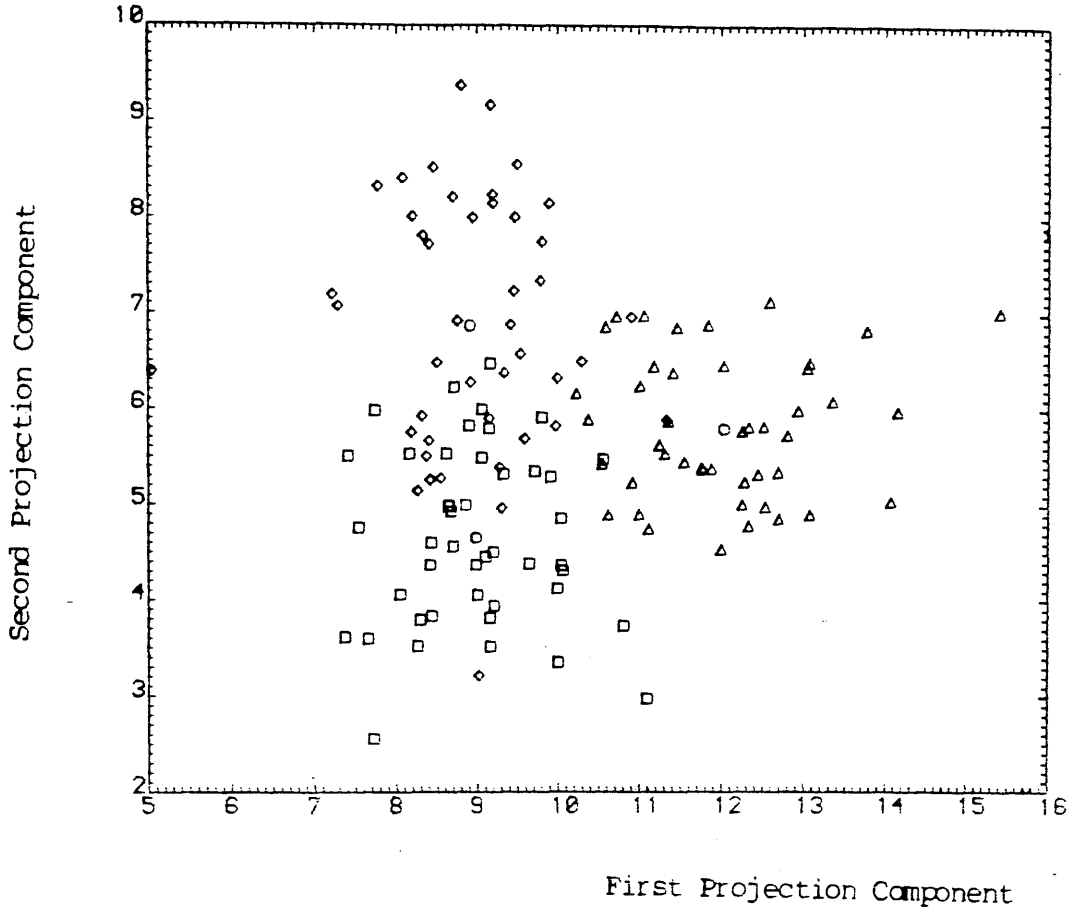


Figure 7.6 A two-dimensional projection of 135 20-dimensional lung sound feature vectors using the K-Y transformation. Each feature vector is normalized with respect to the largest estimated averaged power. Symbols as shown in figure 7.5.

"Does the new inspiration belong to the normal class?" or "Are there any relationships within the set of projected data?". (Another practical application of this visual aid has been shown in section 7.4 where different normalization techniques for the feature vectors can be compared visually.)

Five linear mapping algorithms have been implemented in this study:

(a) Kittler-Young (K-Y) transformation (section 4.2.3),

(b) Karhunen-Loeve (K-L) transformation (section 4.2.2),

(c) Fisher (F-S) transformation (section 4.2.4),

(d) Fukunaga-Mantock (F-M) transformation with the un-modified weighting function (equation 14 in section 4.2.5), and

(e) F-M transformation with the modified weighting function (equation 15 in section 4.2.5).

Figures 7.6 to 7.10 show the different 2-dimensional projections achieved by using the above 5 methods for 135 20-dimensional lung sound feature vectors, 45 from each group of subjects studied. It can be seen that the K-Y transformation provides the best 2-dimenional projection for this particular set of data.

Furthermore, figure 7.6 shows that there are 3 slightly overlapped clusters in the projection, each corresponding to one group (or class) of the subjects studied. A greater overlapping can also be seen between the exposed and the asbestosis patients. (See Appendix 7.A for further comments on figure 7.6.) This is to be expected because, as mentioned in section 2.5.1, the initial pathological change of asbestosis is an alveolitis. The degree

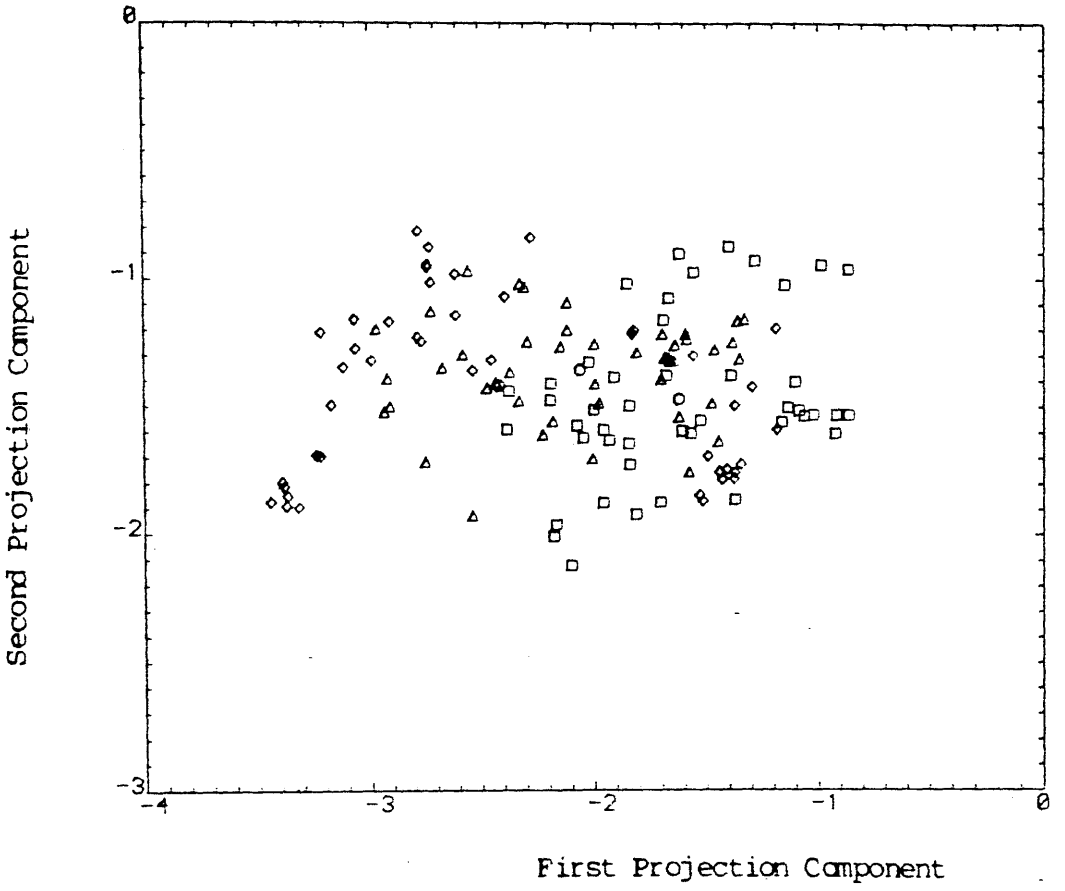


Figure 7.7 A two-dimensional projection of 135 20-dimensional lung sound data using the K-L transformation. Symbols as shown in figure 7.5.

