OUTLIER MANAGEMENT IN INTELLIGENT DATA ANALYSIS

J. Gongxian Cheng

A thesis submitted in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy

Department of Computer Science

Birkbeck College

University of London

September 2000

ABSTRACT

In spite of many statistical methods for outlier detection and for robust analysis, there is little work on further analysis of outliers themselves to determine their origins. For example, there are "good" outliers that provide useful information that can lead to the discovery of new knowledge, or "bad" outliers that include noisy data points. Successfully distinguishing between different types of outliers is an important issue in many applications, including fraud detection, medical tests, process analysis and scientific discovery. It requires not only an understanding of the mathematical properties of data but also relevant knowledge in the domain context in which the outliers occur.

This thesis presents a novel attempt in automating the use of domain knowledge in helping distinguish between different types of outliers. Two complementary knowledge-based outlier analysis strategies are proposed: one using knowledge regarding how "normal data" should be distributed in a domain of interest in order to identify "good" outliers, and the other using the understanding of "bad" outliers. This kind of knowledge-based outlier analysis is a useful extension to existing work in both statistical and computing communities on outlier detection. In addition, a novel way of visualising and detecting outliers using self-organising maps is proposed, an active control of data quality is introduced in the data collection stage, and an interactive procedure for knowledge discovery from noisy data is suggested.

The methods proposed for outlier management is applied to a class of medical screening applications, where data were collected under different clinical environments, including GP clinics and large-scale field investigations. Several evaluation strategies are proposed to assess various aspects of the proposed methods. Extensive experiments have demonstrated that that problem-solving results are improved under these proposed outlier management methods. A number of examples are discussed to explain how the proposed methods can be applied to other applications as a general methodology in outlier management.

Dedicated to the memory of my Aunt, Maofen.

Acknowledgements

It has been a long journey since I started the work presented in this thesis. So when it comes to the time to write this page, it would not be easy to enumerate all of the support that I have received throughout my studies. I might not even know some of the people involved, but I am sure that without many of them, the thesis would not have become a reality.

I especially appreciate the numerous inspirational and passionate brainstorming sessions with Dr Xiaohui Liu and Dr John Wu, which have contributed a great deal to this thesis.

I am also most grateful to Professor Barrie Jones, who not only has secured part of the funding for this study, but more importantly, has also offered unreserved support throughout my studies at the University. He has also helped review many of my papers and has organised several trial runs with the IDA system, which collected critically important test data. My gratitude also goes to Professor George Loizou. Not only has he supported my studies, but he has also been a major driving force for the project.

I must also thank Dr Roger Johnson, Dr Ken Thomas, Dr Trevor Fenner, Dr Roger Mitton and Dr Phil Docking at Birkbeck College, Professor Roger Hitchings and Dr Richard Wormald at the Moorfields Eye Hospital, Professor Gordon Johnson and Professor Fred Fitzke at the Institute of Ophthalmology, University College London, and Dr Simon Cousens at London School of Hygiene and Tropical Medicine, who have all played important roles at different stages in my studies.

I wish to thank Dr Stephen Corcoran for collecting GP data from the MRC pilot study, and Professor A. Abiose at the National Eye Centre of Kaduna, Nigeria, Mr B. Adeniyi and other members of the onchocerciasis research team, for collecting the onchocercal data from field trials in Nigeria, and Grace Wu for collecting glaucomatous data at the Moorfields Eye Hospital.

I am also thankful to Jonathan Collins, Michael Hickman, Sylvie Jami and Kwa-Wen Cho, who provided help for many experiments related to this thesis.

The United Nations Development Programme (UNDP) and the British Council for Prevention of Blindness have provided funding for my study; UK's Medical Research Council and the UNDP, the World Bank and the World Health Organisation have funded the large scale of field trials for this system. Vital data were obtained from these trials.

Words are not enough to express my thanks to my supervisor, Dr Xiaohui Liu, who has been vigorously supporting my studies, guiding my thesis, and making my life in London so enjoyable. The same thanks also goes to Dr John Wu and his family – it has been such a precious experience with you, that it makes me remember forever.

This thesis would not be possible without my wife Chunyu. Thank you for your love and encouragement. I would also like to thank our beloved Sabrina for being quiet and understanding during the writing of the thesis.

Last but by no means least, thanks to my parents for loving me, believing in me, and empowering my thought throughout my life.

Declaration

The work presented in the thesis is my own except where it is explicitly stated otherwise.

The papers have been jointly published with my supervisor, Dr X. Liu, and with fellow research workers, notably Dr J. Wu. However, I have personally conducted the research reported in the thesis, and am responsible for the results obtained.

Gongxian Cheng: ____________ Xiaohui Liu: ____________

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

ACRONYMS

Acronym	***Descriptions***
AI	Artificial Intelligence
CCVP	Computer Controlled Video Perimetry
CPT	Conditional Probability Tables
CRT	Cathode Ray Tube
DAG	Directed Acyclic Graph
GP	General Practitioner
IDA	Intelligent Data Analysis
IUI	Intelligent User Interface
ISO	International Standard Organisation
LCD	Liquid Crystal Display
METRO	Making Eye Test Reliable for Ophthalmology
MRC	Medical Research Council
OLAP	On-Line Analytic Processing
ROC	Receiver Operator Characteristic
SAT	Single Amplitude Trial
SOM	Self-Organising Maps
TDIDT	Top Down Induction of Decision Trees
VQ	Vector Quantisation
WHO	World Health Organisation

Chapter 1 Introduction

1.1 Intelligent Data Analysis

Intelligent data analysis (IDA) is an interdisciplinary study concerned with the effective analysis of data. It requires careful thinking at every stage of an analysis process, intelligent application of relevant domain expertise regarding both data and subject matters, and critical assessment and selection of relevant analysis methods. Although statistics has been the traditional method for data analysis, the challenge of extracting useful information from large quantities of on-line data has called for advanced computational analysis methods.

For the last decade or so, the size of machine-readable data sets has increased dramatically and the problem of "data explosion" has become apparent. On the other hand, recent developments in computing have provided the infrastructure for fast data access; processing and storage devices continue to become cheaper and more powerful; networks provide more data accessibility; widely available personal computers and mobile computing devices enable easier data collection and processing; and techniques such as On-Line Analytic Processing (OLAP) [35] allow rapid retrieval of data from data warehouses. In addition, many of the advanced computational methods for extracting information from large quantities of data, or "data mining" methods, are beginning to mature, e.g., artificial neural networks, Bayesian networks, decision trees, genetic algorithms, and statistical pattern recognition [139]. These developments have created a new range of challenges and opportunities for intelligent systems in data analysis [64, 115].

In the quest for intelligent data analysis, the issue of data quality has been found to be one of the particularly important ones. Data are now viewed as key organisational resources and the use of high-quality data for decision-making has received increasing attention [161]. It is commonly accepted that one of the most difficult and costly tasks in large-scale data analysis is trying to obtain clean and reliable data. Many have estimated that as much as 50% to 70% of a project's effort is typically spent on this part of the process [116].

Data quality is a relative concept. Nowadays multiple uses of data for different purposes are common in the context of data warehouses. The true meanings of items in the initial databases may be lost in the transition to a data warehouse as the initial data gatherers or database developers may no longer be involved in the transition. Moreover, data with quality considered reasonable for one task may not be of suitable quality for another task.

There are many potential quality problems with real-world data. First, data may be *noisy* for a variety of reasons: faulty data collection instruments, problems with data entry, transmission errors, confusion over the correct units of measurements etc. Second, data may be *missing* due to problems with manual data entry, sensitive nature of data required, or for reasons of economy. Third, data items may be *inconsistent* in that they may be recorded in different formats or in an inappropriate reporting period. Fourth, even when the data are accurate, complete, consistent, and so forth, they may not be timely. This is often a crucial consideration in many real-time applications such as foreign exchange, the stock market, and process control. Last but not least, it is not always easy to separate noisy data items from the rest of the data. For example, anomalous data points are often measurement or recording errors but some of them can represent phenomena of interest, something significant from the viewpoint of the application domain. To distinguish between these two requires a careful application of relevant domain knowledge.

In response to this opportunity, data cleaning companies are being created, and data quality groups are being set up in institutions. Since the use of the "wrong" data or low-quality data often leads to erroneous analysis results, research on data quality has attracted a significant amount of attention from different communities, including information systems, management, computing, and statistics [178, 168, 145, 106, 143].

1.2 Outlier Management

1.2.1 Issues and Challenges

A strange data value that stands out because it is not like the rest of the data in some sense is commonly called an *outlier*. An outlier may appear as an extreme value or a peculiar combination of the values in multivariate data. The handling of outlying or anomalous observations in a data set is one of the most important tasks in intelligent data analysis, because outlying observations can have a considerable influence on the analysis results. Altman has given an example in regression analysis where the presence of a single outlier has greatly altered the regression result ([4], figure 7.2). A number of regression diagnostics have been developed to identify the statistical influence of outliers [52, 15], which can be used to check the significance of outlier influence in a given application.

Although outliers are often measurement or recording errors, some of them can represent phenomena of interest, something significant from the viewpoint of the application domain. Consequently, simply rejecting all outliers may lose useful information, and lead to inaccurate or incorrect results in data analysis tasks. For example, in fraud detection, suspicious credit card transactions may indeed be fraudulent, but could also be those looking-suspicious, but legitimate ones. In hand-written character recognition, a good outlier might be an atypical but legitimate pattern, while a bad outlier might be a "garbage" pattern. In [60], Guyon and Stork acknowledged the importance of explicitly treating outliers, and pointed out the problems of using robust analysis methods in this type of applications. Matic *et al.* has shown that by individually examining outliers and only cleaning those error patterns, the classification error has decreased from 11% to 6.7%, whereas the error rate would increase if all outlying patterns were cleaned [122]. In process control, abnormal multivariate time series observations collected from a plant might be an indication of excessive stress of certain equipments, or simply a result of changes made by operators. In the former case, appropriate action should promptly be taken to prevent the escalation of the abnormal events leading to possible plant shutdown, while nothing much needs to be done for the latter [102].

In certain situations, outliers may lead to the discovery of unexpected knowledge [19]. In the 1880s when the English physicist Rayleigh measured nitrogen from different sources, he found that there were small discrepancies among the density measurements. After closer examination, he discovered that the density of nitrogen obtained from the atmosphere was always greater than the nitrogen derived from its chemical compounds by a small but definite margin. He reasoned from this anomaly that the aerial nitrogen must contain a small amount of a denser gas. This discovery eventually led to the successful isolation of the gas argon, for which he was awarded the Nobel Prize for Physics in 1904.

In the last few decades, many effective statistical analysis methods have been developed to deal with outliers. However, statistics alone may provide insignificant or insufficient help to some applications, where different types of outliers share similar statistical properties regardless of their nature. In [8], Barnet and Lewis have given an example for the case when Mr Hadlum appealed against the failure of an early petition for divorce. His claim was based on an alleged adultery by Mrs Hadlum, the evidence for which consisted of the fact that Mrs Hadlum gave birth to a child 349 days after Mr Hadlum had left the country to serve the nation. The average gestation period for a human female is 280 days, so 349 days appeared surprisingly long (an outlier). The judges ruled that 349 days, although improbable, was not beyond the range of scientific possibility. However, the alleged adultery could equally be true. It is apparent that more background knowledge was needed, and statistics alone was not sufficient to make the judgement in this case.

Because outliers often share similar statistical properties, it would be difficult to distinguish between them without relevant domain knowledge that leads to a basic understanding of why they are outlying and what the underlying data generation mechanism is. In order to judge whether an outlier is informative or useful in a practical context, other information is often needed, such as relevant domain or common-sense knowledge, or the experience of data analysts in relation to judging outlying data points, etc. To date, the progress in the explicit management of outliers has been largely restricted to the automated detection but manual analysis of outliers, as in the investigation of credit card fraud and inside dealing at stock markets [57], in hand-written character recognition [61], or in the study of customer behaviour [121]. Only after the knowledge becomes available and represented in a computable format, is it then possible to develop automated methods to prevent the useful outliers from being precluded by the IDA process.

Because of the IDA's unique connection between statistics and AI, it becomes IDA researchers' special challenge and responsibility to analyse outliers detected by statistical or other processes, in a domain-sensitive manner. This dissertation will focus on outlier management, a fundamental issue in addressing the data quality problem.

1.2.2 An Intelligent Data Analysis Approach

Over the last few years, I have developed strategies to analyse and manage outliers based on domain knowledge. The approach assumes that detected outliers should not be automatically deleted, but should be further analysed using relevant knowledge in connection with particular problem-solving tasks.

The strategies I have proposed include the application of various AI techniques to different analysis tasks. These techniques include knowledge acquisition, knowledge representation, and supervised and unsupervised machine learning techniques. Application-specific expertise is acquired and represented, and then applied to detected outliers. Domain knowledge is used to model the data based on the established problem-solving objectives and the form of knowledge available. I will show that outliers can be distinguished by those knowledge-based models. The combined strengths of artificial intelligence and statistics make it possible to achieve such an objective towards the "application-sensitive" outlier analysis – a process of both fields working in harmony that I consider an integral part of an IDA system.

The study in the thesis extends the existing work in both statistical and computing communities on outlier detection by distinguishing between different types of outliers. The type of knowledge-based outlier analysis method is a novel approach to outlier management and constitutes a useful contribution to the data quality issue in intelligent data analysis research.

Apart from the outlier analysis strategies, I have also introduced a method to detect outliers in multivariate data using self-organising maps (SOM) [95]. This method is based on the "neighbourhood preservation" property of the SOM, that similar data points are mapped onto close output nodes on the two-dimensional map. It facilitates outlier visualisation and identification so that the outliers can be explicitly analysed. In particular, because the SOM directly maps the input data into clusters on the output map, it becomes possible to "explain" the clusters by investigating the associated input data. The clusters on the output map can then have "physical meanings," and the information may be used to visualise and interpret the relationship between the main population and the outliers, which are mapped onto different areas on the output map.

I have also recognised that reducing noise at the source is another important step in outlier management. If the data quality can be improved while the data are collected, the task of outlier detection and analysis may be easier and perhaps less dependent on the data model to detect and distinguish the outliers. This reduces the risk that too many outliers would have to be analysed, especially when domain knowledge is incomplete.

Since the ultimate goal of outlier management is to improve the problem-solving tasks in application domains, the entire process is completed by applying the method to a challenging real-world application where the proper management of outliers may considerably affect the system performance. The outlier management methodology can be summarised as follow steps.

1) Reducing noise at source
2) Detecting outliers with the SOM
3) Reasoning about outliers using domain-specific knowledge
4) Using the better quality data to solve domain problems, for example, classification, knowledge discovery, etc.

It is important to note that even though I have developed methods for all the steps above, the methods are not necessarily associated with each other. In applications where one or more steps are relevant, the methods may be applied individually, and some steps may choose other existing techniques, or they may be skipped entirely. For example, one may use a statistical approach to detect outliers in step 2), and still applies knowledge-based strategies proposed by the thesis to analyse them to make the decision of rejecting bad outliers in step 3), provided sufficient domain knowledge can be acquired. In other applications where the distinction between "good" and "bad" outliers does not significantly influence the data analysis results, step 3) may be skipped altogether and simply rejecting or accommodating the outliers will achieve reasonable results. The methodology will be systematically discussed, and various circumstances under which this methodology can be applied in Section 2.4.

1.2.3 Applying the Approach to Real-World Problems

Visual field testing provides the eye care practitioner with essential information regarding the early detection of major blindness-causing diseases such as glaucoma. Testing of the visual field is normally performed using an expensive, specially designed instrument whose use currently is mostly restricted to eye hospitals. However, it is of great importance that visual field testing be offered to subjects at the earliest possible stage. By the time a patient has displayed overt symptoms and been referred to a hospital for an eye examination, it is possible that the visual field loss is already at an advanced stage and cannot be easily treated. To remedy the situation, personal computers (PCs) have been exploited as affordable test machines.

One of the major technical challenges is to obtain reliable test data for analysis, given that the data collected by the PCs are much noisier than those collected by the specially designed instrument. When a PC is used in public environments such as GP clinics or field investigations, and is run by amateur operators or patients themselves, noise is typically involved in the collected data. As patients' testing behaviour varies, noise in the collected data may be significant. On the other hand, outliers caused by countless unknown pathological conditions also have an equal chance of occurring, which is especially common in this new form of tests. An application such as this therefore provides an excellent platform for developing and testing the outlier management techniques proposed by the thesis.

My research initially targeted a more reliable way of detecting outliers [112, 185]. The SOM-based method allowed me to individually analyse outliers on a domain-specific basis, where I discovered that outliers could be caused by different factors [179, 110]. The research work then quickly evolved to the task of analysing and discriminating among the detected outliers for “real measurements” or “real noise.” Such discrimination was made possible by acquiring application-specific expertise about either how the noisy data points or the real measurements are distributed. The knowledge was extended from its initial incomplete set with machine learning, and was used to model the data in different ways. Depending on the types of knowledge available, the data models may be formed to explain either real measurements [114] or real noise [184] among data points. By testing the outliers against the data models, I was able to see their different characteristics, and application-sensitive data cleaning then became possible. Evaluating results from my experiments have shown immediate evidence of improvement in data quality.

It is worth emphasising that my research did not only focus on analysing outliers. I put efforts into improving the data quality in the data collection stage [29, 111], so that it reduces the chances of mishandling outliers at a later stage. I have also used the noise-eliminating method to help the discovery of new knowledge from cleaned data [28, 30, 31], which demonstrates how improved data quality is important to real world problem solving tasks. Moreover, I have implemented an integrated IDA system for screening some of the most common blinding diseases and applying such a tool to clinical data collected in different public environments [29, 107]. Understandably, data cleaning is one of the key capabilities of this tool. The thesis demonstrates how integrated intelligent data analysis can be effective and useful by applying the approach to real world applications. By receiving favourable results from its field trials under real clinical environments, the outlier analysis approach was validated in another important way.

Although my research was first motivated by this medical application, the proposed methodology is sufficiently general that they may be applied to other types of applications. I will further discuss the various applicable and non-applicable circumstances in Section 2.4.

1.3 Key Contributions

I consider that the thesis has made three key contributions to the knowledge in IDA, although other contributions will also be discussed in Chapter 9.

First, the idea of explicitly distinguishing between different types of outliers using domain knowledge is a useful extension to both statistical and computing communities for outlier management. Earlier statisticians first developed a number of statistical tests to determine whether to retain or reject outliers. They later realised that any hard assumptions and thresholds imposed would not be reasonable to all applications. Therefore, a class of "robust" approaches was proposed and explicit outlier treatment received much less attention over the past few decades [8]. In these methods, the influences of outliers are *accommodated*, so that inferences drawn from data are not seriously distorted by the presence of outliers. However, the accommodation or robust methods may not be suited for those application areas where explicit treatment (i.e. rejection or retain) of outliers is desirable (see Section 1.2.1 for examples and Section 2.1.3 for general discussion).

In this thesis, I believe that the approach of explicit outlier distinction deserves more attention than it currently receives: new computational techniques are being developed which are beginning to provide the necessary capability for advancing this important area of research. I suggest that the determination of "good" or "bad" outliers is possible if proper domain knowledge can be used and I realise that the variety of machine learning techniques have become sufficiently mature to facilitate the necessary domain knowledge acquisition. Based on the knowledge model, outlier distinction could be made automatically. This will have important implications on those data mining applications where the automation of outlier analysis could lead to important benefits. It is hoped that the work reported in this thesis will help stimulate, or perhaps more accurately, renew research in this important area.

Second, I have proposed two complementary strategies for reasoning with outliers with a view to establish their origins. The first strategy aims to identify "good" outliers by building a model of how "normal data" should be distributed in a domain of interest. The second strategy, however, tries to construct a model for capturing the behaviour of “bad” outliers and test outliers against this model. Each of these two strategies could be applied to a range of applications (see Section 2.4.2).

Third, in spite of numerous existing outlier detection methods, outliers in multivariate data are still more difficult to identify and to explain than the less structured situations such as univariate samples. The SOM-based outlier detection method that I have developed allows to visualise the data on a two-dimensional map, and outliers can be identified more easily and may be explained using the maps with specific domain meanings. These properties facilitate the understanding of multivariate data, and the participation of domain experts. The SOM is also suited for outlier detection in different applications (see Section 2.4.1).

1.4 A Brief Description of Contents

In Chapter 2, I will start by covering the background information. Research in intelligent data analysis, and in particular outlier management, will be reviewed. I will also describe backgrounds and terminologies regarding visual field testing and the associated eye diseases, in order to better understand the various application-domain specific discussions throughout this dissertation. I will end this chapter with a detailed discussion of the methodology for its properties and applicability.

Chapter 3 presents the method that I have used in step 2) of the outlier management process (as in Section 1.2.2). I will describe the use of self-organising maps (SOM), an unsupervised machine learning algorithm, to detect and analyse outliers. Its visualisation properties allow individual analysis of outliers, and thereby to discriminate them further.

Chapter 4 presents one of the outlier analysis strategies that I have used in step 3) of the outlier management process (as in Section 1.2.2). I introduce a strategy to identify “good” outliers by modelling real measurements. Such a strategy is particularly effective if there is knowledge about how real measurements should manifest themselves under various circumstances. Any surprises outside the model, therefore, can be considered as noise.

A complementary strategy for knowledge-based outlier analysis, which I have also used in step 3) of the outlier management process, is presented in Chapter 5, where the application data are in a different nature and the available knowledge is in a different form. The strategy involves the building of a "noise model," and any outliers that do not fit the model will be considered as potentially useful information, and included for further investigation.

Chapter 6 demonstrates the work of reducing noise in the data collection stage (step 1) of the outlier management process, see Section 1.2.2) in the context of collecting visual field data in clinical environments.

Reducing noise from data is only the first part towards intelligent data analysis. Its success can only be validated when such "cleaned data" are used in real-world problem-solving tasks. Chapter 7 presents such an example, where outlier analysis methodologies are used to help knowledge discovery from noisy data. It demonstrates my work in step 4) of the outlier management process.

Chapter 8 systematically describes an IDA system, which integrates the outlier analysis strategies, noise reduction methods, and the knowledge discovery techniques. To extend the validation further, I present results received from several real community investigations using the integrated IDA system. I also evaluate the system using several key software quality criteria such as *functionality*, *reliability*, and *efficiency*.

Finally in Chapter 9, the thesis is concluded by summarising the proposed outlier management methodology and the contributions made by the dissertation. Further work is also outlined in this chapter.

Chapter 2 Background

In this chapter, I will discuss both related work and the proposed methodology. A variety of intelligent data analysis techniques will be reviewed. In particular, outlier management will be summarised in various aspects including their detection methods, causes, and treatments. I will also provide background information and terminologies regarding the application for diagnosing eye diseases. I will conclude this chapter with Section 2.4 – a detailed discussion of the properties of the proposed outlier management methodology.

2.1 Outlier Management

2.1.1 Why Do Outliers Occur?

An outlying observation, or “outlier,” is one that appears to deviate markedly from other members of the sample in which it occurs [58]. Such outliers do not fit with the pattern that we have in mind, at the outset of our enquiry, of what constitutes a reasonable set of data. We have subjective doubts about the propriety of the outlying values both in relation to the specific data set that we have obtained and in relation to our initial views of an appropriate model to describe the generation of our data.

So why do outliers occur? They may have arisen for purely deterministic reasons: a reading, recording, or a calculating error in the data. The remedy for these situations is clear: the offending sample values should be removed from the sample or replaced by corrected values. In less clear-cut circumstances where we suspect, but cannot guarantee, such a tangible explanation for an outlier, no such obvious remedy is available to us, and we have no alternative but to regard the outlier as being of a random nature [8].

In [62], outliers are classified into one of four classes. First, an outlier may arise from a procedural error, such as a data entry error or a mistake in coding. These outliers should be identified in the data cleaning stage, but if overlooked, they should be eliminated or recorded as missing values.

Second, an outlier is the observation that occurs as the result of an extraordinary event, which then is an explanation for the uniqueness of the observation. In this case the researcher must decide whether the extraordinary event should be represented in the sample. If so, the outlier should be retained in the analysis; if not, it should be deleted.

Third, outliers may represent extraordinary observations for which the researcher has no explanation. Although these are the outliers most likely to be omitted, they may be retained if the researcher feels they represent a valid segment of the population.

Finally, outliers may be observations that fall within the ordinary range of values on each of the variables but are unique in their combination of values across the variables. In these situations, the researcher should be very careful in analysing why these observations are outlying. Only when specific evidence is available that discounts an outlier as a valid member of the population should it be deleted.

2.1.2 Outlier Detection

Many statistical techniques have been proposed to detect outliers and comprehensive texts on this topic are those by Hawkins [67], Barnet and Lewis [8]. Outliers can be identified from a univariate, bivariate, or multivariate perspective.

The univariate perspective for identifying outliers examines the distribution of observations and selects as outliers those cases falling at the outer ranges of the distribution. The detection method is relatively straightforward and the primary issue is to establish the threshold for designation of an outlier. For example, some would advocate the heuristic that defines a value more than three standard deviations away from the mean as an outlier.

In addition to the univariate assessment, pairs of variables can be assessed jointly through a scatterplot. Cases that fall markedly outside the range of other observations can be noted as isolated points in the scatterplot. To assist in determining the expected range of observations, an ellipse representing a specified confidence interval (varying between 50 and 90 percent of the distribution) for a bivariate normal distribution can be superimposed over the scatterplot.

This provides a graphical portrayal of the confidence limits and facilitates identification of the outliers [62].

Specific methods are needed for multivariate or highly structured data. Fortunately, a number of statistical methods have been developed for this purpose, such as those using the sub-ordering principle [9], graphical and pictorial methods [91], principal components analysis methods [68], and the application of simple test statistics [55]. Some more formal approaches provide models for the data, and test the hypotheses of certain observations being outliers against the alternative that they are part of the main body of data [67].

Apart from statistical outlier detection, there are also methods based on information theory, which assumes that outliers are most surprising and therefore have the most information gain [61]. Neural network based outlier detection was also developed [112, 130]. In addition, a variety of AI techniques have been used to help detect outliers in datasets, including Bayesian methods, rule-based systems, decision trees, and nearest neighbour classifiers [127, 169]. In applying these methods, the challenge is often to balance two things: the blind removal of outliers, which may result in an inaccurate and often too simplistic model, and an over-fitted model of high complexity, which generalises poorly to data beyond the training set (i.e. by including all the outliers).

Robust decision trees detect and remove outliers in categorical data [85]. A C4.5 decision tree [140] is used, during the training phase, to build a model from the whole data set. This is followed by a pruning phase where nodes are removed whenever this leads to a higher estimated error rate. A rule-based system for outlier detection is introduced in [138] to check the data quality of patient records in Austrian hospitals. Some other computational methods for outlier detection include [5, 22, 92, 122].

2.1.3 Methods for Handling Outliers

Although a considerable amount of work on outlier detection has been done in the statistical community, relatively little work has been done on how to decide whether outliers should be retained or rejected. One statistical approach advocates an explicit examination or testing of an outlier with a view to determining whether it should be rejected or taken as a welcome identification of unsuspected factors of practical importance [59]. However, this judgement is extremely difficult to make for statistical methods because different types of outliers often share similar characteristics from a statistical point of view [8]. This has led to conclusions by some statisticians that "statistical techniques can be used to detect suspicious values, but should not be used to determine what happens to them [4]."

In view of the difficulties with explicit examination of outliers, a majority of current statistical work adopts an alternative approach that neither rejects nor welcomes an outlier, but accommodates it [81]. This approach is characterised by the development of a variety of statistical estimation or testing procedures, which are robust against or relatively unaffected by outliers. In these procedures, outliers themselves are no longer of prime concern. This approach assumes that outliers are somehow undesirable objects and their effect or influence ought to be minimised. Unfortunately, this assumption contravenes many practical applications, because outliers can actually represent unsuspected factors of practical importance and can therefore contain valuable information. In these situations, the influence of outliers should be emphasised rather than limited or minimised [122]. Any attempt to systematically minimise the influence of outliers without due consideration to these applications can lead to loss of valuable information, often crucial for problem solving.

In order to successfully distinguish the outliers, various types of information are normally needed. These should not only include various data characteristics and the context in which the outliers occur, but also relevant domain knowledge. The procedure for analysing outliers has been experimentally shown to be subjective, depending on the above-mentioned factors [37]. The analyst is normally given the task of judging which suspicious values are obviously impossible and which, while physically possible, should be viewed with caution. For example, Matic *et al.* suggested a method that the outliers are examined manually to determine whether they should be included or discarded [122]. However in the context of data mining where a large number of cases are normally involved, the number of suspicious cases would be sizeable too, and manual analysis would become inefficient.

2.2 Diversity of IDA Methods

2.2.1 Artificial Neural Networks

The development of artificial neural networks has been inspired in part by the observation that biological learning systems are built out of a very large number of interconnected neurons. This work dates back to the very early days of computer science. In 1943 McCulloch and Pitts' model of artificial neurons [123] caused much excitement, which led to the exploration of variations of this model. In the early 1960s, Widrow and Hoff investigated perceptron networks ("Adelins") and the delta rule [175]. By the late 60's it became clear that single-layer perceptron networks had very limited capabilities [125]. Hopfield [79] analysed asymmetric networks using statistical mechanics and analogies from physics, and the Boltzmann Machine [75] tightened the link between statistical mechanics and neural network theory even further. Perhaps the most widely used artificial neural networks are back-propagation networks [23, 149, 173] and self-organising maps (SOM) [95, 96], which are powerful supervised and unsupervised learning methods, respectively.

The back-propagation network has a "teacher" who supervises the learning by providing correct output values for each input. The resultant network can then be used to map unknown input values to appropriate output values. Consider a neural network with a set of input neurons, a set of output neurons, and a set of links, via some intermediate neurons connecting the input and output neurons. The back-propagating algorithm allows a correct mapping between input and output patterns. Typically, the weights on the links are initially set to small random values. Then a set of training inputs is presented sequentially to the network. After each input has propagated through the network and an output has been produced, a "teacher" compares the value at each output with the correct values, and the weights in the network are adjusted in order to reduce the difference between the network's current output and the correct output. The back-propagation network has been used to implement applications in many domains for a variety of problems, including bioinformatics, control, speech recognition and credit scoring [6, 69].

On the other hand, Kohonen's SOM automatically model the features found in the input data and reflects these features in topological maps. The resulting maps form local neighbourhoods that act as feature classifiers on the set of input patterns in such a way that similar input patterns are mapped onto close neighbourhoods in the maps. A typical architecture of the SOM consists of two layers. The input layer is a vector of N nodes for presenting the input patterns to the network, and the output layer is typically a two-dimensional array of M output nodes for forming feature maps. Each input pattern produces one "winner node" in the output layer and similar input patterns produce geometrically close winner nodes. The applications of the SOM are widespread [86], including biological modelling [133], vector quantisation [153], and combinatorial optimisation [46]. I will provide a more detailed description of how the SOM is applied to an application in the next chapter.

2.2.2 Bayesian Networks

A Bayesian network [103, 136, 84] is a directed, acyclic graph (DAG) that encodes probabilistic relationships among variables of interest. The process of using the Bayesian network for problem-solving is to find the appropriate structure of the DAG and the conditional probability tables (CPT) associated with each node in the DAG. When the structure of the DAG is known, say from the domain expert, only CPTs need to be calculated from the given data. However, if the network structure is unknown, one has to find that structure, as well as its associated CPTs, which best fits the training data. To do so, one needs a metric to score candidate networks and a procedure to search among possible structures. A variety of score metrics and search procedures has been proposed in the literature [24, 70].

It has been argued that there are several distinct advantages of using Bayesian networks for large-scale data analysis and modelling [70]. First, since the arcs connecting the nodes in the DAG can be thought of as representing direct causal relationships, a Bayesian network can be used to learn such relationships. Second, because the model has both causal and probabilistic semantics, it is an ideal representation for combining prior knowledge and data. Third, Bayesian networks can be used both for supervised learning [82] and for unsupervised learning [26]. For these reasons, we are beginning to see more Bayesian networks in practical applications [53, 119].

In AI, work on Bayesian networks can be traced back to [137] in which "message-passing" algorithms for trees were developed. An algorithm for learning "polytrees" with unknown structure and fully observable variables is given in [136]. Early work on learning Bayesian networks was done in [39], extended by [71] for recovering the structure of general networks in the fully observable case. A deep statistical analysis of Bayesian networks is provided in [156] for the fixed structure, fully observable case. Work on missing data can be found in [41, 54, 143].

2.2.3 Extracting Rules from Data

Many real-world problem-solving tasks are classification – assigning cases to categories or classes determined by their attributes. For instance, given the categories of "football player" and "netball player," one might try to assign an individual to one of these two groups and sex might be the primary attribute in deciding the assignment.[1] Here are two ways of building a classification model, which will allow the classification of previously unseen cases.

First, the model may be built by interviewing domain experts to elicit classification rules. For example, the expert might say, "if someone is a female, then she is more likely to play netball," and "if someone is a male, then he is more likely to play football." Early expert systems or knowledge-based systems relied heavily, many exclusively, on the acquisition of relevant rules from the experts. Despite being one of the most successful sub-fields of AI for over a decade with many impressive systems developed [154, 43, 47], these systems suffered from the knowledge acquisition brittleness: when a new case falls outside the experts' considerations, these systems fail to give an appropriate answer.

Second, the classification model may be constructed inductively from numerous recorded classifications. In the football versus netball example, values of a few attributes might be available for a group of individuals with known classifications. Apart from sex, other attributes might include age, married or not, hairstyle etc. So how can one construct a

[1] The netball/football example is adopted from Professor Max Bramer's lecture notes.

classification model from these recorded classifications and attributes? C4.5 [140], a descendant of an early program by the same author, called ID3 [141], is probably the most commonly used program for inducing classification rules from a set of labelled training data. Essentially, C4.5 recursively selects an attribute by which the training set is split into non-empty subsets for each value of the attribute. During any stage of the tree building, if all cases in a training set belong to the same class, then the value of the class is returned. The key requirements for using this approach are that the classes should be discrete and pre-defined, and that there should be plenty of cases, far more than classes, expressed in terms of a fixed collection of attributes. For applications involving continuous classes, the Classification and Regression Trees (CART), developed in the statistical community, may be used [21]. See [151, 127] for overviews of different rule induction methods.

Recently there has been much work on the extraction of so-called "association rules" from databases [2]. Association rules are statements of the form "x% of customers who bought items A and B also bought the item C." Many algorithms have been invented to extract various kinds of rules from data [120, 3], and they have been found particularly useful for analysing basket data in retail applications for the purposes of cross-marketing, store layout, catalogue design and customer segmentation.

2.2.4 Evolutionary Computation

The idea that evolution could be used as an optimisation tool for engineering problems was studied in the early days of computing. For instance, Rechenberg [144] introduced "evolution strategies" and applied them to the optimisation of real-valued parameters for devices such as airfoils. Fogel *et al.* [51] developed "evolutionary programming," in which finite-state machines were used as the representation scheme for candidate solutions. Holland [78] invented "genetic algorithms," aiming to formally study the phenomenon of adaptation as it occurs in nature and to develop ways in which the mechanisms of natural adaptation might be imported into computer systems. All these approaches shared a common idea: evolving a population of candidate solutions to a given problem by using operators inspired by natural genetic variation and natural selection. These approaches form the backbone of the field of "evolutionary computation [126]."