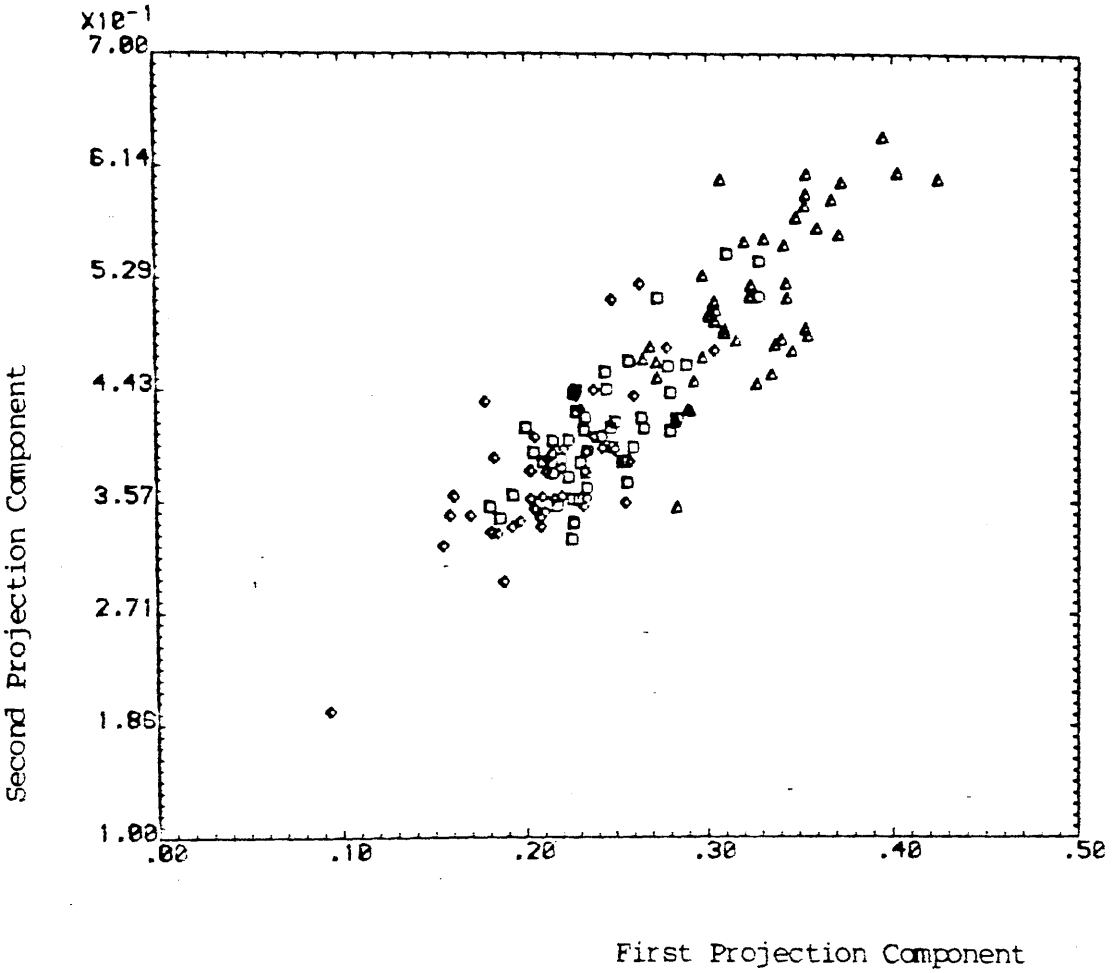


Figure 7.8 A two-dimensional projection of 135 20-dimensional lung sound data using the F-S transformation. Symbols as shown in figure 7.5.

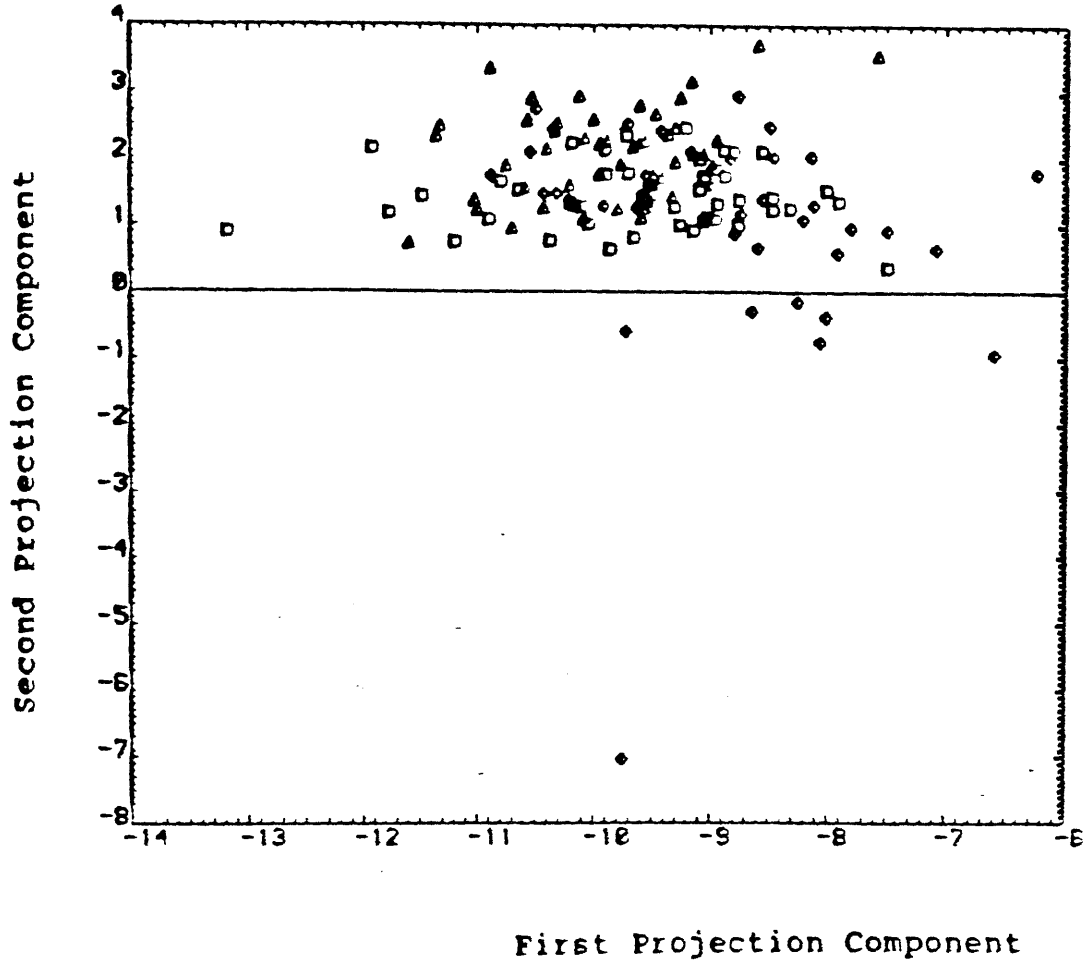


Figure 7.9 A two-dimensional projection of 135 20-dimensional lung sound data using the F-M transformation with the original weighting function. Symbols as shown in figure 7.5.