The field of genetic algorithms has progressed a long way since Holland's pioneer work, especially in the last decade. We have witnessed an increasing number of interesting practical applications, including "genetic programming," the evolution of computer programs [99], the prediction of protein structure [152], and the prediction of dynamic systems behaviour [134]. The basic idea of genetic algorithms may be formulated as the problem of "search" – search for solutions. One starts by generating a set of candidate solutions (the initial population) for a given problem. The candidate solutions are then evaluated according to some fitness criteria. On the basis of the evaluation results one decides which candidates will be kept and which will be discarded. Further variants are produced by using appropriate operators such as crossover and mutation on the surviving candidates. After a certain number of iterations (generations), the system converges – one then hopes that the best surviving individual represents a near-optimum or reasonable solution.

Genetic algorithms can play an important role in IDA since data analysis can often be formulated as search problems – for example, a search for the next step in exploratory data analysis, a search for the most appropriate model explaining the data, a search for a particular structure, etc. In exploratory data analysis, at each stage there is often a set of possible operations that could be performed and what to do next often depends on the results obtained so far, the problem-solving context, data characteristics and the analyst's strategy. In addition, given each data set there are often a large number of possible fitting models. Genetic algorithms are a strong contender for many classes of search problems [56, 126].

2.2.5 Other IDA Methods

In the above sections, I have briefly discussed some of the advanced IDA methods, which have been under rapid development for the last decade. These methods have been applied to a wide range of practical applications. However, it should be noted that there are many other methods which have much to contribute to data analysis, including case-based reasoning [146, 98], fuzzy and rough sets [192, 7, 135], inductive logic programming [129, 142], support vector machines [167], and visualisation [40, 131]. Of course, one should not forget that a vast volume of literature on data analysis can be found in statistics and pattern recognition [42, 118, 101, 63].

2.3 Diagnosing Eye Diseases

Due to the nature of our medical applications, there is much specialised clinical or ophthalmic terminology throughout this dissertation. Because many concepts are crucial for understanding the research work in this thesis, I try to systematically but also briefly explain them in this section. Explanations for those terms marked with underline in this section can also be found in the glossary at the end of this dissertation.

2.3.1 Optic Nerve Diseases

The optic nerve contains the retinal nerve fibres at the back of the eye, which carry visual impulses from the retina to the brain. Vision would be affected when the function of the optic nerve becomes abnormal. Among the eye diseases that cause vision impairment or blindness, the cause and cure of optic nerve diseases are least known. Medical researchers have been making enormous efforts in studying optic nerve diseases. The application of my study focused on two types of the most common optic nerve diseases, namely glaucoma and optic neuritis. They are severe threats to blindness in developed and developing countries respectively, and their impacts will grow with increased life expectancy [162].

Glaucoma. It is a condition sometimes associated with high pressure in the eye, which over many years can damage the retinal nerve fibres. A badly affected person only notices what is being directly looked at, but misses objects to the sides, similar to wearing blinkers, which is particularly hazardous in driving or navigating. Established glaucomatous damage is irreversible, and some severe cases may eventually result in complete blindness. The earlier the condition is detected, the better the chance of preserving sight with treatment. Glaucoma affects approximately one in 30 people over the age of 40 and is the second-leading cause of irreversible blindness in the developed world [162].

The diagnosis of glaucoma normally requires a variety of clinical information such as disk appearance, intra-ocular pressure, and visual field. However, only visual field testing, called perimetry, provides early detection of the disease [80].

Optic neuritis. It is the most common optic nerve disease affecting young people in developing countries, particularly in Africa and Central/South America. It is typically associated with onchocerciasis, a major epidemic disease in these areas. When the WHO onchocerciasis control programme was started, it estimated that between 20 and 30 million people were infected by onchocerciasis throughout the world [174]. The average age of the first attack is 31 years, though teenagers and people over 40 might develop this disease for the first time as well. It is an inflammation in the optic nerve system, causing blurred central vision (the vision used to read and see fine details), reduced colour perception, and reduced sensation of light brightness. The exact cause of optic neuritis is still little known.

Ideally, it should be possible to identify early cases of optic neuritis, so that treatment can be given and further visual function loss may be prevented [1]. However, patients who are at risk of optic nerve disease are not easy to screen by simple traditional visual function tests such as visual acuity. Visual field testing devices, which detect early signs of the diseases more effectively, are seldom available in the epidemic regions. The provision of simple, low cost but scientifically sound technology for developing countries is a priority [162].

2.3.2 The Conventional Visual-Field Testing Procedure

Typical symptoms of optic nerve diseases include abnormality in the visual field – the entire portion of the retina in the eye where objects are visible. When the visual field becomes affected, some parts are less sensitive to external stimulus, such as light or movement. Ophthalmologists therefore, usually rely on an eye-testing procedure, namely perimetry, to diagnose optic nerve diseases [80, 105].

Perimetry systematically examines the visual field at various levels of light sensitivity [105]. The procedure involves a perimeter – a special device, and a trained specialist called a perimetrist to control and monitor the test. The perimeter presents light stimuli varying in sizes or brightness at a number of locations across the visual field. In order to examine different locations in the visual field, the perimeter has a central point (the fixation point) at which the observer has to gaze. When stimuli are presented one by one at each location, the sensitivity of the corresponding retinal areas is tested. Based on the observer's responses to the stimuli, the perimeter can quantify the distribution of visual functions in the visual field, in the form of sensitivity values. Figure 1 depicts such a conventional perimetry.

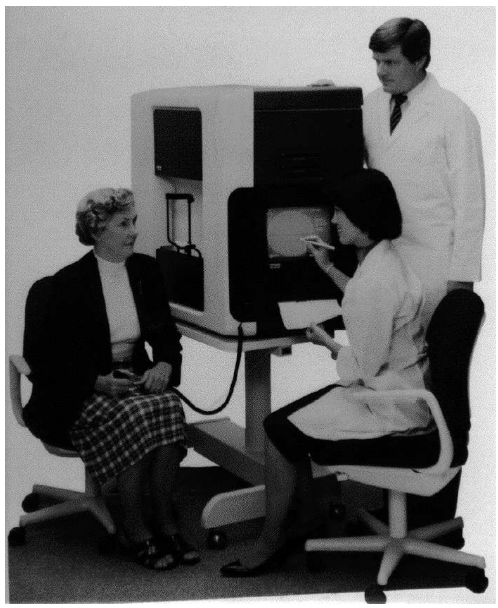

Figure 1. A conventional perimetry

A number of behavioural factors could contribute to the accuracy of the sensitivity measured from such a test, however. As a psychophysical test, it not only measures physiological status, but also deals with psychological behaviour. It is naturally subject to behavioural fluctuation problems from patients [76, 17, 88], and there have been a number of research attempts to address the issue [13, 158, 159, 124]. For example, the precision of the retinal location being examined relies on how well the observer focuses on the fixation point. A fixation loss might occur if the eyeball moves between two stimuli testing two different retina locations. A false negative response will result if the observer is distracted and fails to respond despite the stimulus having been seen. Vice versa, a false positive response can occur if the observer responds without seeing the stimulus. The perimeter offers several facilities to inspect these reliability factors and they have been used to indicate the reliability of test data [87].

- The observer's fixation is checked by presenting a stimulus in the blind spot. A fixation loss has resulted if the observer responds.
- Occasionally during the test, the perimeter presents a stimulus that is much brighter than normal, but the observer does not respond. This is a case of a false negative response.
- At other times, the projector moves as if to present a stimulus but does not do so. A false positive response has occurred if the observer responds.

Perimetry of this type has been successfully used in clinical environments to help diagnosis [72]. However, the perimeter is specially designed, expensive, and with limited availability. Its use is mostly restricted to eye hospitals [163] and it seems unlikely that the use of such devices by opticians will increase much [157]. These factors prevent them from being used for community eye screening and massive field epidemiological investigation. On the other hand, people should undergo visual field testing at the earliest possible stage. By the time a patient with overt symptoms has been referred to a hospital for eye examination, the visual loss might already be at an advanced stage and might not be easily treated [155].

2.3.3 The CCVP

As an alternative to the conventional perimetry, the Computer Controlled Video Perimetry (CCVP) was proposed as a software-based perimetry operating on a PC [182]. The CCVP is the first visual stimulation-generating program implemented on portable PCs, which has shown to be a useful measurement for detecting the progress of early glaucoma [177, 188, 180, 181, 183]. Instead of using light stimuli, the test uses motion stimuli measuring motion sensitivity [49] by displaying moving objects on screen. Because of this, it is sometimes referred as *motion sensitivity perimetry*. It examines a number of retinal locations in the visual field, with varying moving scales. The CCVP was designed to be a simple, low cost, and user-friendly system.

Figure 2 is an example of a CCVP screen layout. The test screen consists of a number of vertical bars with the same shape, colour and intensity, but different in size. The controlling program moves a vertical bar horizontally at a pre-designed moment, and expects a response from the observer. The observer, always focusing on the circle (the fixation point) at the right side of the screen, responds by pressing a key or a mouse button. Moving vertical bars in a certain area of the screen would lead to the examination of the corresponding parts of the visual field.

Note that not all the vertical bars are used in the CCVP to examine motion sensitivity. The six numbers (1 – 6) in Figure 2 are not displayed on screen. Rather, they illustrate the testing locations that the particular version of the CCVP uses. The six locations in this case – *(L1, L2, L3, L4, L5 and L6)* are considered sufficient to cover different motion functional areas on the retina [182].

By moving the bars at different scales, the sensitivity of the retinal locations may be determined in further detailed levels. At each location, there may be up to four levels of movement scales – *(S1, S2, S3, S4)*, with *S1*, *S2*, *S3* being different movement scales and *S4* being flicker. However, the more popular version of the CCVP only presents *S2* at each location.

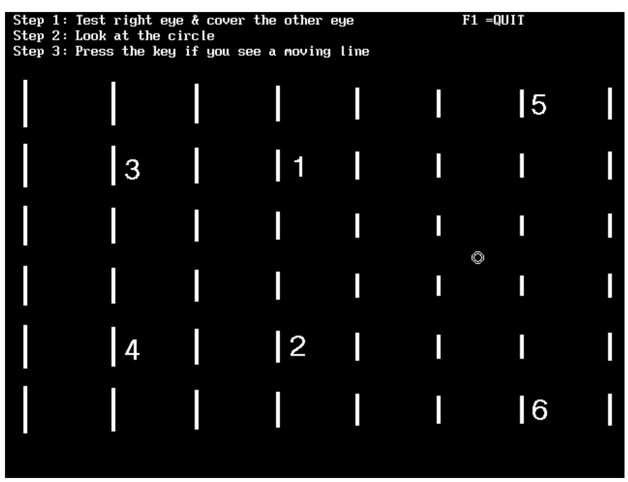

Figure 2. An example of the CCVP screen layout

Figure 3 is an illustration of how a subject self-tests with the CCVP. The subject uses his arm to support his chin in front of the screen, in order to maintain proper fixation. Like the conventional perimetry, the reliability of measurements may be affected due to the nature of any psychophysical tests. Unlike conventional perimetry, however, the CCVP does not have dedicated equipment to monitor the fixation loss or other behavioural instability during a test session.

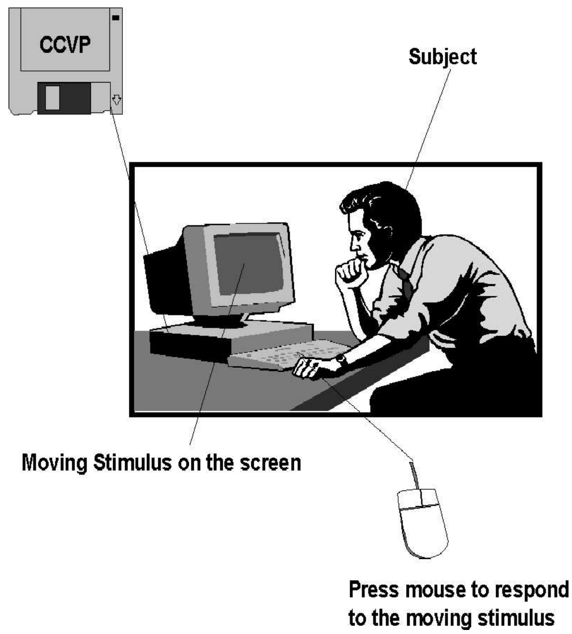

Figure 3. The CCVP test

To facilitate the collection of reliable test results, the CCVP has adopted a strategy called Single Amplitude Trial (SAT). The theory for SAT is to measure the observer's responses to the same test condition for a given number of repeating cycles, in order to reduce the effects of extraneous noise in a test [159, 160]. The SAT requires all the testing locations to be examined repeatedly with a pre-designed sequence during an entire test session. It assumes that the fluctuation is so small that the responses to the given stimulus will be constant during a test. A reliable observer's behaviour pattern should be different from an unreliable pattern detected by this strategy. To keep the same test conditions, therefore, the sequence of stimulus presentations used a look-up table rather than random order. In the table, there are two types of sequences for all stimulus locations:

P1: L1→L2→L3→L4→L5→L6

P2: L2→L4→L6→L1→L3→L5

For the test with four levels of stimuli and r repeated testing cycles, the whole sequence for the CCVP would be:

$(S4P1 \rightarrow S2P2 \rightarrow S1P1 \rightarrow S3P2) \times r$

At the end of this CCVP test, r number of data matrices v_i are produced, each of which records the observer's responses during a single cycle and has the following form. Each element x_{ijk} is either *1* with the stimulus seen, or *0* with the stimulus unseen, where i is the repeating cycle number, j is the location number, and k is the stimulus level. Equation I represents the data matrices for the particular version of the CCVP, which tests six retinal locations and four levels of movement scales, as described above.

Equation I.

$$v_i = \begin{pmatrix} x_{i11}, x_{i12}, \ldots, x_{i14} \\ x_{i21}, x_{i22}, \ldots x_{i24} \\ \ldots \\ x_{i61}, x_{i62}, \ldots, x_{i64} \end{pmatrix} \qquad (1 \le i \le r)$$

Note that the matrices become vectors if the CCVP uses only one stimulus level.

As far as each location j is concerned, there will be an average *motion sensitivity* value s_{jk} for each level of stimulus k. It is calculated by dividing the number of positive responses by the total number of repeating cycles, as in Equation II. A higher sensitivity value means that the corresponding location is more sensitive to a motion stimulus. Clinicians rely on these sensitivity values to perform diagnoses.

Equation II.

$$s_{jk} = \frac{1}{r}\sum_{i=1}^{r} x_{ijk}$$

The entire test result would be represented as a visual function pattern P with the sensitivity values calculated:

Equation III. $$P = \begin{pmatrix} s_{11}, s_{12}, \ldots, s_{14} \\ s_{21}, s_{22}, \ldots, s_{24} \\ \ldots \\ s_{61}, s_{62}, \ldots, s_{64} \end{pmatrix}$$

The response time, the time between the moment the stimulus is displayed and the moment the response is received, is also recorded by the CCVP. The response time is set to zero if the stimulus is not seen.

A few critical issues must be addressed although the CCVP has a number of advantages over the conventional perimetry. First, because the CCVP is subject to a substantial amount of noise, it is vital to improve the data quality before they can be relied on for clinical decision-making. Second, as the CCVP was a newly proposed alternative perimetry, one of the most crucial tasks would be to validate its performance, e.g., how accurately it can reflect patients' visual functions and how sensitive it is to detecting early optic nerve defects. Third, there is a need to address practical issues and to integrate the techniques into an application that can be used in real clinical environments.

In Chapter 4 to Chapter 6, I will systematically demonstrate how my noise management strategies with AI techniques have addressed the first issue. In Chapter 7, I will show that my research has also helped the CCVP to address the second issue, after test results are made more reliable. Finally in Chapter 8, I will summarise how my IDA research has contributed to the CCVP in establishing it to be a public self-screening tool for eye diseases.

2.3.4 AI in Diagnostic Visual Field Analysis

Much work has been done to apply various AI techniques to the diagnosis of visual function losses. For example, Weiss and his colleagues developed an expert system called CASNET [170, 171] for the diagnosis and treatment of glaucoma. The system was developed using a qualitative model of glaucoma and successfully evaluated by a large and varied group of ophthalmologists. The diverse opinions of consultants were utilised to establish a knowledge base detailing how various important pieces of information regarding visual field, disk appearance, intra-ocular pressure, personal information, etc. are associated with different types of glaucoma.

Among the various diagnostic information about glaucoma and other optic nerve diseases, the outcome of visual field analysis is crucial to its early detection [105, 76]. Consequently, it is not surprising that there have been a lot of AI attempts to make this analysis easier and more accurate. A majority of these attempts can be classified into two categories: expert systems and neural networks.

The expert systems approach to interpreting visual field data typically involves the use of standard knowledge of perimetry utilising the anatomy of visual pathways and the experience of individuals in analysing the data. Two such systems are Krakau [100], Boglind and Wanger [104]. Recently the use of neural network techniques to analyse visual field data has become increasingly popular. A number of networks have been trained to associate the perimetry data patterns of individual tests with various kinds of visual function losses [132, 89]. These networks are typically trained using the back-propagation technique [149] and past cases where diagnostic decisions were already made for each individual case.

One of the major problems with using the above methods to interpret the visual field data is that these data, collected from patients, commonly contain the measurement noise. As perimeters generally rely on patient feedbacks in reaction to visual stimulation, noise is usually caused by behavioural fluctuation such as learning effect, fatigue, inattention, failure of fixation, etc. It implies that the data cannot be guaranteed to accurately reflect the visual function of an individual, therefore making the clinician's diagnostic process a challenging one.

To help the clinician deal with the data reliability problem, a common strategy is to use repeated measurements, which examines locations on the retina following pre-determined repetitious sequences. The redundant information collected by repeated measurements facilitates noise identification and has the potential of improving data quality [186]. In addition, perimeters often provide facilities for inspecting various reliability factors, such as fixation losses, false positive and false negative responses, and for quantifying the data variance, such as short-term fluctuation [124]. This statistical information is expected to make clinician aware of the potential noise in the data and therefore he or she may consider these indications when interpreting the data.

Data interpretation using these indications is nonetheless a difficult one. First, although the indications offer a global measurement about the reliability of an individual test, they do not tell *when* and *where* the noise actually occurred during the test session, which is extremely useful for eliminating noise and monitoring the patient's test behaviour. Second, this kind of reliability information cannot be used to explain *why* the noise occurred, for example, whether it is due to the patient's learning effect, fatigue, anxiety, or to pathological reasons. Third, statistical methods used to obtain reliability information often involve certain statistical assumptions about the data, which may not always hold. Perhaps more important, the clinician has to either discard all the test data if too much noise is involved, or use the noisy data for analysis and diagnosis. In the latter case, even though the extent of the noise in the data is known, the clinician's interpretation is ultimately a subjective one, depending on his or her experience and knowledge.

These difficulties associated with conventional techniques have led to the seek of an alternative way to deal with the noise in the test data. In the next chapter, I will present a new computational method that enables us to show when, where, and to a large extent, why the noise occurred in the test data. The method is a combination of data visualisation techniques and domain knowledge about eye diseases and psychophysical testing behaviour. It can demonstrate patients' temporal behaviour during the test session. This has provided the first step to distinguish between behavioural instability and pathological fluctuations. With the real noise separated and filtered out, the clinician can then make more accurate diagnostic decisions based on quality test data.

2.4 The Proposed Methodology for Outlier Analysis

The objective of my study is to develop computational methods for *explicitly* managing outliers such that the methods should not only identify where the outliers are, but also carefully analyse them to distinguish between different types of outliers. For example, “good” outliers are normally informative and useful in helping with problem solving activities, while “bad” outliers are non-informative in practice, i.e. not useful to a particular problem-solving task. Here I propose a general methodology for explicitly distinguishing between different types of outliers in data:

1) Visualisation and detection of outliers in multivariate data using self-organising maps.

2) Reason about outliers using domain knowledge. In particular, two strategies have been proposed:

   i) To identify “good” outliers, namely to model how “normal data” should be distributed in a domain of interest, and to reason about outliers based on this model.

   ii) To use knowledge regarding the understanding of “bad” outliers to help reason about outliers. In other words, it attempts to model the behaviour of these outliers and accept data outside of the norms if it is not accounted for by the “bad outlier model.”

As pointed out in Section 1.2.2, not all the problems require both components although the above methodology consists of both outlier detection and outlier analysis. They can be used independently to deal with each of these situations. In certain applications, for example, outlier detection itself is the focus, while outlier analysis may be the only task when we are given particular abnormal (outlying) situations. Similarly, the outlier analysis may be applied to statistically detected outliers. The following two sections will explain how the proposed outlier detection and analysis procedures work, and their characteristics.

2.4.1 Outlier Detection by Self-Organising Maps

This method is based on the property of the SOM that similar input vectors from a multi-dimensional space are mapped onto close output nodes on a two-dimensional map. Consequently, those input data sharing similar properties are clustered on the output map. Because all data points can be visualised on the output map, the main population and outliers among input data, which are mapped onto different areas on the output map, may be identified and explicitly analysed. In addition, because each cluster on the output map is mapped from certain input data, it is possible to "explain" the output map by investigating the associated input data and assign the map with "physical meanings." We may use this information to further interpret the relationship between the main population and the outliers.

In practice, this method is capable of operating at different levels for multivariate data, detecting different types of outliers. Suppose that we have data containing information on each football player, such as goals, fouls committed, penalties conceded, time played during each game, etc. during a season. Each data vector represents the performance of a player in a particular game. We could use the SOM to visualise how well a player played each game during this season, by mapping all data vectors of this player onto the two-dimensional map. Consequently, certain outliers could be identified, e.g., a player had very poor performance in a game. This could be indicated by the phenomenon that the output node associated with the specific player data vector is outlying on the output map. Additionally, because the temporal behaviour of the individual is visualised, domain experts may be able to identify a particular pattern from each of the players. For example, a player had poor performance during the initial several games but gradually improved towards the end of the season, or vice versa. In Chapter 3, an example of how the SOM can be used for modelling a subject's visual test behaviour is given.

Similarly, the SOM could be used in the same way to model a group of individuals to study whether an individual is outlying. For example, if we have each data vector representing the overall performance of a football player over the entire season, the SOM would allow us to detecting some outstanding or less competent players in a league. Of course, one could conceive that the SOM could also be used to study whether the performance of a football team is outlying, if an appropriate data representation scheme is used.

In short, the SOM not only makes it possible to identify outlying individual instances or classes, but also allows the behavioural study for an individual class over a period of time. This serves as a useful tool for studying temporal behaviour apart from serving the purpose of detecting and visualising outlying data. Chapter 5 will show that this feature facilitates the labelling of multivariate data by domain experts.

Note that the SOM-based outlier detection method assumes that the interesting property is more stable in the data and therefore in the main population, and it exposes the outliers for further analysis. For example, the property that a utility demand (gas, water, or electricity) time series should follow certain trends (daily, weekly or monthly) is more stable than the occasional fluctuation in data caused by special events (e.g., extreme weather, major historic events etc.), or data recording errors. This method would not work if data patterns in an application are scattered around with no regularity, because there may not be a main population representing the interesting properties.

2.4.2 Two Outlier Analysis Strategies

In this thesis, two general strategies for distinguishing good and bad outliers are proposed. These two strategies are complementary in nature, and together, they cover a range of application areas.

The first strategy attempts to identify “good” outliers, i.e. to model how “normal data” should be distributed in a domain of interest, and to check outliers against this model. Chapter 4 will present such an approach based on examples provided by the domain expert, from which a model with four classes of real measurements was constructed using a supervised learning algorithm. Such a model is then applied to identify the good outliers.

This type of analysis is suited to applications where sufficient examples covering the "real measurements" or "normal patterns" are available. For example, in the hand-written character recognition experiments [122], human supervisors played the role of distinguishing good outliers (rare but valid patterns) from bad outliers (garbage patterns). However, it is possible that one could construct the model for normal writing patterns and check suspicious patterns (outliers) against the model, and the distinction process may be automated. This "normal pattern" model may be built using a sufficient number of training samples involving all kinds of valid written patterns representing a given character. Distinguishing fraudulent transactions from those looking-suspicious, but legitimate ones is another type of application where this modelling approach may be applicable. Given a large number of suspicious credit card transactions every day, there is a practical need to automate the process of outlier distinction. The corresponding "normal pattern" model may be built using a large number of legitimate transactions whether they are looking suspicious or not. In Chapter 4, I will show how this model is constructed and applied to outlier analysis in a set of glaucomatous test data.

Instead of using the knowledge from the application domain regarding how the normal data should be distributed, the second strategy for analysing outliers uses knowledge regarding the understanding of "bad" outliers or "undesirable behaviour." In other words, it attempts to model the behaviour of these outliers and accept data outside of the norms if it is not accounted for by the bad outlier model.

This type of analysis is particularly useful to applications where sufficient examples covering the "bad" outliers or "abnormal patterns" are available. For example, we could try to use this modelling approach to distinguish causes for abnormal multivariate time series observations collected from a plant as mentioned above. There may be plenty of time series patterns associated with a limited number of frequently occurred faults in the plant, which could be used to construct the bad outlier model. Chapter 5 presents another case where this type of model is constructed using an inductive learning method, and the model was applied to the analysis of the CCVP data.

It should be noted that the use of domain knowledge to "explain" outliers can be achieved by different means, depending largely on the availability of the relevant knowledge, how the knowledge is acquired and which form it takes. For example, if the knowledge is in the form of a set of rules acquired from domain experts, then reasoning techniques associated with knowledge-based systems may be used. On the other hand, if the knowledge is a set of examples provided by an expert, some machine learning methods may be used to extract the structure of the data and form useful knowledge. The vast literature on knowledge-based systems and machine learning plays the role of providing methods, guidance, and solutions for the construction of such kind of knowledge-based models.

As to when to apply each strategy, it often depends on the availability of relevant domain knowledge. For example, both strategies could be in principle applied to the case of credit card fraud transaction. However, it appears that the approach of modelling "normal patterns" may be more realistic here since fraudulent behaviour is often changing fast: the offenders often come up with new methods to beat the system. In this thesis, the bad outlier modelling method has been found more suitable for making outlier distinction within the CCVP data collected from a GP clinic for the following two reasons. First, the normal pattern modelling approach would have been difficult, given one would need to have to elicit domain knowledge regarding how different eye diseases manifest themselves on the visual field data, because of the large number of possible eye diseases with each of the GP patients. Second, the knowledge regarding the bad outliers can be acquired more easily because the limited number of typical behavioural fluctuations, such as fatigue, learning effect, and inattention are found to cover large portion of the bad outliers (see Chapter 5).

2.4.3 Remarks

The proposed methodology and its associated outlier detection and outlier analysis strategies have been discussed in the above sections, together with an assessment of their applicability. This methodology needs to be applied to real-world applications in order to demonstrate how effectively it manages outliers. This thesis has chosen the CCVP data in the context of screening for blinding diseases for the study of outliers. In this type of applications, outliers often occur in the collected data with different origins, and appropriate analysis is important for preventing the elimination of those outliers carrying useful clinical information. I will show that the “bad” outliers tend to appear in the form of measurement noise as opposed to “good” outliers, which often represent real measurements of the visual functions. The next six chapters will be devoted to the application of the methodology to this type of challenging medical applications involving the screening of different types of disease such as glaucoma and optic neuritis using data collected in a variety of environments such as GP clinics and large-scale field investigations.

Chapter 3 Outlier Visualisation and Detection

In this chapter, I shall present my work on outlier detection and visualisation. Using the CCVP as an example application, an approach of detecting and analysing outliers by self-organising maps [95] is discussed. I shall demonstrate that the self-organising maps enable the analysis of various test behavioural factors, which are the primary cause of measurement noise in this particular application. I will show why outliers can be interesting features or measurement noise caused by random, undesirable factors. I will also demonstrate how measurement noise may be separated from other outliers by analysing its possible causes.

3.1 Self-Organising Maps (SOM)

To overcome the difficulties associated with conventional techniques, we are essentially looking for a way of understanding when, where, and how outliers occur in a data set. In this section, I introduce a computational method for this purpose. In particular, Kohonen's self-organising network [95, 96, 97] is used to model noisy data. Transition trajectories on the maps are utilised to provide graphical information about the nature of the outliers. This visualisation of outlier patterns establishes a foundation for identifying measurement noise among outliers in the data.

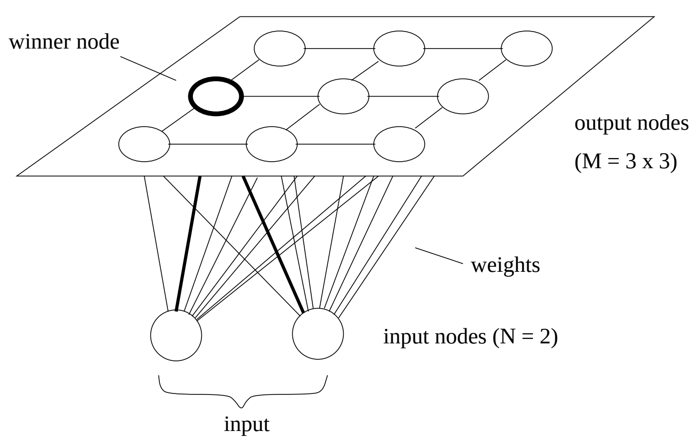

Figure 4. A self-organising network

A self-organising network can be defined with a set of input nodes $V = \{v_1, v_2, \ldots, v_N\}$, a set of output nodes $C = \{c_1, c_2, \ldots, c_M\}$, a set of weight parameters $W = \{w_{11}, w_{12}, \ldots, w_{ij}, \ldots, w_{NM}\}$ $(1 \leq i \leq$

N, $1 \leq j \leq M$, $0 \leq w_{ij} \leq 1$), and a map topology that defines the distances between any given two output nodes. The input nodes represent an N-dimensional vector and the output nodes are usually arranged in a 2-dimensional array to form a “map.” The way in which output nodes are interconnected is called “map topology.” Common map topologies include hexagon lattice and rectangular lattice. One illustrative example of a self-organising network is shown in Figure 4 where N=2, M=9 and the map topology is rectangular lattice. Each input node is fully connected to every output node via a variable connection, and hence there are $N \times M$ connections. A weight parameter is associated with each of these connections. For a given SOM, an input vector v associates with one *winner node* c_j on the output map. A winner node is an output node that has minimum distance to the input vector.

Equation IV. $$\forall i, \left\| v - c_j \right\| \leq \left\| v - c_i \right\|$$

The distance between an input vector and an output node is defined as

Equation V. $$\left\| v - c_j \right\| = \sqrt{\sum_{k=1}^{N} \left(v - w_{jk} \right)^2}$$

An SOM is established by “learning” or “self-organising” on a given set of input N-dimensional data. The self-organising algorithm calibrates the weight parameters W initialised by a set of given values. The calibration is done by reiteratively applying input samples cyclically or in a randomly permuted order. For each iteration t, the weight parameters are updated as

Equation VI. $$\frac{dw_i}{dt} = \alpha(t)\gamma(i,t)(v - w_i) \qquad (1 \leq i \leq N \cdot M)$$

The α is a monotonic decay function with t (Equation VII), and γ is a “neighbourhood function,” which defines how weight parameters associated with neighbouring output nodes are affected. Two popular types of the neighbourhood function are “bubble adaptation” [95] or a Gaussian-type function [148] (Equation VIII). Here a, b, c and k are some positive constants, and $\left\| c_i - c_j \right\|$ is the distance between the learning node c_i and the current winner node c_j on the output map.

Equation VII. $$\frac{d\alpha(t)}{dt} = -a \cdot e^{-bt}$$

Equation VIII. $$\gamma(i,t) = c \cdot e^{-\frac{\|c_i - c_j\|^2}{k\alpha^2(t)}}$$

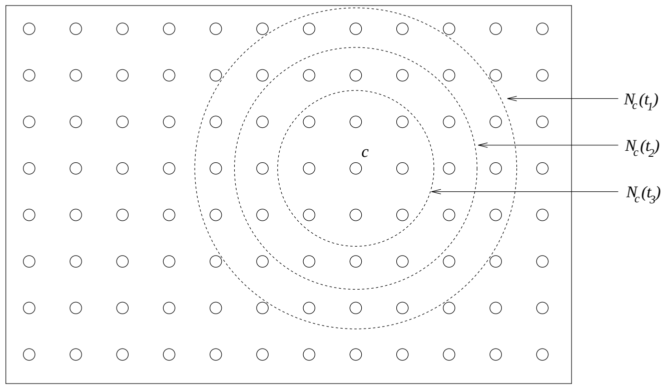

Figure 5. Three neighbourhoods around the winner node *c*

The neighbourhood set N_c is a set of nodes that surround the winner node *c*. These nodes in N_c will be affected with weight modifications, apart from those changes made to the winner node, as defined in Equation VI. These weight changes are made to increase the matching between the nodes in N_c and the corresponding input vectors. As the update of weight parameters proceeds, the size of the neighbourhood set is slowly decreased to a predefined limit, for instance, a single node. Figure 5 shows three different neighbourhood sets around the winner node *c* at different time steps in the algorithm where $t_1 < t_2 < t_3$. In the discussion above, the apparent assumption is that the distance between two output nodes is measured by Euclidean distance.

This process leads to several important properties [96]. First, similar inputs are mapped onto closely neighboured nodes. Second, each particular node represents one special input cluster. Another important property of the maps is their vector quantisation ability, i.e. their ability to project high dimensional input data onto a two-dimensional map. This makes visualisation of complex data possible, e.g., speech data [94] or symbolic descriptions of objects [147].

Perhaps the most significant property in many applications is that similar input vectors are mapped to geometrically close winner nodes on the output map. This is called *neighbourhood* *preservations*, which has turned out to be very useful in visualising and analysing outliers, as we shall see below.

As the feature of neighbourhood preservations is the essential concern about the quality of the SOM produced, I have used the *topographical product* [11] as a measurement for this purpose.

Equation IX. $$TP = \frac{1}{2M(M-1)} \sum_{j=1}^{M} \sum_{k=1}^{M-1} \frac{1}{k} \log\left(\prod_{l=1}^{k} R_1(j,l) R_2(j,l) \right)$$

Where M represents the number of output nodes and R_1 and R_2 can be calculated as follows:

Equation X. $$R_1(j,k) = \frac{d^V\left(w_j, w_{n_k^A(j)}\right)}{d^V\left(w_j, w_{n_k^V(j)}\right)}$$

Equation XI. $$R_2(j,k) = \frac{d^A\left(j, n_k^A(j)\right)}{d^A\left(j, n_k^V(j)\right)}$$

$n_k^A(j)$ denotes the *k*th nearest neighbour of j in the output space and $n_k^V(j)$ denotes the *k*th nearest neighbour of j in the input space. $d^V\left(w_j, w_{n_k^A(j)}\right)$ is the distances measured in the input space between w_j and $w_{n_k^A(j)}$ and $d^A\left(j, n_k^A(j)\right)$ is the distances measured in the output space between j and *k*th nearest neighbour of j, $n_k^A(j)$.

The topographical product is a magnitude indication of neighbourhood violation. The smaller its value is, the better the neighbourhood preservation would become [11].

3.2 Visualising and Identifying Outliers by the SOM

In this section, I shall describe how the SOM is used to analyse temporal behaviour by visualising the test data over time. I have experimented the method with the CCVP test data collected from patients in a hospital and in massive field investigations [185, 179]. The results have much interested both AI researchers and ophthalmic clinicians, and a great deal of knowledge about psychological testing behaviour has been acquired [110].

As discussed in Chapter 2, the CCVP tests each patient with a fixed number of repeating cycles during a single test. For each cycle, one or more types of stimuli are displayed following a pre-determined order at a number of designated locations on the retina. It would therefore be natural to view the testing process as a patient going through a fixed number of identical test cycles. It would also be natural to assume that the responses from one test cycle should be similar to another. This assumption, however, is often violated by behavioural factors such as learning effects, fatigue, etc., which lead to significant deviations among patients' responses. The redundant information from the repeating cycles, therefore, would help identify the outliers involved in the data.