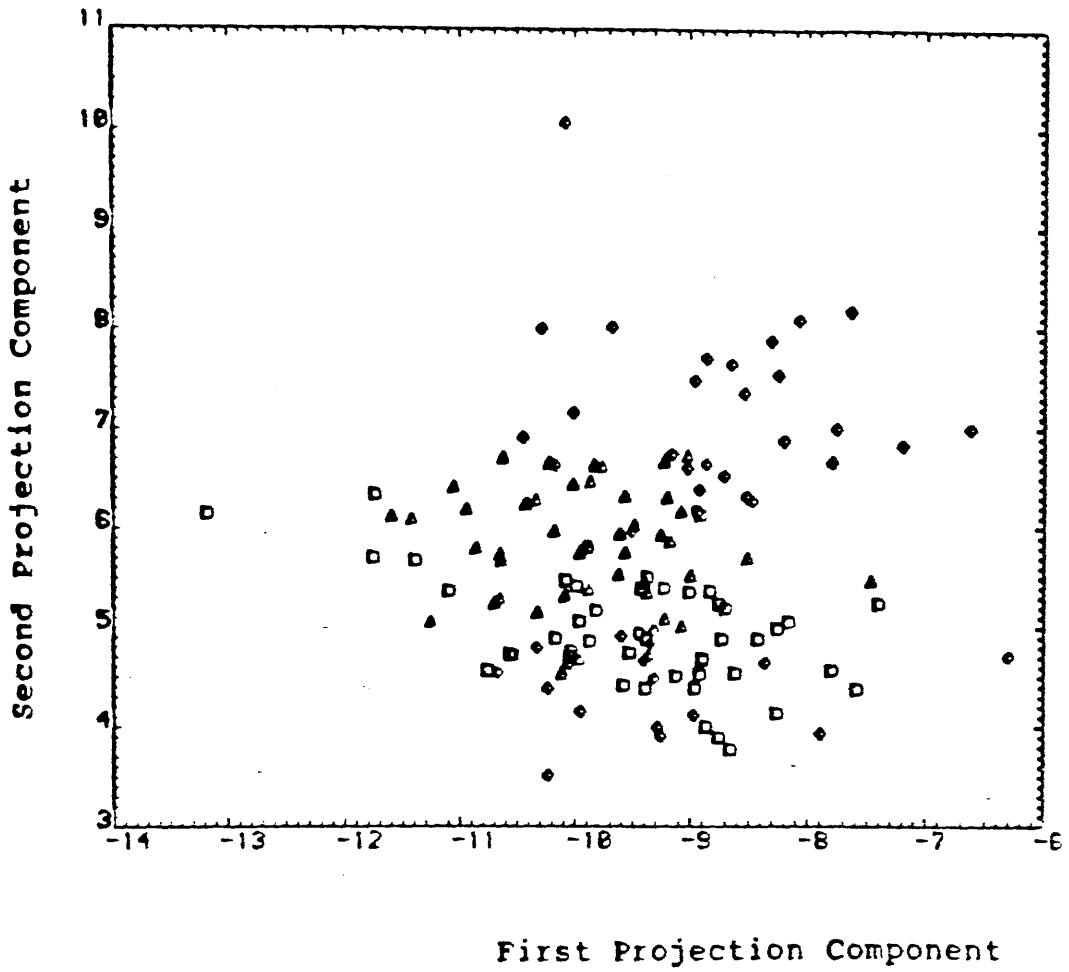


Figure 7.10 A two-dimensional projection of 135 20-dimensional lung sound data using the F-M transformation with the modified weighting function. Symbols as shown in figure 7.5.

or intensity of the alveolitis will dictate the progress of the disease, which varies from individual to individual. This similar initial pathological change in the lung parenchyma may account, in part, for the greater overlapping of the clusters between the two classes.

Each component of the eigenvector (column of the mapping matrix) can be considered as a weight for the corresponding component in the feature vector. Thus each component of the transformed feature vector is a linear weighted sum of the original 20-dimensional feature vectors. Hence, the relative importance of each component in the 20-dimensional feature vector can be assessed by examinating the corresponding component in the eigenvector. This also indicates the relative importance of each frequency interval from which the feature is generated. The components of each of the two eigenvectors formed by the K-Y transformation are listed in table 7.1. Careful inspection of these components suggested that both the high and low frequency components are of importance. Anderson et al (1986) point out that this may indicate that differences in breath sounds as well as the presence of crackles are responsible for the separation of the three groups in figure 7.6. It also brings out the potential of this type of analysis technique.

It is also worth noticing, in figure 7.9, the effect of outliers on the resulting projection when using the original F-M transformation, i.e. method (d). This effect can be reduced by using the modified weighting function suggested in section 4.2.5,

Table 7.1 Components of the 2 eigenvectors formed by the K-Y transformation.

Components	First Eigenvector	Second Eigenvector
1	6.86	3.84
2	-0.57	4.89
3	8.39	-5.76
4	-3.90	-2.75
5	0.34	-3.65
6	-6.00	6.53
7	1.73	-1.10
8	-6.56	6.58
9	7.06	4.19
10	7.70	3.45
11	8.12	-12.18
12	-9.16	-1.28
13	-2.21	-2.05
14	4.67	2.19
15	-5.55	0.46
16	-11.63	1.08
17	-1.94	8.95
18	7.79	-0.56
19	3.70	-0.66
20	2.21	-4.82

as shown in figure 7.10.

7.6 Nearest Neighbour Classification

Section 7.5 has established that the mapping module can provide the medical professionals with some useful visual displays and, thus, also suggests the possibility of using such displays in examinating patients, provided a suitable mapping matrix has been found. In that case, the operator of the non-invasive examination system can, through these displays, determine whether a patient has asbestosis or not. (It must again be emphasised that the proposed system should not be considered as a replacement for other diagnostic tools. Its aim is to provide a simple, routine, noninvasive and relatively reliable indicator for asbestosis, thus off-loading some of the unnecessary burden for the medical professionals and hopefully enabling the physicians to pay more attention to those patients who are suffering from the ailment.)

If the environment is perfectly known during the design stage, these displays will probably be sufficient. Unfortunately, the environment is usually only partially known to the designer. Thus, when the system becomes operational, it may be necessary to detect the new or changing environment (if such information is available). One of the simplest ways is to use the nearest neighbour (NN) classification rules. (Reasons for preferring NN to other types of classification rules were discussed in section 5.2.1.) Here, the classification scheme is

used to indicate whether the existing information about the system is sufficient of not. A simple indicator will be the number of errors and/or rejections (if a reject option is used) committed by the NN classification rule. However, when the proposed system is commissioned, the new lung sound signals from some new subjects (or using the terminology introduced in chapter 5, i.e. the test samples) will be unclassified (i.e. the operator will not know to which group the new subject belongs). Thus it is impossible to know whether a test sample has been classified correctly or not. Nevertheless, the system that is proposed here is off-line in nature. Therefore, it is possible to estimate the average probability of error (or error rate) for a batch of test samples (Fukunaga and Kessel, 1971). If the existing information is inadequate, the average error rate should increase above a threshold t, which is predefined by the designer. At that point, some other mechanisms can be invoked to reveal the new information, for example the data exploratory technique used by Urquhart (1983). Unsupervised learning algorithms can also be used to reveal such information (chapter 8).

The first task in constructing this module is to select a NN classification rule that will give the smallest expected probability of error. Unfortunately, the number of classified lung sound feature vectors (prototype samples) available in this study is too small to permit the implementation of those methods discussed in section 5.2.2.4. Although it is possible to implement those discussed in sections 5.2.2.2, 5.2.2.3 and 6.3, the small number of prototypes almost immediately precludes any

sensible comparison between these algorithms (even if the leave-one-out method is used to evaluate the probability of error, Devijver and Kittler, 1982). Nevertheless, section 6.2 provides an interesting theoretical study on two of the algorithms (i.e. the k-NN rule, section 5.2.2.2.1, and the distance-weighted k-NN rule, section 5.2.2.3.3). It has been shown in section 6.2.3 by a simple numerical example that the distance-weighted k-NN rule may have a slightly smaller expected probability of error than the k-NN rule when the number of prototypes is finite. It is, of course, possible to reduce the probability of error further by removing some of the "bad" prototypes using editing algorithms (see Luk, 1987). This is important because when the proposed system becomes operational, it is very likely that all the modules would be implemented in a dedicated microcomputer with limited amount of memory. Therefore, the number of prototypes that can be included will be limited. Furthermore, it is also possible to speed up the searching process in the nearest neighbour rule by implementing one of the algorithms surveyed in Luk (1987). (The author would recommand an algorithm by Yunck, 1976, because it is simple and does not involve any multiplication operations - possibly an important factor when using a dedicated microcomputer without a specialized mathmatical co-processor.)

The second task in constructing this module is to estimate the averaged probability of error for a set of unclassified test samples. This is in fact one of the most difficult questions facing researchers in the field of NN

classificaton. Most of the proposed solutions have been centred on the relatively simple two-class problem. Nevertheless, Devijver (1985) has recently suggested a method of finding an unbiased estimate of the probability of error for the 1-NN rule in a multiclass environment, which is based on the idea used by Fukunaga and Kessel (1971). The concept is very simple because it simply counts the number of nearest neighbours in each class. Again, due to the small number of prototypes available in this study, it is impossible to experiment with it on the lung sound data. However an experiment has been conducted using 3 classes of bivariate data similar to those described by Dudani (1976). In this experiment, the prototype set consisted of 600 samples, 200 from each class; and there were 5 independently generated test sets, each containing 3000 samples. The estimation was, therefore, executed 5 times, each time with a different test set, and the results were averaged. Figure 7.11 shows the averaged probability of error versus the number of nearest neighbours used for the estimation. It can be seen that this method does not seem to converge as expected from the theoretical study by Devijver (1985).

7.7 Remarks

In this study, the frequency range between 0 and 2000Hz has been evenly divided into 20 intervals and a feature is generated from each interval. A projection of the 3 classes of data onto a 2-dimensional subspace was found not to exhibit excessive interpenetration provided the K-Y transformation was

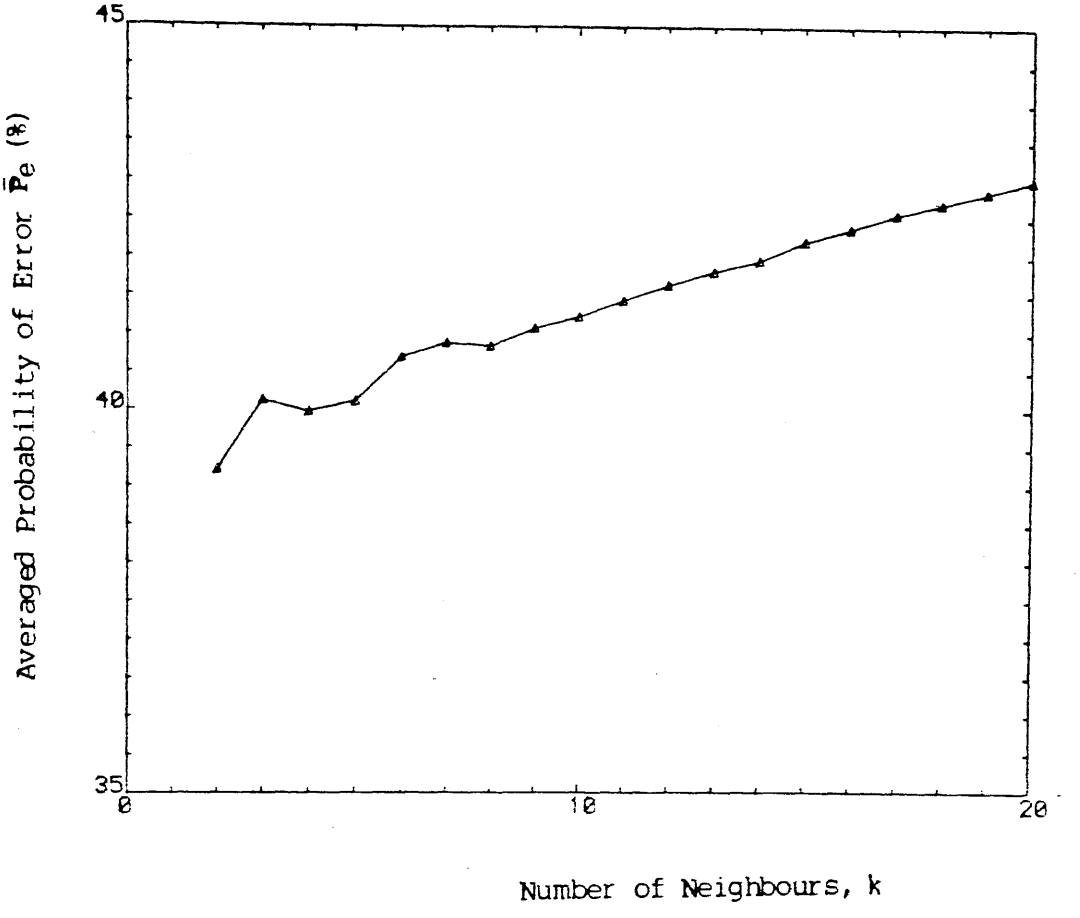


Figure 7.11 Averaged probability of error $\bar{P}_e$ versus number of nearest neighbour k used for the estimation.

used. However, it is felt that the present approach may have deemphasized the lower frequency components, especially those between 0 and 400Hz which from the previous study (Urquhart, 1983) have been shown to be relatively important. It is likely that a finer division at the lower frequencies may yield some useful features, which when combined with the features derived from the higher frequencies, may provide a better display than that in figure 7.6. This, however, will require further investigation.