The proposed method consists of three steps. First, Kohonen's learning technique is used to find a network capable of generating maps that reflect the patient's test behaviour. In particular, each response pattern for each repeating cycle is used as an input vector to the SOM and a corresponding winner node is produced on the output map. As far as each test case is concerned, the number of winner nodes on the output map would equal the number of repeating cycles conducted for a particular individual. Paths connecting these winner nodes in the order of the repeating cycles would thus constitute a *transition trajectory*, which graphically illustrates on a 2-dimentional map how the individual's behaviour changed from one cycle to the other.

Second, an effort is made to find a network that shows better *neighbourhood preservations*, i.e. similar input vectors are mapped onto identical or closely neighbouring nodes on the output map. This step is important as we want to map similar response patterns from patients onto similar nodes. As pointed out previously, I have used the topographical product as a measurement for this purpose.

Finally, the behaviour maps for all test cases are generated using the network produced in the second step, which is then analysed to help identify the measurement noise. They are also used to determine the origins of the noise, for example, whether it is because of learning effects, fatigue, or other factors.

Kohonen's SOM applied to the test data now follows:

Definitions: Let the number of nodes be M. Let W be a set of connection weight vectors where each output node c_j $(1 \leq j \leq M)$ is associated with a connection weight vector of N-dimensional of the form $w_j = (w_{j1}, \ldots, w_{jN})$ where w_{jk} represents the connection weight between the input node k and the output node c_j.

1) *Initialise the map topology and size of* the *output map.* I have experimented with map topology of both hexagon and rectangular lattice, and results were similar.

2) *Initialise the weights parameters.* Initialise the connection weight parameters to random values over the interval [*0.0*, *1.0*], and normalise both the input vectors and the connection weight parameters.

3) *Present new input*. Set i to $i+1$ and present input vector v_i.

4) *Select minimum distance*. The distance between the input vector v_i and each output node c_j is computed by Equation V. Designate the winner node with minimum distance to be c.

5) *Update weight parameters*. Adjust the weight parameters of the winner node c and its neighbouring nodes c_j, as specified in Equation VI. As for the neighbourhood function, I have found that a Gaussian-type neighbourhood function gives much better neighbourhood preservations than a neighbourhood iteration set (i.e. the "bubble" adaptation). This finding is consistent with some other studies [117, 44].

6) *Repeat by going to step 3).* This iterative process continues until the given iteration number is reached.

7) *Calculate the topographic product of the resultant network (see Equation IX).*

8) *Find a well-performed network.* If the neighbourhood preservation of the present network is not satisfactory as indicated by the current topographic product, go to step 2). Otherwise,

9) *Use the network to produce one map for each individual test with r repeating cycles.* There is one winner node corresponding to each cycle v_i (see Equation I). All r winner nodes constitute a transition trajectory, which shows how the individual behaved from cycle to cycle.

The resultant map is then used to analyse outliers and identify the measurement noise among them. Individual trajectories and their associated maps are classified into a number of categories, e.g., reliable, absolutely unreliable, unreliable due to fatigue, unreliable due to learning curve, etc.

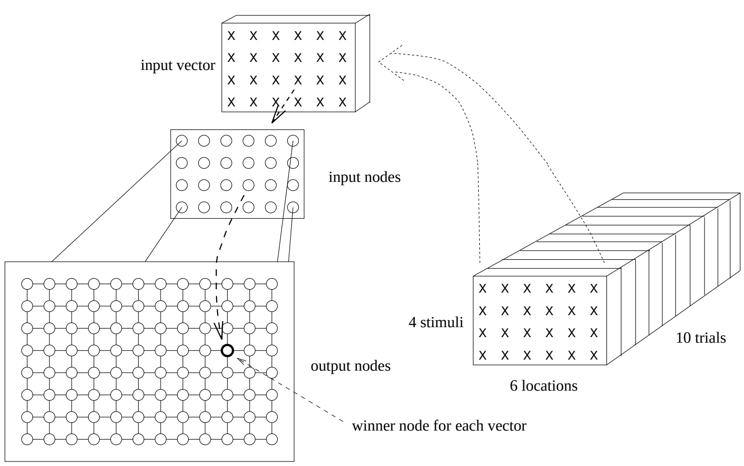

Figure 6. Applying the SOM to the CCVP

Figure 6 illustrates the application of the SOM to a particular type of the CCVP test, where there are 4 types of stimuli on 6 locations with 10 test cycles. The basic idea is that the data set corresponding to each cycle is used as an input vector. Since each input vector produces a winner node, the 10 input vectors for each CCVP test would produce 10 winner nodes. They constitute a transition trajectory in the map as shown in Figure 7.

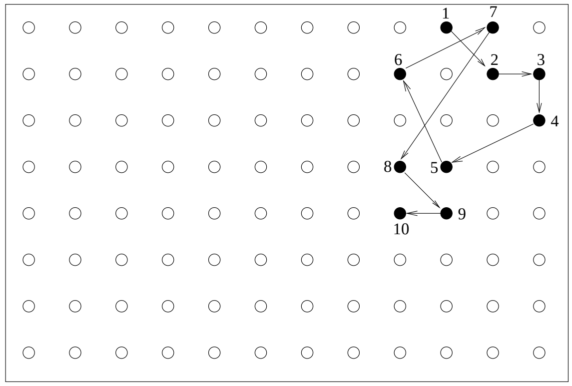

Figure 7. A transition trajectory in the output map

Several observations can be made with this method. The input vectors are quantified to 0 and 1 from binary attributes of either "seen" or "unseen," respectively. As Euclidean distance is used to elect the winner node among all output nodes, the values representing the binary attributes become insignificant once they are normalised [96]. Nevertheless, this assumes that all elements in input vectors are equally weighted when calculating the similarity between two input vectors. In other words, the attributes of any testing locations are equally regarded, without considering any correlation among them. This is generally an inaccurate assumption, as I will discover in Chapter 7. In my experiment, however, I have found that the spatial order of the SOM has captured the internal relationships without *a priori* knowledge about correlation in attributes. In an example demonstrated in [96], a similar result was also obtained without respecting inherent correlation among attributes.

In addition, I have found that the frequencies of input vectors, in addition to their similarities in between, are also reflected by the SOM. For example, the most frequently occurring vector, $(1, 1, 1, 1, 1, 1)^t$, is usually mapped to one of the four corner nodes. Its nearest neighbouring vectors, such as $(1, 0, 1, 1, 1, 1)^t$, are usually projected onto output nodes that are some distance away. The output nodes in between do not respond to any existing input vectors. This indicates that the SOM has learned from the input samples, and that the difference between the "all-seen" vectors and any other vectors is of great importance. Furthermore, it seems that the SOM has also learned the differences among testing locations. My study of disease pattern discovery from the SOM has provided further evidence for this feature, as described in Chapter 7.

Those findings show that the SOM is able to produce non-linear mapping based on statistical characteristics among data. I consider these features important in my domain-specific outlier management (see also discussions in Section 2.1.3).

3.3 Analysing Outliers and Noise Patterns

However, what does the map really tell us apart from the winner nodes and the transition trajectory? As mentioned previously, one of the most important features of the SOM is that similar input vectors lead to close winner nodes. So at least we know that different output nodes will have somewhat different input vectors associated with them. In other words, output nodes represent characteristics of different groups of input vectors, as a result of the learning process from a particular data set. Furthermore, there can be various "physical meanings" of the map by looking into output nodes in different aspects. In the case of perimetry testing, we are primarily interested in the information on visual functionality, i.e. how well the patient can see the stimuli on the screen.

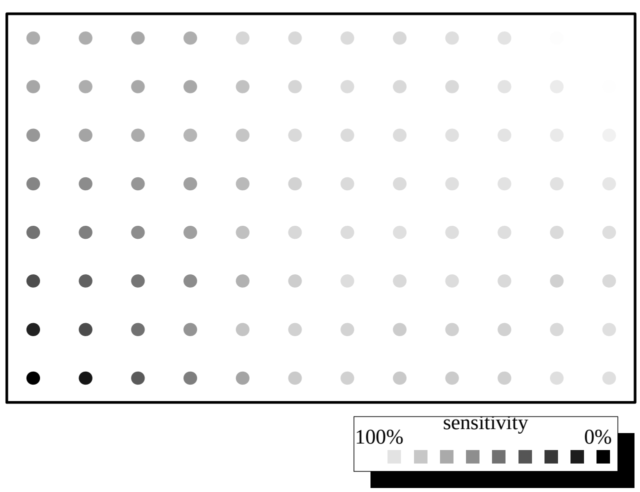

Figure 8. One of the physical meanings of the map – the average motion sensitivity

Figure 8 is an example of one of the physical meanings of the map, which is produced by calculating the average motion sensitivity (Equation II) of *all* the input vectors associated with each output node. I have found that the average motion sensitivity calculated this way is very close to the "alternative average motion sensitivity" calculated from the weight parameters w_{ij} associated with the output node (see Equation XII, where N is the number of input dimensions, and w_{ij} are weight parameters). This indicates that the output node, characterised by its weight parameters, can be seen as a "representative" vector of all the associated input vectors. Further descriptions of this phenomenon are provided in Chapter 7.

Equation XII. $$p_j = \frac{1}{N}\sum_{i=1}^{N} w_{ij}$$

Two observations can be made from the "physical meanings map." First, the top right region is most sensitive, i.e. patterns from patients who can see most of the time are mapped into this region. This is indicated by the most faded grey level of the nodes on the map. An extreme case is the node on the top right corner of the map (white), reflecting the pattern of seeing all the time, i.e. the corresponding input is a vector of 24 elements of all *1*s. Second, the motion sensitivity gradually decreases towards the bottom left region, indicated by darker grey levels. An extreme case is the node on the bottom left corner (black), reflecting patterns from patients who cannot see at all.

These two observations are largely due to the characteristics of the SOM: similar input patterns would be mapped onto typographically close nodes on the output map. For example, an input vector of all *1*s and that of all *0*s, when mapped onto the map, has the largest geometrical distance between them.

Having obtained and understood the map, the interpretation of the transition trajectory becomes the most important task. Does this trajectory tell us that the patient has performed a reliable test? If this is the case, most parts of the trajectory should be within a small neighbourhood on the map as one would expect the patient's behaviour to be similar from one test cycle to another. Or does this trajectory tell us that it is a totally unreliable test? In this case, the winner nodes might be very irregular, and the direction by which the trajectory travels is diverse and unpredictable. Or does this trajectory only tell us that this test is in between the above two cases? But then can we find out which portions of the test are reliable and which are not? And can we figure out the reasons for the unreliable portion? For example, is it due to fatigue, learning, or something else?

In principle, there are several different ways the interpretation task might be achieved. First, one can employ some basic algorithms using the geometry positions of the nodes and their relative distances to answer some of those questions raised above. For example, if the algorithm finds that the majority of the winner nodes are centred around a particular region, other winner nodes would then constitute the outliers, such as nodes 9 and 10 in Figure 9. The outlier winner nodes can then be further examined using domain-specific knowledge to determine their possible causes.

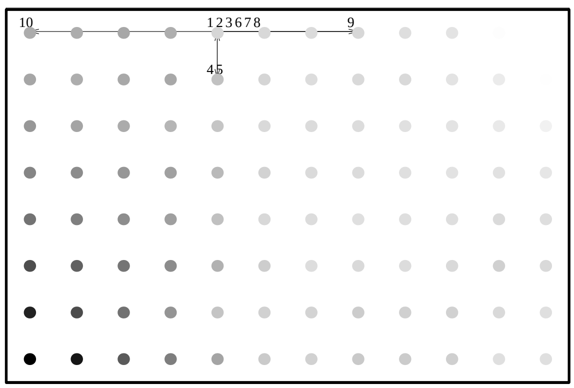

Figure 9. An example of reliable test with little noise

Second, one can also ask the expert to interpret those trajectories using the physical meanings of the maps and fundamental knowledge of the perimetry test. For example in Figure 10, the trajectory shows that the winner nodes of the first six test cycles are all on the same node with a high sensitivity value, which indicates the patient could see most of the time. However, the winner nodes of the next four cycles moved away from the first nodes towards lower sensitivity regions, which indicates that the patient could not see as clearly as earlier in the test. This appears to be a case of fatigue.

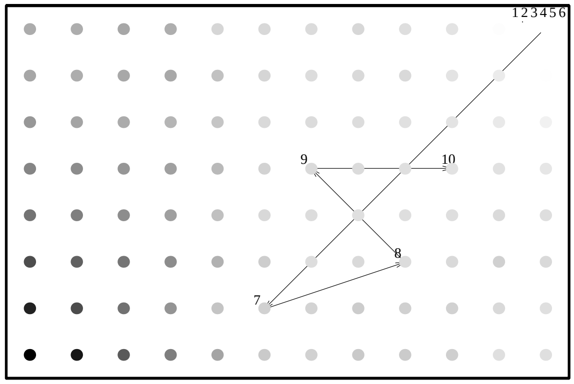

Figure 10. An example of fatigue

3.4 Experimental Results

In this section, the experimental results are presented, applying the proposed method to a set of clinical test data from the Moorfields Eye Hospital, London. In particular, the transition trajectories for 270 test records were produced and then analysed to identify the noisy data. All these transition trajectories were also used to determine which categories each individual test belongs to, e.g., whether the test is reliable (little measurement noise is involved), or unreliable (significant amount of noise is found). Among the unreliable ones, the trajectories and their associated maps are used to further distinguish between different causes, e.g., whether there is noise because of the learning effects, fatigue, or others. Below, I shall elaborate on the classifications of the psychophysical behaviour, bearing in mind that almost all the tests involve a certain amount of noise.

Reliable: A test is reliable if there is little measurement noise involved. In this case, the test results should be similar from cycle to cycle with perhaps few exceptions. This implies that an overwhelming majority of winner nodes should be centred around a small neighbourhood on the map. One of these examples is shown in Figure 9 where the first eight winner nodes are gathered on two nodes on the map.

Totally Unreliable: A test is totally unreliable if the motion sensitivity varies a great deal from one cycle to the other and there are no regularities in the variations. In this case, the winner nodes might be irregular, and the direction by which the trajectory travels is diverse and unpredictable. Figure 11 shows one example of such a case.

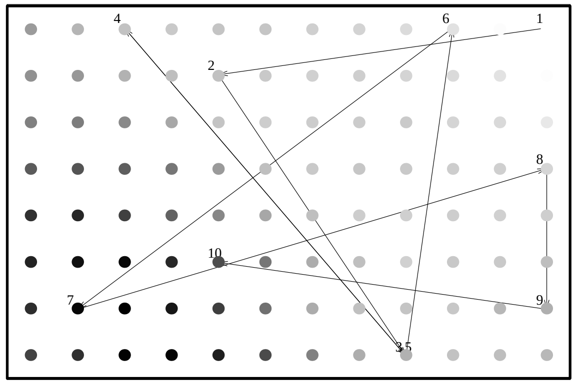

Figure 11. An example of an unreliable test

Fatigue: A test is unreliable because of fatigue if the sensitivity values of the initial several test cycles are high and similar to each other while the sensitivity values of the remaining cycles decrease over time. In this case, the winner nodes of the initial cycles tend to be in a small neighbourhood and the winner nodes of the subsequent cycles tend to move away from the small neighbourhood to areas where the sensitivity values are lower. One example of such a case has already been given in Figure 10.

Learning: A test is unreliable because of the learning effect if the sensitivity values of the initial few cycles do not show much regularity, but the sensitivity values of the later cycles gradually become similar. In this case, the winner nodes of the initial cycles perhaps are irregular, but are gradually gathered around a small neighbourhood on the map. One example of such a case can be seen in Figure 12.

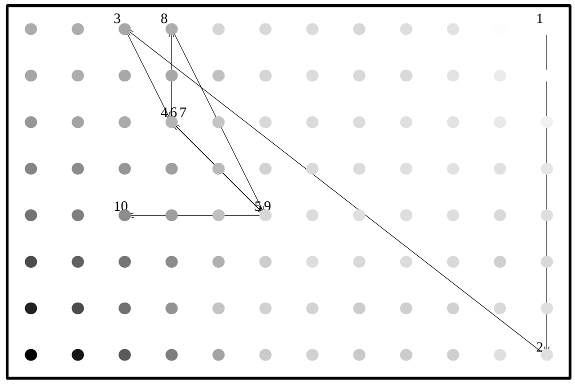

Figure 12. An example of learning effect

Others: This category includes those maps that the expert could not easily interpret. The outliers might be caused by multiple behavioural factors in a single test, and there are many other reasons for the outliers to occur, for example, inattention, failure of fixation, and pathological reasons.

Inattention: It is probably the most common reason noise occurred, whether there is little or a lot of noise in the test. That is, inattention has probably contributed, to various degrees, to all the categories of test behaviour classified above. Although a typical example is given in Figure 13 (where the 9th cycle is clearly caused by inattention), inattention could happen in almost all other situations, e.g., the 10th cycle in Figure 9 and the 7th cycle in Figure 10. Since it is likely that the phenomenon of inattention occurs in most test cases, I have not classified it into an independent behavioural category.

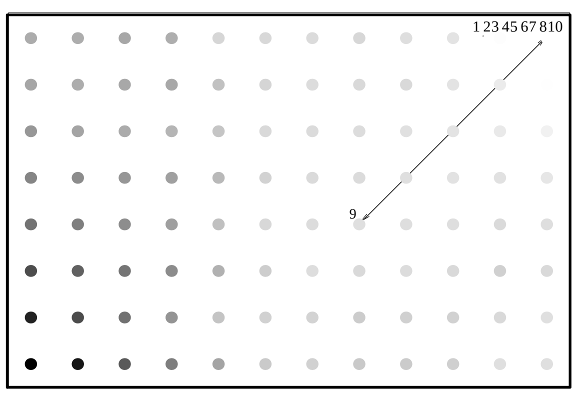

Figure 13. An example of inattention

With the CCVP it is difficult to monitor and assess fixation, i.e. whether a patient is staring at the fixation point on the test screen. If a patient can see in a certain area on the retina during one cycle, but cannot see any stimuli in the same area during the next cycle, there is good reason to suspect that the patient failed to focus on the fixation point. Because the CCVP only examines six locations spread over the visual field, it is difficult to identify fixation loss due to smaller eyeball movements. However, fixation loss has fewer significant effects for the CCVP because motion sensitivity is supposed to be associated with bigger areas instead of precise locations on the retina [182].