Another problem related with the mapping module is the variance of the mapping matrix. As mentioned in section 4.2.5, since the mapping matrix is a function of the estimated scatter matrix (section 4.2.1), the variance of the mapping matrix is proportional to the variance of the estimated scatter matrix; and the variance of the estimated scatter matrix depends on the number and the dimensionality of the prototypes that were used to estimated it (figure 4.7). Thus it follows that the mapping matrix also depends on these two variables. In this study, the dimensionality of the feature vector is 20 and the number of prototypes is only 135. It is therefore not surprising that the mapping matrix in figure 7.6 may not work well with unseen data.

To see this effect, another 45 prototypes (15 from each class of subjects) that have not been used in the above study, were used. Figure 7.12 shows a 2 dimensional projection formed by using the eigenvectors listed in table 7.1 for all 180 20-dimensional lung sound feature vectors. It can be seen that

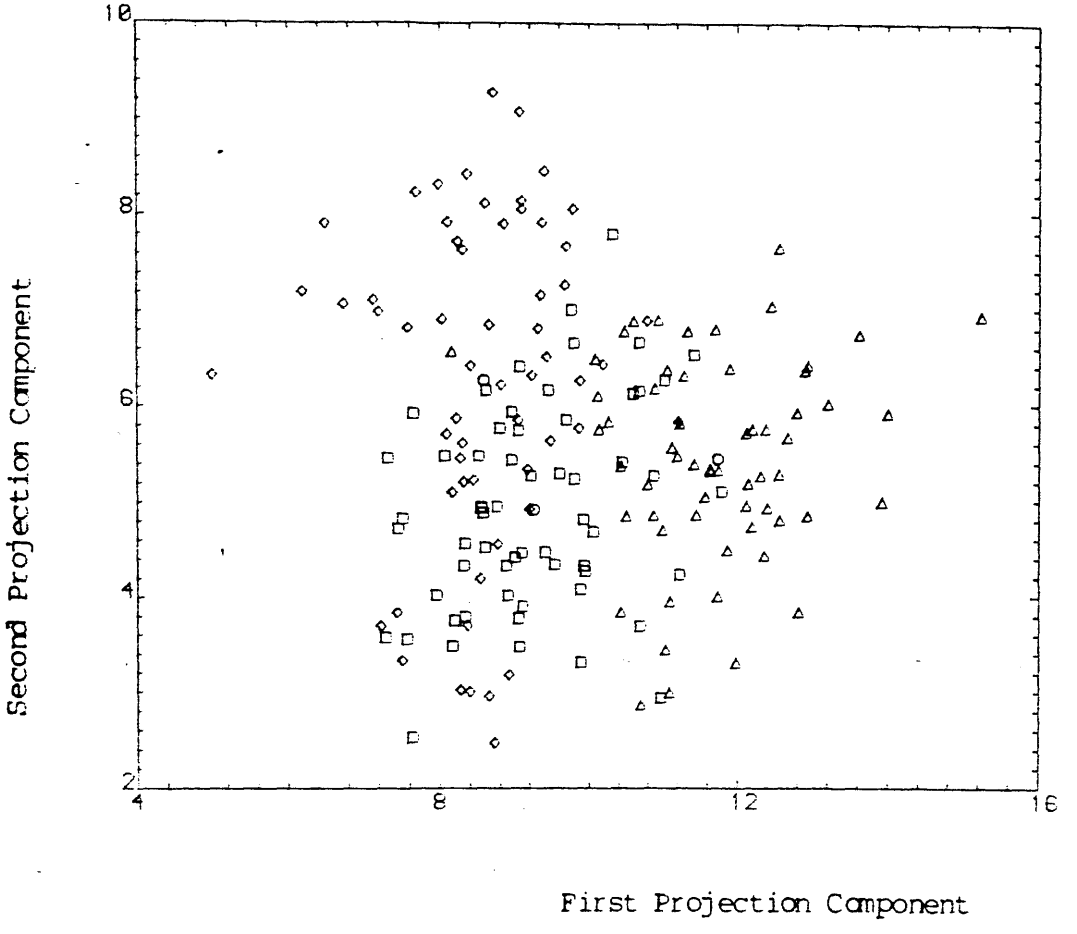


Figure 7.12 A two dimensional projection of 180 20-dimensional lung sound data using the mapping matrix formed by the 2 eigenvectors listed in table 7.1. Symbols as shown in figure 7.5.

there is more overlapping between the three classes (compare testing a classifier on the union of the training set and an unseen data set). Thus, the mapping matrix formed by using the eigenvectors listed in table 7.1 does not provide the best separation between the classes. Figure 7.13 shows the new transformed 2 dimensional space when K-Y transformation is applied again on all 180 20-dimensional lung sound data. It can be seen the whole projection has been rotated and better separation between the three classes is achieved when compared with figure 7.12.

One very important factor that has been neglected during this study is the age difference between the normal subjects and those asbestosis and exposed patients. All the normal subjects are between 25 and 35, whereas those asbestosis and exposed patients are between 50 and 70. Thus, the good separability between normal subjects and those asbestosis and exposed patients in figures 7.6 and 7.13 may, in part, due to this age effect.

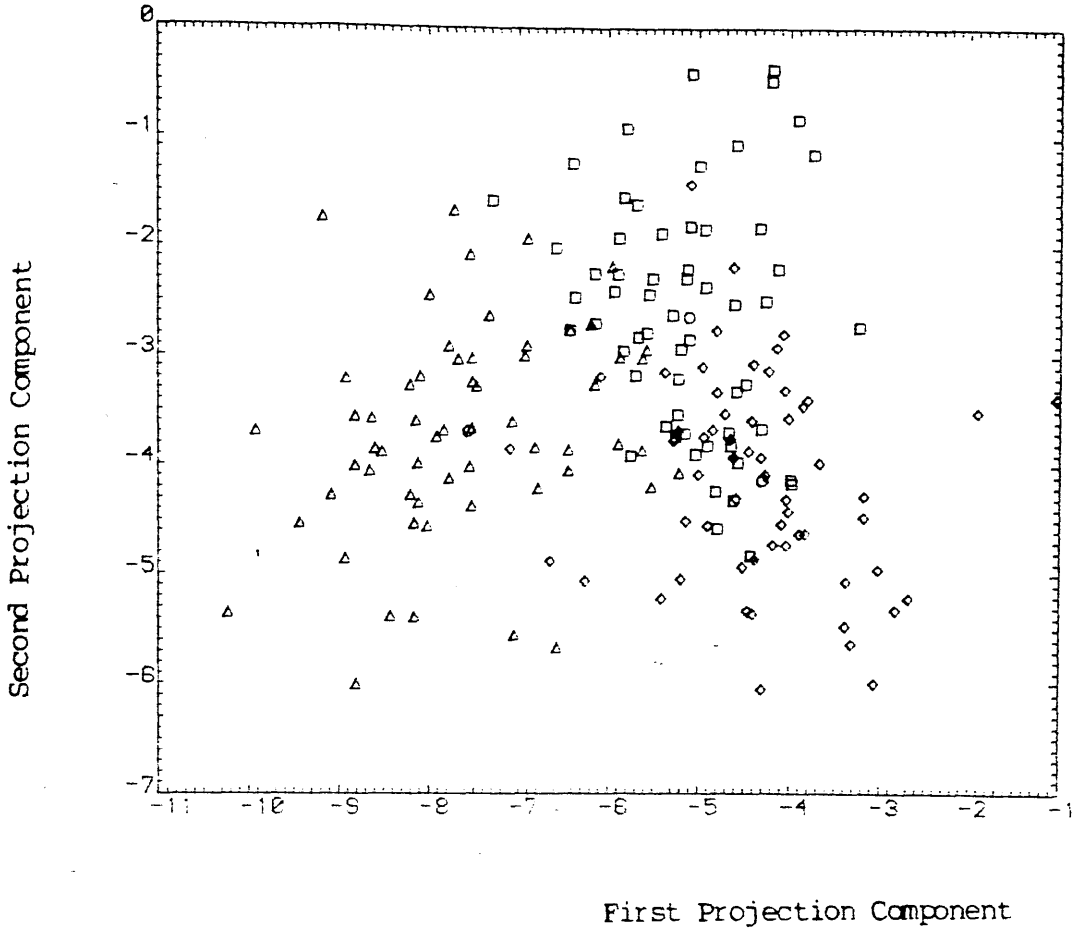


Figure 7.13 A two dimensional projection of 180 20-dimensional lung sound data using the K-Y transformation. Symbols as shown in figure 7.5.

Appendix 7.A: Further comments on figure 7.6

In figure 7.6, three slightly overlapped clusters are observed. In that figure each point in the projected space corresponds to an inspiration from one of the subjects studied. Unfortunately, it is impossible to infer from that figure which set of points originates from which subject. This is important because a set of points from a particular one subject may form a localized cluster inside one of the three main clusters observed in figure 7.6, and hence the observed separability between the three groups of subjects does not necessarily represent a natural separation between the groups, but may be heavily influenced by one or more subjects under study.

Figure 7.6 was therefore replotted in figure 7.14 with different symbols representing different subjects: normal subjects are identified by different upper case characters, exposed subjects are denoted by lower case characters and asbestosis patients are represented by Greek characters. It can be seen that the characters representing different subjects are quite widely spaced within their respective main clusters although the asbestosis subjects tend to scatter in a much wider area than either exposed or normal subjects. Considerable overlapping can be observed between two exposed subjects (represented by lower case a and e) with two asbestosis subjects (represented by the Greek characters δ and ε). Thus it still appears that separation is by disease class rather than by individual, but again in view of the small numbers of patients the results must be regarded as preliminary.

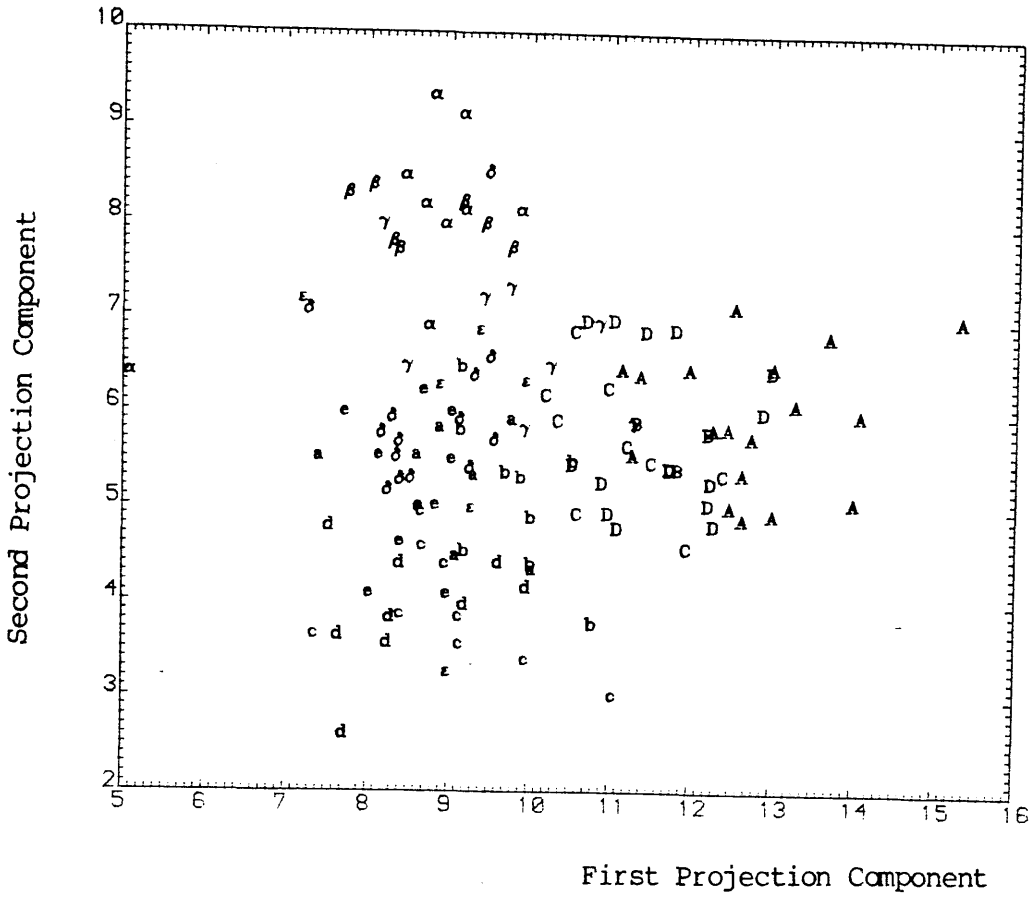


Figure 7.14 A replot of the 135 2-dimensional projected lung sound feature vectors from figure 7.6 with different symbols representing different subjects: normal subjects are identified by different upper case characters, exposed subjects are denoted by lower case characters and asbestosis patients are represented by various Greek characters.

References

[01] Akaike, H. (1974). "A new look at the statistical model identification". **IEEE Trans. on Automatic Control**, **AC-19**, 716-723.

[02] Anderson, K., Luk, A., Macleod, J.E.S. and Moran, F. (1986). "The application of pattern recognition and signal processing techniques in the diagnosis of asbestosis". **Thorax**, **41**, 715.

[03] Benedetto, G., Dalmasso, F., Guarene, M.M., Righini, G. and Spagnolo, R. (1983). "A method for the acoustical analysis of respiratory crackles in crytogenic fibrosing alveolitis". **IEEE Trans. on Biomedical Engineering**, **BME-30**, 620-623.

[04] Campbell, D.D. (1983). Personal communication.

[05] Carter, G.C. and Nuttall, A.H. (1980). "On the weighted overlapped segment averaging method for power spectral estimation". **Proceedings of the IEEE**, **68**, 1352-1354.

[06] Devijver, P.A. and Kittler, J. (1982). **Pattern Recognition: A Statistical Approach**, Prentice-Hall International, London.

[07] Devijver, P.A. (1985). "A multiclass k-NN approach to Bayes risk estimation". **Pattern Recognition Letters**, **3**, 1-6.

[08] Dudani, S.A. (1976). "The distance-weighted k-nearest neighbour rule". **IEEE Trans. on Systems, Man, and Cybernetics**, **SMC-6**, 325-327.

[09] Ertel, P.Y., Lawrence, M., Brown, R.K. and Stern, A.M. (1966a). "Stethoscope acoustics: I the doctor and his stethoscope". **Circulation**, **34**, 889-898.