False Noise: It is important to note that "false noise" can occur among outliers, because of a certain pathological condition of the patient. Even if the patient has fully concentrated during the test, there might still be outliers due to short-term pathological fluctuations. These outliers cannot be treated as noise, but should be included in diagnostic decision-making. In the absence of further deterministic analysis, they were classified into the reliable category if there is no significant measurement noise involved; other more difficult cases were classified into the category of "others."

The results of the above analysis are summarised in Table 1. Note that noise is involved in virtually every individual test, although its magnitude may vary from one to another. In addition, clinicians are provided with another useful piece of information – the trajectory and its associated map. They demonstrate the test behaviour graphically and have great potential in the clinical decision making process.

Behaviour	***Number of cases***	***Percentage***
Reliable	145	53.7%
Totally unreliable	55	20.4%
Fatigue	28	10.4%
Learning	9	3.3%
Others	33	12.2%
Total	270	100%

Table 1. Behaviour classifications

3.5 Concluding Remarks

In this chapter, I introduce an outlier detection method using the SOM. Based on its neighbourhood preservation feature, the SOM is capable of mapping high dimensional data onto a two-dimensional map where the data points can be visualised and the outliers can be identified. It can be used in general to detect outliers in multivariate data where learning samples are sufficient.

Using this outlier detection method, we can explicitly analyse outlying data and identify measurement noise among them. The method is used to model the temporal behaviour and to demonstrate when, where, and to a large extent, why noise occurs in data. The results from this study suggest that the explicit treatment of outliers in data is promising. As briefly mentioned before, unstable tests may be caused by either pathological or behavioural factors of the individual test. The method proposed in this chapter has provided the foundation for identifying and analysing outliers in the data, collected from the CCVP. However, it is obvious that data analysis alone, regardless how advanced, is not sufficient for this task. In the next two chapters I will present my research about the distinctions between the real noise caused by fatigue, inattention, learning etc. and the false noise caused by pathological factors. We shall see that it is crucial to apply domain-specific knowledge to the identified outliers.

Chapter 4 Outlier Discrimination by Modelling Real Measurements

In the previous chapter, I suggested how multivariate data could be visualised and analysed in detail with the SOM. Not only has it enabled to detect outliers, but it also allows to tell how and where outliers are. In the context of analysing the CCVP data, I have noted that not all outliers are necessarily measurement noise. In this chapter, I will propose a strategy to explicitly identify outliers and then carefully analyse them to distinguish noisy outliers from noise-free outliers. I will then apply this strategy to a large quantity of visual testing data. Finally I will evaluate this strategy, which shows that it provides a satisfactory way of locating and rejecting noise in the test data.

4.1 A Two-Step Strategy for Outlier Discrimination

One fundamental assumption made when a new self-organising neural network was proposed [14] is that interesting properties in data are more stable than the noise [128]. Consider the example of the visual function test for a normal person who does not have visual function loss. The property that the person is able to see the stimuli on the screen most of the time would be more stable than the occasional fluctuation in data caused by errors for whatever reasons. I have adopted this assumption as the basic principle for identifying *real* measurement noise, to which I shall refer as the *noise identification principle*.

Suppose that repeated measurements are designed where the same observation is made over a certain number of cycles, and consider the CCVP as an example. A normal person might be distracted in the middle of a test, for instance, the fifth cycle. This would result in poorer motion sensitivity values within the visual field, leading to fluctuation in the data. This type of fluctuation, however, should not affect the overall results of the visual function, since the person is able to see the stimuli most of the other times during the test. The main task here is to identify the common feature exhibited by most of the times, i.e. the person can see the stimuli most of the time. The part of the data inconsistent with this feature, i.e. the fifth measurement, will then be exposed and consequently suspected as noise.

As discussed in Chapter 3, I consider unsupervised learning algorithms to be the natural candidate to find a computational method capable of detecting common features among data. They are known to be capable of extracting meaningful features, which reflect the inherent relationships between different parts of the data [48]. For example, we can use an unsupervised learning algorithm such as the SOM [95] to let the data self-organise in such a way that the more stable part of the data is clustered to reflect certain interesting features. Those parts of data that are inconsistent with those features will be separated from the stable cluster, and are considered outliers.

It should be emphasised that the outliers should not necessarily be the measurement noise – they can actually be the real measurements reflecting the real values of an attribute. In the example of diagnosing glaucoma using visual field data, the fluctuation in the data can be caused by behavioural factors such as fatigue and inattention, but can also be caused by the pathological condition of the patient. Consider a glaucoma patient undergoing a visual field test. It is quite possible that there will still be fluctuations in the responses at certain testing locations, even if the patient has fully concentrated during the test. The nature of the disease has dictated the responses. The elimination of these responses would lead to the loss of much useful diagnostic information, and worse still, could lead to incorrect conclusions about the patient's pathological status.

Therefore it is necessary to check whether the outliers are indeed the measurement noise. Nevertheless, this task is difficult to fulfil using the data alone, since there are often many possible explanations for fluctuation in the same data set, as discussed above. The use of appropriate domain-specific knowledge, however, has potential in resolving this difficulty. For example, the knowledge of how diseases such as glaucoma manifest themselves on the test data is crucial for identifying the measurement noise. With this type of knowledge, we can then have a better chance of finding out the component among the outliers which is caused by pathological factors.

The above discussions lead to a general strategy for identifying the measurement noise in data, which consists of two steps. First, an unsupervised learning algorithm is used to cluster the more stable part of the data. This algorithm should be able to detect some common features among those data. Those outliers, which are inconsistent with those features, then become suspects of measurement noise.

Second, knowledge in application domains, together with knowledge about the relationships among data, is used to check whether the outliers are indeed the measurement noise. This type of domain-specific knowledge, though it may be acquired from experts, is often incomplete. For example, only a partial understanding has been obtained about how diseases like glaucoma manifest themselves on any visual field test data [182]. Therefore, it is often desirable to apply machine learning methods to the initially incomplete knowledge in order to generalise over unknown situations. One such example is shown in the next section.

4.2 Outlier Discrimination

4.2.1 Outlier Identification

As discussed in Chapter 3, outliers may be detected by visualising and identifying the more stable part of the data; outliers can then be exposed as a result. The method for identifying the more stable part of the CCVP data is to model the patient's test behaviour using the SOM. Data clusters are then visualised or calculated. This method consists of three steps. First, Kohonen's learning technique [95] is used to train a network capable of generating maps, which reflect the patient's test behaviour. Second, an effort is made to find a network that shows better neighbourhood preservations, i.e. similar input patterns are mapped onto identical or closely neighbouring nodes on the output map. Having obtained a well-performed network, the final step is to generate the behaviour maps for individual patients and analyse these maps to identify the more stable part of the data. As described in Chapter 3, these maps graphically illustrate how the patient's behaviour changed from one test cycle to another (Figure 7).

As one of the key SOM features is that similar input vectors would lead to similar winner nodes, here we have the general rule for identifying the more stable part of the data: if most of the winner nodes are centred around one particular region, then the input data vectors associated with these nodes constitute the more stable part of the data. These vectors share one common feature: they are similar to each other, judged to a large extent by a distance measurement such as the geometry distance on the output maps.

The above rule can be implemented by algorithms using the geometry positions of the nodes and their relative distances. The approach taken here is to search for a maximum set of nodes on the output map that occupies the smallest topographical area. In particular, an evaluation function is defined in Equation XIII for this purpose and the objective is to find a subset of winner nodes, C_S, which minimises the value of $F(C_S)$.

Equation XIII. $$F(C_S) = \frac{A(C_S(k))}{k^2} \qquad \left(k = r, r-1, \ldots, \lfloor r/2+1 \rfloor\right)$$

Where r is the total number of winner nodes, A denotes the topographical area in the map occupied by a subset of winner nodes, and $C_S(k)$ represents a subset of winner nodes with k members.

4.2.2 Modelling the Real Measurements

Let us now examine the outliers, for example, the data vectors associated with winner nodes 9 and 10 in Figure 9, and see whether some of these vectors are the measurement noise. For this particular application, it is especially interesting to find out whether data points among outliers are caused by the pathological condition of the patient during the CCVP test.

To achieve this goal, a deep understanding of how diseases manifest themselves on the data is essential. Here I have used both knowledge about inherent relationships among data and domain knowledge from the expert to obtain this understanding.

The knowledge about data is reflected on the maps produced by the SOM. For example, each node on the output map is likely to have a number of input vectors associated with it. These input vectors in turn determine the physical meanings of the node such as average motion sensitivity (Equation II), the typical patterns that the node represents, etc. (Figure 8). Using these physical meanings, the domain expert can try to group those input patterns that have the same or similar pathological meanings. In the case here, an input pattern consists of a vector of six elements, each of which represents whether the patient sees a certain stimulus location on the computer screen (Figure 2).

The expert identified four groups of patterns representing different pathological conditions. Group A is composed of input patterns reflecting that the patient under test is showing the early sign of upper hemifield[2] damage. Group B consists of patterns demonstrating that the upper hemifield of the patient is probably already damaged. Group C and D are made of patterns similar to group A and B, except they are used to represent the two different stages of the lower hemifield damage. Any two patterns that fall into the same group will be considered as having a similar pathological condition.

Take group A as an example. It contains the following four patterns:

{ (*1, 1, 0, 1, 1, 1*), (*1, 1, 1, 1, 0, 1*), (*1, 1, 0, 1, 0, 1*), (*1, 1, 0, 1, 0, 0*) }

These have been identified as possible patterns for a glaucoma patient showing early signs of upper hemifield damage. Two factors have been taken into consideration by the expert when selecting these patterns. First, the domain knowledge about early upper hemifield damage is used. For instance, damages in the upper hemifield locations 3 and 5 were included in those patterns. Damage in location 1 is not included, however, because it often indicates the upper hemifield is probably already damaged [182]. Second, the physical meanings of the output map are used, especially for how typical input patterns are associated with output nodes. For example, the above four patterns are mapped in a topographically connected area on the output map.

[2] Half of the visual field. See glossary.

These pathological groups are then used to check whether the outliers are measurement noise. This step was simply implemented as follows. First, those input vectors that are mapped to the outlier nodes are identified. Then they are checked for whether each of these vectors belongs to the same pathological groups as those patterns that were recognised as the more stable part of the data. If yes, it is then treated as a real measurement; otherwise, it is measurement noise.

One of the major difficulties in applying this method is that the patterns that make up those pathological groups are not complete. They (27 in total) are only a subset of all the possible patterns ($2^6 = 64$). Therefore, when there is a new pattern occurring, the above method cannot be applied. One of the main reasons why the expert could not classify all the patterns into those four groups is that the CCVP was a newly introduced test and the reflection of glaucoma patients and suspects on the CCVP data was not fully understood.

To overcome this difficulty, machine learning methods can be applied to generalise from those 27 classification examples provided by the expert. In particular, I have used the back-propagation algorithm [150] for this purpose. The input nodes represent the locations within the visual field, output nodes are those pathological groups, and three hidden nodes are used in the fully configured network.

The trained back-propagation network is able to reach 100% accuracy for the training examples (27) and to further classify another 26 patterns among the known classes, A, B, C and D. The remaining patterns are regarded as the unknown class since they have no significant output signal in output nodes. They have been found to be much more likely to appear in the outliers than in the more stable part of the data (Table 2).

Group	***Initial classification by the expert***	***Expanded classification by back-propagation***
A	4	8
B	7	22
C	4	5
D	9	15
Other	3	10
Unclassified	37	4
Total	64	64

Table 2. Classification of the four pathological groups

One of the interesting observations is that patterns within each of the resultant groups tend to be clustered in a topographically connected area, demonstrating the same property of the initial groups (Figure 14).[3]

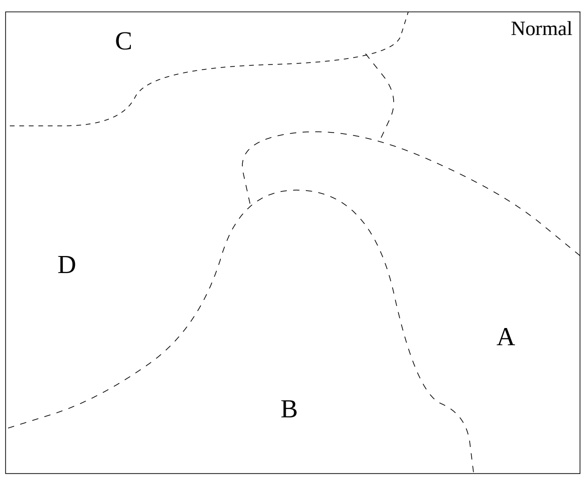

Figure 14. The four initial groups are expanded and are mapped onto the SOM

It should be noted that the example described above is a rather simple one in which there are only 2^6 possible input patterns. This particular version of the CCVP test has been chosen for its simplicity in order to make it easier to describe the general ideas in implementing the noise

[3] There is not an explicit area representing the patterns in the group "other" as they are mapped into some isolated output nodes in the areas of B and D.

identification principle. In fact, there is another popular version of the CCVP, which also tests the six locations within the visual field by ten repeating cycles, using, however, four levels of stimuli. Therefore, the data vectors produced within this test contain 24 elements instead of 6. Consequently, there are 2^{24} possible input patterns. I have also experimented with large quantities of data from this test using the proposed noise identification strategy. The results are similar to those of the simpler test described in the next section.

4.3 Evaluation

4.3.1 Several Observations Regarding the Strategy

The noise identification strategy is based on the assumption that the interesting properties in data are more stable than the noise. Below are several observations regarding this strategy.

First, explicit identification and deletion of measurement noise in data may be a necessary step before the data can be properly explored, as shown in this application. In particular, I have found that noise deletion can offer great assistance to the clinician in diagnosing those otherwise ambiguous cases (see 4.3.2). In a separate experiment with learning disease patterns from the CCVP data (Chapter 7), I have found that many useful features not initially found from the raw CCVP data, were uncovered from the data after the measurement noise was deleted.

Second, the use of domain knowledge supplied by the expert is of special concern as this type of knowledge involves a substantial number of subjective elements, and is often incomplete, as shown in this application. It should be pointed out that this strategy might not be effective for those applications where there is little relevant knowledge but a lot of *false noise*, i.e. many outliers actually reflect real measurements. Where there is little concern about the false noise situation, however, an unsupervised learning algorithm can be used directly to identify the measurement noise, in this case, all the outliers identified.

Third, other supervised machine learning techniques should also be applicable with the principle of learning from expert examples presented in Section 4.2.2. In particular, the C4.5

decision trees algorithm [140], which I use in Chapter 5, was also found effective. In the context of supervised machine learning however, there is an essential difference in terms of knowledge representation between the two methods, though less important in this application. The expertise model that the neural network learns is in a "black box" fashion while the decision trees generate a set of explicit rules. Such rules may be further interpreted and therefore they benefit experts themselves by expanding their knowledge.

Finally, no claim is made that this strategy can be used to identify all the measurement noise in data, or that all the noise identified is the real one. This depends on the ability of the chosen algorithms to accurately cluster those data items with common features and the quality of domain knowledge used to exclude the false noise.

4.3.2 The Results

This section presents the results in applying the proposed strategy to a set of clinical test data (2630 data vectors). The data were collected by ophthalmologic specialists from a group of glaucoma patients and suspects at the Moorfields Eye Hospital, London. To find out how successful this strategy is in achieving its objective, I introduce the concept of *reproducibility*.

As glaucoma is a long-term progressive disease, the visual function of the same patient should remain more or less the same during a short period of time. As a routine clinical procedure, identified glaucoma patients are regularly asked to conduct follow-up tests to monitor the progress of the treatment. Therefore test patterns from such follow-up tests within this time period should be very similar. However, this is not always true under real situations, as measurement noise is involved in each test, perhaps for different reasons with various appearances. Consequently, it is not surprising that a large number of test patterns from such follow-up tests showed disagreements to various degrees. One example of such follow-up tests is shown in Figure 15.

As one of the main reasons for the disagreement is the measurement noise, it is natural to assume that the motion sensitivity results of the two tests should agree (to a certain degree) after the noise is discarded. This then constitutes a strategy for evaluating the proposed strategy for identifying and eliminating noise from data.

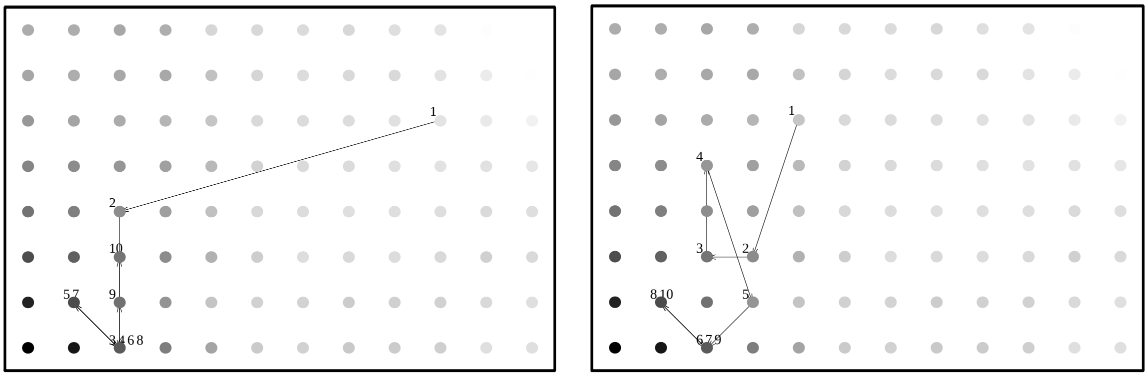

Figure 15. A pair of follow-up tests

There were 91 pairs of records available for this purpose. These tests were conducted within an average time span of less than a month. The average motion sensitivity measurements (percentage values calculated by Equation II) of these tests are contrasted in Figure 16 (left). The dot is used to indicate the result of the first test and the oval is used for the result of the second test. The difference between the two results for each case is illustrated by the line in between them. The same results after the rejection of noise by the proposed method are given in Figure 16 (right).