[10] Ertel, P.Y., Lawrence, M., Brown, R.K. and Stern, A.M. (1966b). "Stethoscope acoustics: II transmission and filtration patterns". **Circulation**, **34**, 899-909.

[11] Fukunaga, K. and Kessel, D.L. (1971). "Estimation of classification error". **IEEE Trans. on Computers**, **C-20**, 1521-1527.

[12] Gavriely, N., Palti, Y. and Alroy, G. (1981). "Spectral characteristics of normal breath sounds". **Journal of Applied Physiology: Respiratory, Environmental and Exercise Physiology**, **50**, 307-314.

[13] Guard, D.R. (1976). "The generation of breath sounds and their transmission through the chest wall". Ph.D. Thesis, University of Southampton.

[14] Howie, K. (1981). "The frequency response of a microphone for measuring human lung sounds". Final year project report, University of Glasgow.

[15] Kay, S.M. and Marple. S.L. (1981). "Spectrum analysis: a modern perspective". **Proceedings of the IEEE**, **69**, 1380-1419.

[16] Kraman, S.S. (1983). "Does the vesicular sound come only from the lungs?". **American Review of Respiratory Diseases**, **128**, 622-626.

[17] Luk, A. (1987). "Nearest neighbour classification". Internal Report, University of Glasgow.

[18] McGhee, J. (1978). Unpublished results.

[19] Mori, M., Kinoshita, N., Morinari, H., Shiraishi, T., Koike, S. and Murao, S. (1978). "Spectral analysis of breath sounds". **Nippon Kyoho Shikau Gawkai Zasshi**, **16**, 503-512.

[20] Murphy, R.L.H. and Sorensen, K. (1973). "Chest auscultation in the diagnosis of pulmonary asbestosis". **Journal of Occupational Medicine**, 15, 272-276.

[21] Nuttall, A.H. (1981). "Some windows with very good sidelobe behaviour". **IEEE Trans. on Acoustics, Speech, and Signal Processing**, **ASSP-29**, 84-91.

[22] Sayers, B.McA. (1975). "Exploring biological signals". **Biomedical Engineering**, **10**, 335-341.

[23] Tierney, D. (1983). "Acoustic transducer for detecting human lung sounds". Final year project report, University of Glasgow.

[24] Urquhart, R.B., McGhee, J., Macleod, J.E.S., Banham, S.W. and Moran, F. (1981). "The diagnostic value of pulmonary sounds: a preliminary study by computer-aided analysis". **Computing Biology and Medicine**, **11**, 129-139.

[25] Urquhart, R.B. (1983). "Some new techniques for pattern recognition research and lung sound signal analysis". Ph.D. Thesis, University of Glasgow.

[26] Yuen, C.K. (1983). "Comments on spectral estimation using combined time and lag weighting". **Proceedings of the IEEE**, **71**, 535-536.

[27] Yuen, C.K. and Fraser, D. (1979). **Digital Spectral Analysis**, CSIRO-Pitman, London.

[28] Yunck, T.P. (1976). "A technigue to identify nearest neighbours". IEEE Trans. on Systems, Man, and Cybernetics, SMC-6, 678-683.

Conclusions and Suggestions for Future Research

This work was supported by the Croucher Foundation, Hong Kong.

Chapter 8: Conclusions and suggestions for future research

8.1 General conclusions on the proposed non-invasive examination system

Throughout the past three years of research, most of the effort has been focused on the feasibility of developing a non-invasive system that can routinely be used either by a physician or by trained para-medical personnel to examine patients who have exposed to asbestos dusts. At the begining of this project, it was decided that the system should be divided into a number of relatively independent modules. This has the advantage of developing the most needed module(s) initially, such as the data acquisition and preprocessing module, and so forming the backbone of the system. From this, other more sophisticated periphery such as the classification and unsupervised learning modules can be added whenever needed. This is particularly significant when one consider the inevitable limited number of patients available during this stage of development and also the demonstration nature of this project. Whenever possible, each module would be implemented with a number of options so that the user could select the desired alternative. For example, at the preprocessing stage, the user would be allowed to select different types of window other than the 4-term minimum Blackman-Harris window used in the WOSA algorithm. The user could, say, select a 4-term minimum Kaiser-Bessel window (Harris, 1978). This approach has the benefit of permitting the addition of some special algorithms at a later stage of the development, or the

re-adaptation of the existing algorithms for other special needs.

Obviously, the most important question in this chapter is to conclude whether it is possible to build the proposed system. It is the author's belief that such a system may be possible to realise in the near future. The main item of hardware would be a dedicated (transportable) microcomputer with a hard-disk of suitable size (cost probably less than £1000). The rest of the hardware cost will be for building a dedicated transducer, an analog-to-digital converter with anti-aliasing filter and possibly a display unit. On the whole, when the system is fully realised, its hardware cost can be considered as inexpensive.

Furthermore, from the results presented in chapter 7, it can be seen that the mapping module itself is possibly adequate, in most cases, for identifying patients with asbestosis provided a suitable mapping matrix can be found in the future. This matrix should give a small false negative and false positive rate when the new samples are identified by the classification module. It must be emphasized that the proposed system would not be used to diagnose asbestosis. Its function is to indicate whether other more troublesome and possibly invasive tests should be used on an otherwise normal subject. Thus, the classification module may also help to assist the decisions reached by a physician or by para-medical personnel. In this case, the decision given by the classification module will be extremely simple. Essentially, the answers will be "yes", "no" and/or

"don't know", depending on whether any reject option is used in the decision. Sections 5.2.2 and 6.3 have provided some very useful nearest neighbour types of algorithms for this classification module. For the proposed system, the "don't know" answer (or the use of reject option) can be used to reduce the false negative rate. For example, if the (k,ℓ_i)-NN rule is used (section 5.2.2.2.3), ℓ_i for the non-asbestosis classes can be adjusted to a value greater than $\lceil k/3 \rceil$, so that in order to be identified as non-asbestosis, a test sample has to have more neighbours in the non-asbestosis classes than in the asbestosis class. In addition, when the system is being developed further, the size of the prototype set will gradually increase. Thus, it may be necessary to select a smaller but better prototype set and/or to speed up the searching time for the set of nearest neighbours. The algorithms surveyed by Luk (1987) may therefore be useful.

The next question is whether the medical profession would accept, in part or as a whole, the proposed system. It is the author's opinion that it really depends on how one looks at the system. If the system is treated only as a screening tool, the author believes that it would be acceptable to the medical profession, provided further developmental work and other fine tuning is performed. If it is viewed as a medical research tool, some of the modules in the system are certainly useful, such as the mapping and classification modules: similar modules have been used in the past to handle a number of medical-related problems (Cox et al, 1972; Mendelson et al, 1973; Nagy, 1968). As a

diagnostic aid, on the other hand, the system, at least in the form now proposed, is less likely to be acceptable. Indeed, the proposed system lacks the grace of the so called knowledge-based systems (Friedland, 1985; Sutherland, 1986), which if properly programmed can give not only a decision but also the reasons behind a decision.