One of the major findings is that the results from the two follow-up tests have much better agreements after the noise is rejected. This is indicated by two observations. First, 80% of the pairs after the rejection of noise have reached almost total agreement (with less than 1.0% error), while only 48% of the pairs agreed in the original data set. This is reflected by the fact that there are many more cases on Figure 16 (right) where the dot and oval are overlapping or very close than those on Figure 16 (left). Second, if one calculates the mean difference between the two tests, 5.6 (95% confidence interval: 4.3 – 7.0) is the figure for the original data, while 3.4 (95% confidence interval: 1.8 – 4.9) is obtained after the noise is eliminated. This is indicated by the observation that the lines between the two tests are generally shortened in Figure 16 (right). These findings have shown that the proposed method does provide an effective way of identifying and discarding the noisy data.

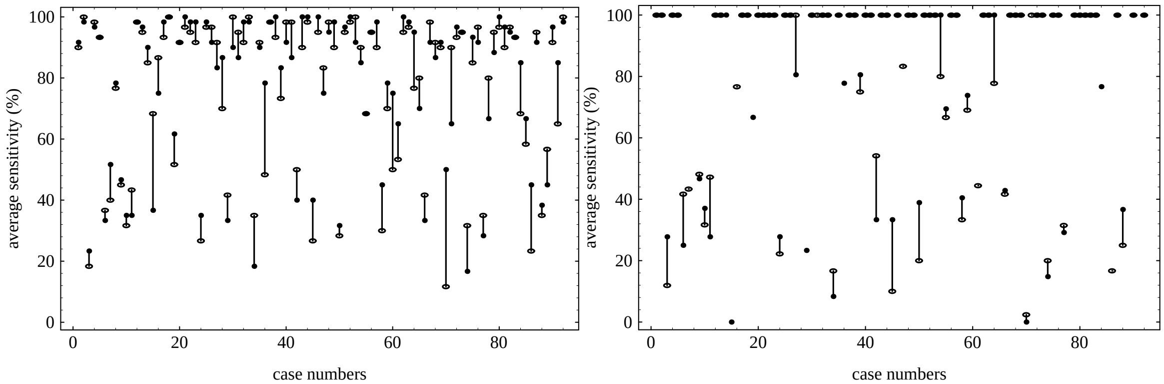

Figure 16. (left) before noise deletion; (right) after noise deletion

In addition, noise deletion may also be of direct diagnostic assistance to the clinician. One of the difficulties for the clinician is that the result from one test suggests that the patient is normal (no glaucoma), while the result from the other test shows that the patient is abnormal (having glaucoma of some kind).[4] Since much better agreement is shown between the two follow-up tests after the deletion of noise, there should be fewer pairs with contradictory test results. This is indeed the case with the experiment data as shown in Figure 16: there are quite a few conflicting cases in Figure 16 (left), while only two such cases exist in Figure 16 (right).

4.4 Concluding Remarks

In this chapter I have introduced a novel way of distinguishing real measurements from noisy data. The principle I adopted to achieve this task is that interesting properties in data are more stable than noise. To implement this principle for this application, the SOM is used to model a patient's behaviour during the CCVP test. The output maps are then used to separate the more stable part of the data from the outliers. Expert knowledge, augmented by supervised machine learning techniques, is used to check whether outliers are measurement noise caused by behavioural factors, or caused by the patient's pathological status.

The evaluation results have shown that the proposed strategy provides a satisfactory way of identifying *real* measurement noise in CCVP test data. Moreover, the explicit identification and elimination of the noise in these data have been found not just desirable, but essential, if the data are to be properly modelled and explored for specific problem solving tasks.

Apart from the strategy proposed in this chapter, I have also studied another strategy by integrating statistical methods with domain knowledge more closely [108]. It is an alternative attempt to model real measurements, i.e. how measurement should be distributed in a domain of interest. This approach is based on one of the most popular multivariate outlier detection method [176]: only a single outlier at a time is detected if the algorithm is applied sequentially [25]. This assumes that an outlier set $\{o_1, o_2, \ldots, o_n\}$ is ordered in that $o_1, o_2, \ldots, o_{i-1}$ must be rejected before o_i $(1 < i \le n)$ can be rejected. It is therefore possible to find out a subset of the outliers that should be rejected by sequentially testing each outlier against each plausible hypothesis, in the form of "knowledge-based" probability distributions. This strategy has the advantage that a number of competing hypotheses may be established, and can be tested against by the sequence of outliers to search for the most probable fit. For example, if several diseases have occurred in the data, a given test case may belong to any one of the disease categories. It must be checked with all the disease patterns before noisy outliers can be rejected correctly [108].

[4] The study [182] has suggested the "golden line" in CCVP that divides the normal and abnormal groups at the average sensitivity value of 0.75.

Chapter 5 Outlier Discrimination by Modelling Noise

In the previous chapter, a strategy for distinguishing between phenomena of interest and measurement noise was proposed and applied to the analysis of a set of visual field test data collected from a group of glaucoma patients in an eye hospital. That strategy attempted to model "real measurements," namely, how measurements should be distributed in a domain of interest (e.g., how glaucoma manifests itself on visual field data). Based on the knowledge acquired, the strategy then rejected outliers that do not fall within the real measurements. As pointed out previously, that strategy is applicable when a sufficient amount of knowledge about how real measurements should be distributed can be acquired. There are, however, many applications where this type of domain knowledge may not be readily available, and yet among outliers "phenomena of interest" still needs to be separated from noise.

In this chapter, I propose an outlier analysis strategy using relevant knowledge and understanding of the *noise* in the data. It attempts to model noise such as that caused by behavioural factors in the visual field test data, and accept data outside of the norms if they are not accounted for by a noise model. The experiment demonstrates that this strategy does significantly better at a diagnostic task than an equivalent approach that either utilises all data, or attempts to reject all non-normal values.

5.1 Strategy for Modelling Noise

The primary objective of this study is to develop computational methods for managing outliers by explicitly examining outliers with a view of either rejecting or welcoming them. The methods should not only successfully identify where the outliers are, but also carefully analyse them and their association with the entire data set to distinguish between the truly noisy and "surprising" but useful ones.

There are several issues regarding this process. First, it assumes that the outliers are caused by either behavioural fluctuations or pathological instabilities. The task of the expert, therefore, is to define and identify those cases, in which outliers are caused by behavioural factors and not by interesting pathological status. Second, such knowledge needs to be acquired, as well as to be represented in a form to make automatic processing possible. In other words, a "noise model" should be established to formulate an algorithm for computational outlier analysis. Here it assumes that such a model is not readily available for most applications (things would become much easier if it was), and it needs to be constructed or learned. In many application domains, the expert would find it easier to identify noise examples than general rules of how noise exhibits in data. Therefore, I proposed noise model construction would be based upon a machine learning process from the examples.

Below I outline the strategy for explicitly identifying and analysing outliers in data.

Definitions:

Let Ω be a p-dimensional sample space.

Let $X = \{ x_1, x_2, ..., x_n \}$ be a set of vectors drawn from Ω.

Let $O = \{ o_1, o_2, ..., o_r \}$ $(1 \leq r < n)$ be a set of outliers in X where $O \subset X$.

Let $C = \{ C_n, C_r \}$ represent two general classes – recognisable noisy cases, C_n, or other cases, C_r.

Let $F = \{f_1, f_2, ..., f_m \}$ be a set of features extracted from X.

Clustering and Outlier Detection:

Given a set of data points X, a clustering algorithm is applied to identify the more stable part of the data. The less stable parts of the data then become a set of outliers, O. Note that there is no guarantee that there will be outliers existing in all the data sets. Sometimes, there may be no major cluster of data points (e.g., there are several clusters or no cluster at all).

Noise Model Construction:

A noise model is constructed that could account for much of recognised measurement noise in a domain of interest. In particular, it assumes that a group of data sets, *Xs*, can be labelled by the expert into two general classes $\{C_n, C_r\}$, based on their relevant domain knowledge and close examination of data sets. This group of labelled data sets, together with a set of features, F, which may be extracted from the data sets, is then used to build the classification model (e.g., a set of classification rules) using an inductive learning technique. Those classification rules leading to the class C_n then become the noise model, *M*.

Eliminating Noisy Outliers:

Each new data set, for instance *X'*, is tested against the noise model *M* generated in the above step. If applicable, then the outliers *O'* within the data set *X'* can be rejected (due to known measurement noise).

I make the following observations regarding the proposed method:

1) The clustering algorithm used in the method can be a traditional statistical clustering algorithm [45], a self-organising neural network [95], or other machine learning methods [48].

2) The construction of the noise model requires a set of representative and labelled instances on which the "noise model" may be built. Each data set can be assigned in a domain of interest to two mutually exclusive classes. Class C_n indicates that the corresponding outliers in this data set are noisy data points, therefore can be deleted. Class C_r says we do not think the outliers in the data set are due to measurement noise, although we might not be able to say whether they are phenomena of interest either.

3) Note that many classification models may be constructed from a set of labelled instances. The classification model as mentioned in the *Noise Model Construction* step refers to the "best" model in terms of its predictive accuracy, its simplicity or interpretability, misclassification costs, or other appropriate criteria for the problem under investigation [172, 36].

4) The success of the method very much depends on the correctness and completeness of the noise model constructed. The correctness of the model depends largely on the quality of domain knowledge – a set of labelled instances in the proposed method, although the

choice of inductive learning algorithm might also matter. On the other hand, if the model does not sufficiently cover all possible types of measurement noise, the data set after cleaning would still contain much noise.

5) The precondition for using the method regarding the availability of relevant knowledge about the distribution of noisy data points (labelled instances) is reasonable in many applications. For example, in time series forecasting, the understanding of “special irregular events” and their effects on the forecasting results can be used to build the corresponding “noise model.” Using the data after removing the effect of the special irregular effects may often increase the forecasting accuracy.

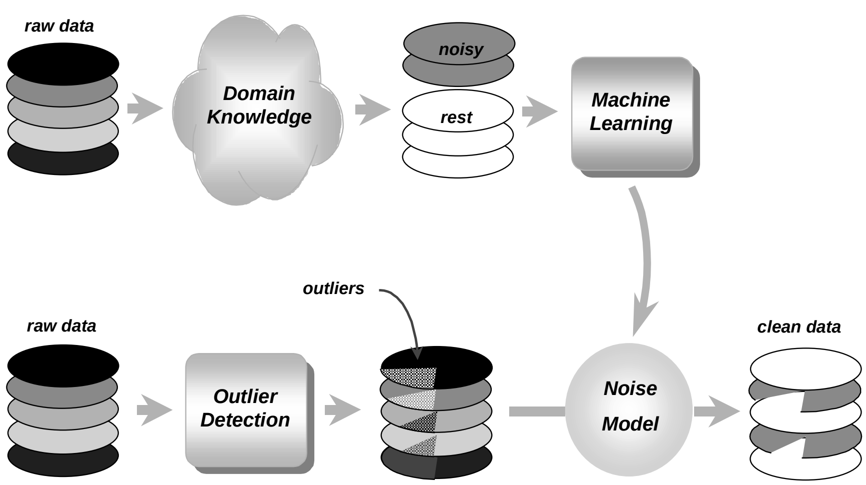

Figure 17. The strategy for modelling noise

5.2 Analysing Outliers Using Noise Model

Figure 17 illustrates how the strategy proposed in the previous section based on noise models works. Suppose that a set of training data, by using relevant domain knowledge, can be labelled into two classes: “noisy” (C_n) and “rest” (C_r). Class “noisy” indicates that the corresponding outliers in the data set are noisy data points, while class “rest” covers all other situations. Given a sufficient amount of training data, one can use any supervised machine learning techniques to build a “noise model” and this model, after validation, can then be used to help distinguish between the two types of outliers.

Note that the labelling of training instances is not always easy, especially with multi-dimensional data. To assist in this process, I have used the SOM to visualise and compress data into a two-dimensional map as discussed in Chapter 3. Data clusters and outliers then become easy to spot, and data are then relatively easily interpreted using the meaning of the map and relevant domain knowledge. So, given a data set, outliers may be detected and can then be tested using the noise model. As a result, noisy outliers can then be deleted, while the rest of the outliers are kept in the data set for further analysis.

5.2.1 Noise Model I: Noise Definition

In the application of the CCVP, noise in data are defined as those data points typically associated with learning effects, fatigue, and inattention. In this connection, we may define outlying data points due to fatigue on the SOM as follows. If the sensitivity values of several initial test cycles are high and similar to each other, and the sensitivity values of the remaining cycles are decreasing over time, then the data points corresponding to the remaining cycles are outlying due to fatigue. In this case, the winner nodes of the initial cycles tend to be in a small neighbourhood and the winner nodes of the last few cycles tend to move away from the small neighbourhood to areas where the sensitivity values are lower. One example of such a case has already been given in Figure 10.

On the other hand, if the sensitivity values of the initial few cycles do not show much regularity but the sensitivity values of the remaining cycles gradually become similar, then the data points corresponding to the initial cycles are outlying due to learning effects. In this case, the winner nodes of the initial cycles perhaps are irregular, but are gradually gathered around a small neighbourhood on the map. One example of such cases can be seen in Figure 12.

Figure 13 demonstrates a typical case of inattention. Clearly the subject had a normal visual field, but was distracted during the ninth measurement cycle. This results in poor sensitivity values for this particular measurement, leading to fluctuation in the data. This type of fluctuation, however, should not affect the overall results of the visual field. Therefore the data collected during this cycle can be dropped.

In all the above cases, decisions regarding whether to delete certain outlying data points are relatively easy. For example, the data points corresponding to measurements 7, 8, 9 and 10 may be deleted in Figure 10, while those corresponding to measurements 1 and 2 in Figure 12 may be cleared. However, things are not always this clear-cut.

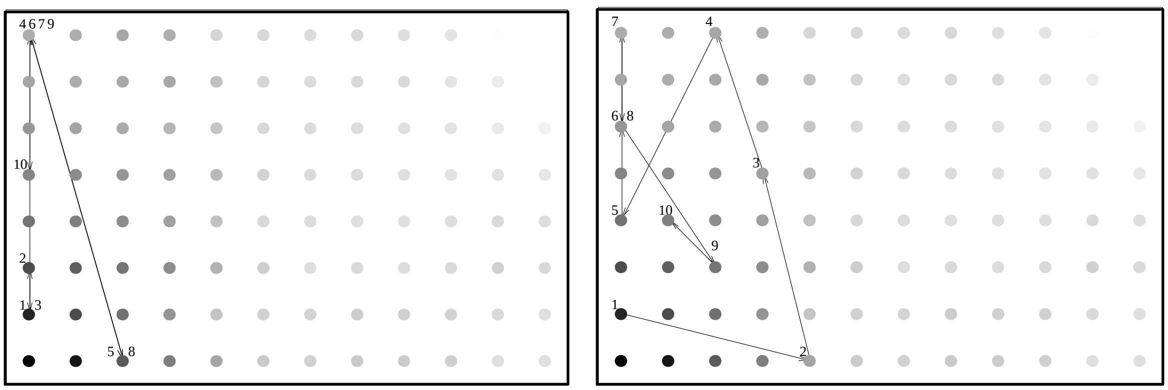

Figure 18. Examples of glaucoma cases with a cluster (left) and without a cluster (right)

Figure 18 demonstrates two test results for the same subject who had been confirmed by an ophthalmologist as a glaucoma patient. Figure 18 (left) does seem to show there is a cluster in the top left corner of the map. However, consider that none of the measurements has shown any high motion sensitivity (in the top right corner area) and that there are six measurements scattered around on the map. There is good reason to believe that these measurements might tell us something about the pathological status of the subject, and should therefore be kept. Meanwhile, Figure 18 (right) does not appear to show any interesting clusters and none of the measurements is very sensitive to motion. In this case, there is no easy way of finding out which measurements are noisy and which are not. All the measurements will therefore be kept for further analysis.

5.2.2 Noise Model II: Construction

The construction of the noise model in this application is as follows:

1) A set of test records (310 in total) was used for the purpose of building the classification models. A domain expert labelled each of the records using the visualised data presented by SOMs and relevant knowledge regarding the visual field test. A record is categorised into either C_n: corresponding outliers are measurement noise caused by one of the three behavioural factors: learning, fatigue and inattention; or C_r, the outliers might represent useful information, or "phenomena of interest."

2) Several features were extracted from the data sets and relevant domain knowledge. These features, together with those labelled instances as discussed in the above step, were used to develop the classification models. In particular I have used Quinlan's C4.5 [140] to learn a set of production rules in the form of decision trees. Experiments were performed to find a set of rules that would minimise the errors on the unseen cases. This includes the division of 310 cases into training and testing cases of various sizes and the use of a more robust method of 10-fold cross validation. It appears that the 10-fold cross validation presents the most confident rules for the data set.

3) Those production rules within the C_n class now become the "noise model," and the model can then be used to delete the corresponding outliers for future test data (see below), when applicable. For any given piece of test data as input, the decision trees will be able to tell whether the test case belongs to C_n or C_r. It is apparent that the confidence of the decisions depends on the quality and coverage of the training samples provided by the expert.

I have experimented with learning the production rules by combining all three behavioural factors into one class C_n, and tried to separate it from the class C_r for the remaining cases. With the labelling of just over 300 cases, useful rules have been learned for deleting "noisy" outliers and also for keeping outliers when it is not sure whether they are noisy or not. C4.5 was able to achieve an error rate of 13.9%, a reasonable result with a small amount of noisy data. Remarkably, one of the learned rules closely resembles what the domain expert had been using in identifying noise cases:

Expert: If the stability[5] > 0.60, average-sensitivity[6] > 0.75, then delete the outliers (the outliers are caused by behavioural factors).