8.2 Conclusions on lung sound research

This project has made a contribution to lung sound research. In particular, the mapping results in chapter 7 have shown that it is possible to separate or discriminate asbestosis patients from non-asbestosis subjects with a low error rate simply by using lung sounds alone. In addition, it has also shown that exposed patients could be separated from normal subjects. This may partly due to the pathological changes as a result of exposure to asbestos fibres in the exposed subjects and may, in part, also be due to the age difference between the two groups. Another finding is that both the high and the low frequency components of the averaged periodogram are important. This may suggest that crackles play an important part in the discrimination process (Anderson et al, 1986). This does not contradict the previous study (Urquhart et al, 1981) which found abnormality at the lower frequencies. In fact, as Kraman (1983) has suggested, this may partly be contributed by the muscle tissue in the chest wall. Thus, the so called "lung sounds" may actually be a combination of sounds generated in the lung, the heart, and other connective tissues, making the interpretation of

lung sounds extremely difficult (Anderson et al, 1987). However, whatever the origin of these sounds, it is clear that they contain useful discriminatory information (Urquhart, 1983).

8.3 Conclusions on nearest neighbour classification

This work has also made a contribution to nearest neighbour (NN) classification. In section 6.2, it has been shown that the error rate for the weighted k-NN rule (section 5.2.2.3.3) may be smaller than for the unweighted k-NN rule (section 5.2.2.2.1) when the size of the prototype set is finite (Macleod et al, 1987). This conclusion may be important when one attempts to select which nearest neighbour algorithm to use in a given application. In addition, Luk and Macleod (1986) have also proposed another nearest neighbour classification scheme which is slightly different from other NN classification rules surveyed in section 5.2.2. Modifications, such as the addition of a reject option, are needed when the size of the prototype set is finite. They have suggested that the proposed rule may be useful in applications where a lower error probability is essential and a somewhat large rejection is tolerable, and when the (k,ℓ)-NN rule cannot achieve the necessary tolerable rejection rate. The non-invasive examination system, if used for preliminary screening as suggested in section 8.1, might constitute a possible application where false negatives are not desirable but false positives may be tolerable.

8.4 Suggestions for future research

One area that is certainly worth some more research effort is in the missing module, i.e. the unsupervised learning module. It is true that unsupervised learning has been around for more than 30 years, yet integrating this technique into the proposed system is by no mean simple. Perhaps, it is possible to follow the same line of research as Dasarathy (1980). His proposed system was capable of detecting a previously undetected class. However, for the proposed non-invasive system, will the problem be as simple as discovering the onset of a new respiratory disease? Or, is it just a simple case of machine malfunction? The problem is really what to look for when this module is invoked.

It is the author's wish that the work on the proposed non-invasive system can be continued because from chapter 7, it can be seen that many modules are still in their infancy. The research that has been performed up to now is really a one-off demonstration work. As suggested in section 8.1, more developmental work and fine tunings are needed before the system can be declared as operational. Undoubtedly, a more user friendly front-end is required. The author would certainly like to see an icon-driven front-end which could be used together with a "mouse" interface to drive the system starting from the data acquisition module.

One look at the frequency response of the transducer

used in this study will certainly suggest that a better one is required in the future. It would obviously be useful to explore different acoustic materials to find the best possible candidate for acquiring lung sound signals. It would also be useful to design a method to estimate the acoustic impedance of the chest wall. The most effective way would be to use cadavers.

It is also possible to develop new feature generation algorithms. This certainly will have an important effect on the subsequent projection in the mapping module, and hence on the classification module as well. The main question is what sort of features are needed and how to generate them. Can some features be generated from the time domain? Can the age, sex, weight, height, and other physiological data be incorporated into the feature generator? Again, the problem cannot be solved unless a definite objective is set for the development. For example, is the future system going to be a diagnostic aid or a screening device or a medical research tool? If it is a diagnostic aid, will it be more profitable to use the so called knowledge-based approaches? These undoubedly should be retained as future research.

Research into the possibility of using lung sound signals to monitor the prognosis of a certain lung disorder is also required. If for example a patient were examined on successive occasions and it was found that the point (assuming a mapping algorithm is used) representing his/her respiration was moving closer towards the cluster formed by that lung disorder,

this might be evidence for believing that the patient was developing that lung disorder. Again, unless some work is performed on this suggestion, it is not known whether the proposal is feasible or not.

In nearest neighbour (NN) classification it would be worth investigating the feasibility of using the proposed NN rule (section 6.3) in editing. If so, how would it compare with other editing algorithms? Some preliminary results (Luk and Macleod, 1985) have suggested that the proposed classification scheme (chapter 6, equations 19) can eliminate more prototypes in one iteration and achieve nearly the same probability of error if subsequent classification was performed by a 1-NN rule (Devijver and Kittler, 1982). Since the reject option is used in that implementation, it could also be worthwhile to study possible relationships between rejection and editing.

References

[01] Anderson, K., Luk, A., Macleod, J.E.S., Moran, F. (1986). "The application of pattern recognition and signal processing techniques in the diagnosis of asbestosis". **Thorax**, **41**, 715.

[02] Anderson, K., Luk, A., Macleod, J.E.S., Moran, F. (1987). "Interpretation of lung sounds". unpublished manuscript.

[03] Cox, J.R., Nolle, F.M. and Arthur, R.M. (1972). "Digital analysis of the electroencephalogram, the blood pressure wave, and the electrocardiogram". **Proceedings of the IEEE**, **60**, 1137-1164.

[04] Dasarathy, B.V. (1980). "Nosing around the neighbourhood: a new system structure and classification rule for recognition in partially exposed environment". **IEEE Trans. on Pattern Analysis and Machine Intelligence**, **PAMI-2**, 67-71, 1980.

[05] Devijver, P.A. and Kittler, J. (1982). **Pattern Recognition: A Statistical Approach**, Prentice-Hall International, London.

[06] Harris, F.J. (1978). "On the use of windows for harmonic analysis with the discrete Fourier transform". **Proceedings of the IEEE**, **66**, 51-83.

[07] Friedland, P. (1985). "Special section on architectures for knowledge-based systems". **Communications of the ACM**, **28**, 903.

[08] Kraman, S.S (1983). "Does the vesicular sound come only from the lungs?". **American Review of Respiratory Diseases, 128**, 622-626.

[09] Luk, A. and Macleod, J.E.S. (1985). "An alternative nearest neighbour classificaton scheme". Presented in the British Pattern Recognition Association Third International Conference, St. Andrews, 25-27 September, 1985.

[10] Luk, A. (1987). "Nearest neighbour classification". Internal Report, University of Glasgow.

[11] Macleod, J.E.S., Luk, A. and Titterington, D.M. (1987). "A re-examination of the distance weighted k-NN classification rule". **IEEE Trans. on Systems, Man, and Cybernetics**, accepted for publication.

[12] Mendelson, M.L., Mayall, B.H., Bogart, E., Moore, D.H., II and Perry, B.H. (1973). "DNA content and DNA-based centromeric index of the 24 human chromosomes". **Science, 179**, 1126-1129.

[13] Nagy, G. (1968). "State of the art in pattern recognition". **Proceedings of the IEEE, 56**, 836-862.

[14] Sutherland, J.W. (1986). "Assessing the artificial intelligence contribution to decision technology". **IEEE Trans. on Systems, Man, and Cybernetics, SMC-16**, 3-20.