Given **T** as any test record, the rule says:

C4.5: $\forall$**T**, stability(**T**) > 0.60,

stability(**T**) $\leq$ 0.90,

average-sensitivity(**T**) > 0.766667

$\rightarrow$ **T** $\in$ class C_n (with the certainty of 78%)

The rule seems to be a little cleverer in that it has an upper limit for stability (0.90). For instance, if the stability is (or very close to) 1.00, then there might be no measurement noise after all. Therefore it should not belong to the class C_n.

5.2.3 Evaluation

This section presents the results of applying the noise model generated above to the elimination of those noisy outliers in a set of visual field test data. The corresponding test was recently conducted by ophthalmologic specialists and GP clinicians in a large urban general practice in North London for a glaucoma case findings study [188]. All patients aged 40 years or older who routinely attended the practice for a three-month period during the pilot study were offered the test. In this particular study, the situation is more complicated than that associated with the data described in Chapter 4 where the test cases were either normal or involved glaucoma at a certain stage. The composition of possible diseases in this study includes not only glaucoma but also different types of conditions such as cataract [188]. For this reason, the knowledge about how a single disease manifests on the test data, even if it can be acquired, will not provide much help in distinguishing outliers. On the other hand, the expert found it relatively easy to identify a set of cases with various behavioural fluctuations, such as inattention or fatigue.

---

[5] The proportion for the stable part of data to the entire data (see Section 6.2).

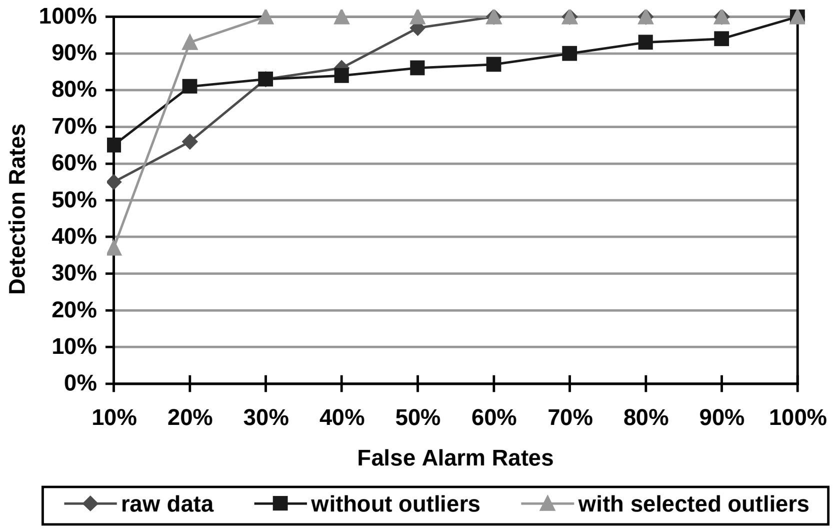

Figure 19. Detection rates versus false alarm rates for three data sets

Among those patients screened by the CCVP, 78 were later assessed clinically in the practice by an ophthalmologist. Among the tested eyes from the 78 people, 22 eyes were later assessed as having glaucoma, 81 eyes were confirmed as normal without any disease, and the rest were diagnosed as having other types of ocular abnormalities. The noise model was applied to these 103 eyes and Figure 19 summarises the results of examining the *discriminating power* of the CCVP in terms of its diagnostic *detection rate* versus *false alarm rate*. Three different data sets are shown in Figure 19: the original 103 eyes corresponding to all the glaucoma and normal eyes, the data set obtained after all outliers are deleted from those test records, and the data set with selected outliers (after applying the noise model to eliminate noisy outliers).

The Receiver Operator Characteristic, or ROC analysis [65], is used to assess the test's diagnostic performance by displaying pairs of false alarms and detection rates throughout the whole range of the CCVP's measurements. While the curves shifted towards the upper left of the diagram, performance of the test is improved in the sense of maximizing detection rates and minimising false alarms. The decision threshold used for discriminating between normal and glaucoma eyes is the average motion sensitivity as in Equation II. The curves are plotted when the decision threshold is changed throughout the range from 0 to 1.0. For example, the

[6] See Equation II in Section 2.3.

cut-off threshold value of 0.7 has been found to enable the data with selected outliers to achieve a detection rate of 90% and a false alarm rate of 20%.

From Figure 19, it is clear that the data with selected outliers performed better than the other two in terms of maximising the detection rate and minimising false alarms. For example, this group can achieve 100% detection rate, while the corresponding false alarm rate is 30%. This is equivalent to saying that none of the subjects suffering from glaucoma would have escaped notice and only 30% of those normal subjects would have been unnecessarily referred for further examination. To reach 100% detection rate by using the raw data, however, 60% of normal subjects would receive false alarms. In comparison with the data with selected outliers, this doubles the number of people who would be referred and further examined unnecessarily.

It turned out that the data obtained by deleting all the outlying data points (without analysis) performed the worst in this experiment. This suggests that a great number of outlying data points among the glaucoma patients were indeed reflections of the pathological status of the patients. Deleting those outliers leads to loss of critical diagnostic information.

5.3 Concluding Remarks

Distinguishing between noisy and noise-free outliers is a difficult problem and I have demonstrated how domain knowledge regarding the distribution of noisy data points may be used to address this problem. Instead of relying on the knowledge about “real measurements” in data, the strategy that I present in this chapter acquires the knowledge about the noise. This type of knowledge might be acquired more easily and more reliably in some applications. As shown in this chapter, the knowledge about noise may be learned from examples provided by the domain expert. By using the classification rules generated from the inductive machine learning technique, the noisy outliers can be separated from the noise-free ones.

An important observation from the work reported in here is that due care must be taken when deleting outlying data points. As with the visual field data from a GP clinic, the results obtained using “outlier-free” data are even worse than those from the raw data. Appropriate analysis should be performed using relevant domain knowledge in this regard.

Chapter 6 Noise Reduction at Source

In this chapter, research work in the thesis will be presented as part of the project METRO [109, 113] to help manage measurement noise for the CCVP. METRO is a joint project between Birkbeck College, the Institute of Ophthalmology and Moorfields Eye Hospital. Its main objective was to make the CCVP a more reliable test. In addition to the studies for outlier analysis described in previous chapters, there was also an effort at reducing noise at other stages of the IDA process and in particular, at the data collection stage. Several techniques have been developed for this purpose.

6.1 Minimise Measurement Noise Systematically

As pointed out previously, the difficulty in obtaining reliable test results needs to be overcome before the CCVP can be widely used. The main effort in making the CCVP a more reliable test is to search for effective and efficient ways of handling the measurement noise. The approach is to minimise the chances of introducing noise wherever possible and to identify and delete noise after it occurs. It is apparent from the discussions in the previous two chapters, that the outlier analysis strategies would be more useful if a certain level of data quality can be achieved. For example, the number of "totally unreliable" cases (see Section 3.4) would be reduced if data were collected more reliably. As a systematic development towards the objective of minimising measurement noise, therefore, it is an obvious choice that noise should first be reduced from the source, i.e. while data are being collected.

There has been extensive interaction with domain experts to identify factors that may contribute to the noise, apart from analysing measurement noise from the CCVP data using the previously described methods. By investigating ways to reduce the effects of these factors, therefore, it is possible to decrease noise and higher quality data may be obtained. Several significant factors cause most of the noise in data [111].

As discussed in Chapter 3, the CCVP data necessarily contain a great deal of measurement noise caused by individuals' behavioural fluctuation. Because learning effect, inattention and fatigue are among the most common reasons for introducing noise, these undesirable behavioural phenomena should be kept to the minimum wherever possible. For example, Collins [38] carried out a number of experiments to find the friendly test interface and most effective testing strategies. This will help keep the subject alert and reduce the possibility of inattention. An automatic training and self-adaptive session is carried out before each test until certain criteria have been satisfied so that the test will proceed with better confidence and learning effect will be reduced. Furthermore, because the original CCVP SAT strategy always measures visual function with a fixed number of repeating cycles, it may introduce unnecessary fatigue effects for individuals who do not need to go over the entire testing cycles to obtain their reliable test results. Cho [33] developed a dynamic testing strategy to help determine the appropriate number of repeated measurements for each individual CCVP test and therefore the noise caused by fatigue may be reduced. These measures will be presented in more detail in the "intelligent user interface" in the following section.

In addition to the above measures to decrease the chance of noise occurrences, I have introduced another strategy by providing the on-line assessment of the data quality. Here is the basic idea: if a test has been deemed "unreliable," it would be most effective to perform an immediate follow-up test. Nonetheless, the follow-up criterion must be carefully determined, because 1) data must be collected efficiently in real-world applications and 2) fatigue effect in psychophysical tests reduces data reliability and therefore limits the number of repeats. In the next section, we shall describe an indicator to estimate data quality on-line. Based on the estimation, we employ a set of decision rules as the control strategy to direct the follow-up tests as described in Section 6.3. This will be followed by the discussion of experimental results in Section 6.4.

6.2 Intelligent User Interface

This section describes the study and implementation of the intelligent user interface (IUI), which helps reduce noise caused by behavioural factors [107]. The work presented in this section (6.2) is a collective effort in METRO and not necessarily done by the author. I however, consider it beneficial to be presented for the integrity of the project and to better understand the thesis. Contributions from other people are explicitly stated and the origins of these individual pieces of work are referred to in the bibliography accordingly.

Figure 20 illustrates how the medical operator normally manages the standard visual-field test and how the intelligent user interface helps provide high quality measurements. Normally, the operator invites a subject for a test, explains the test to the subject, and enters the subject's details. The operator then monitors and controls the test, and reports the results to the doctor.

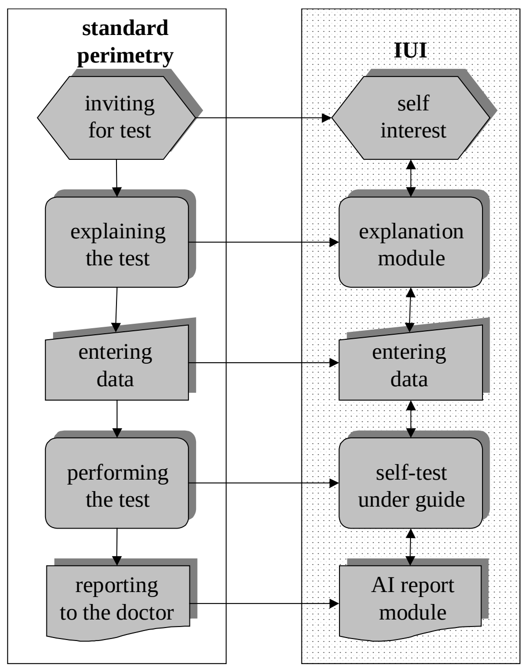

Figure 20. The role of the intelligent user interface (IUI)

For the CCVP with IUI, a subject examines the visual field out of his or her interest in the early detection of possible diseases. An explanation module explains the test and procedure, and the subject enters the data. Self-testing then proceeds under the system's guidance. Finally, the system issues a report, informing the subject whether he or she has passed the test and, if not, which part of the visual field appears to have problems.

One of the key considerations in the development of the test interface was how to handle the problem of human behavioural instability during the test so that the subjects would make fewer mistakes (false positive or false negative responses). The major behavioural factors are learning, inattention, and fatigue.

Learning effects. Subjects are more likely to make mistakes the first time they take the test. Subjects who have done the test before, but who are re-tested after a considerable time interval, might also make mistakes.

The IUI provides self-training sessions for such users before the actual test. The program keeps the subjects in the training session until it judges them familiar enough with the test. It bases this judgement on response time (from when the screen displays a stimulus to when the subject responds). If the response times for most of the responses collected during the test are close to each other and no great irregularity exists, the subject can move on to the actual test.

Inattention. To keep the subject alert during the test, care has been taken to ensure that taking the test is an interesting experience rather than something that the subject "has to do." Techniques include

- A feedback system that uses sound and text to indicate the subject's performance,
- An adaptive fixation point that uses smiling and frowning faces to attract the subject's attention,
- Interesting visual stimuli for testing, and
- Dynamic test strategies for individuals.

For example, one of the original test screens consists of a number of vertical bars with a central circle for the fixation point (see Figure 2 in Section 2.3.3). The program tests a fixed number of bars, using several different stimuli including bar movements and flickers. Some subjects might find the screen layout and the testing stimuli monotonous, and therefore have difficulty concentrating on the test. To tackle this problem, alternative screen layouts and testing stimuli have been developed, for example, using the front view of cars rather than vertical bars. Instead of the simple movement of bars, headlights flashing and windscreen wipers moving are used as stimuli [38]. These developments have led to a test that allows several different screen presentations, each providing a different test image and presenting different degrees of stimuli.

Additionally, a special function is introduced to check whether the subject is really concentrating on the test. The display of the stimuli on the screen is now irregular – for example, stopping for a while after a fixed number of regular displays. This strategy is particularly effective for dealing with those subjects who anticipate the frequency of stimuli presented.

Fatigue. Reducing fatigue was also a focus in METRO. Not every subject needs to go through the same number of measurement cycles. The test results from the first few cycles might provide sufficient information for those with stable testing behaviour. Therefore, it is important to determine whether at a certain stage of the test, a sufficient amount of information has been obtained. If it has, testing can be stopped.

This dynamic testing strategy, developed by Cho ([33], [34]) has these key steps:

1) Calculate the sensitivity values of a few initial measurement cycles.
2) Use these calculated motion sensitivity values to predict the values for all designated numbers of cycles.
3) Calculate the squared error between the sensitivity values of a few initial cycles and the predicted values of all designated number cycles.
4) Compare the squared error with a predetermined tolerance. If the squared error is less than the acceptable tolerance, stop testing. Otherwise, perform another measurement cycle, and repeat steps 1) to 4).

This strategy was tested on a set of CCVP data involving six test locations, four types of testing stimuli and ten measurement cycles. Various techniques such as neural networks, multiple regression and decision tree induction were used to implement step 2). This strategy kept the number of measurement cycles relatively small without having much adverse effect on diagnostic accuracy. In particular, Quinlan's decision-tree induction program, C4.5 [140], achieved approximately 95% accuracy while saving on average two measurement cycles per test. The prediction (classification) models in the form of decision trees were built using four attributes, eleven classes and several hundred cases.

6.3 Stability Analyser and the Follow-Up Control Strategy

6.3.1 Stability Indicator

In this section, I introduce indicators that may be used to estimate data quality, immediately after data collection is completed. These indicators play an important role in the data collection stage, where data may be collected again soon after it is determined to be of poor quality. For example, a follow-up test can be performed while the subject is in the clinic, so that data may be obtained more reliably.

Two such indicators may be used for this purpose. The first is the *stability indicator*, which measures the amount of outliers contained in a data set. The second is the reliability indicator, which measures the amount of noise involved in a data set. The higher the values of these indicators are, the less the amount of outliers or noise contained in the data.

Since noise is typically involved in a CCVP test and in general, any psychophysical test, the key point in determining a stable test is not the similarity among *all* repeated measurements, but the similarity among *most* of the measurements. To implement a stability analyser used by the CCVP each time a test is conducted, we want to judge whether measurements from a substantial number of repeating cycles are close to each other. This is achieved by an outlier detection algorithm as presented in Section 4.2, using the geometry positions of the output node and their relative distances on the SOM. The approach taken here is to search for a maximum set of nodes on the output map, which occupies the smallest topographical area. Specifically, Equation XIII is used as an evaluation function for this purpose. The objective is to find a set of output nodes, $S \subseteq O$, which minimises the value of $F(S)$. The test is considered unstable if it cannot find an S such that $F(S)$ is sufficiently small. Otherwise, the test is stable and S is called "the stable part of test data". What remains then constitutes outliers. Thus for each test, the stability indicator is defined as the proportion of the stable part of data to the entire data.

On the other hand, with the strategies described in Chapter 4 and Chapter 5, it would be possible to identify *the noise* instead of *all the outliers*. One could argue that it would be more sensible to use the reliability indicator as opposed to the stability indicator as a data quality indicator. There are some problems with the reliability indicator, however. First, the reliability indicator requires outliers to be further analysed, but the strategies of analysing outliers need much expertise in either known diseases or known behaviour patterns. There is no guarantee that the knowledge is solid universally, because the domain knowledge is typically specific only to a particular disease or certain behavioural problems exhibited by certain populations. The scope of such knowledge would be too narrow, should it be applied to the task of making the CCVP a generic eye-screening tool. Second, the CCVP tests are usually carried out as a timely diagnostic routine, which is restricted by the availability of both clinicians and patients, or as a mass field epidemiological investigation, which is a resource extensive activity. Therefore the first-hand information is valuable and cannot afford to be missed. It would be an unrecoverable loss if an unstable test were mistakenly presumed reliable so that an immediate follow-up was not performed. The outlier analysis strategies are more useful when the test data have been already gathered and grouped, when noise cleaning becomes the only way to improve data quality.

For the above reasons, I have used the stability indicator to estimate the data quality while collecting data in the field.

6.3.2 Control Strategy

Once the stability analyser is established, how would it be used to direct the follow-up tests? More specifically, how is the indicator used to decide whether the current observer needs to have a follow-up test? In principle, the instability of the current test would normally indicate that the test needs to be repeated again. However, if this test instability has been found to be due to pathological factors pertaining to the patient, it would be inappropriate to ask the patient to keep repeating the test. Besides, these follow-up tests cannot be carried out indefinitely as the patient will become increasingly tired and error-prone. All in all, we want to adopt a control strategy that can lead to reliable test results without subjecting the individual to more repeated follow-ups than necessary or than can be endured. After extensive data exploration and interaction with domain experts, I have summarised the knowledge into such a strategy below.

If the current test is judged to be stable by the stability analyser, then stop the testing. Otherwise, if it is the first test for the observer, conduct a follow-up test, as more information is required about the observer before any decision can be made. However, if the second test is still not stable, some further possibilities need to be considered.

First, the instability in test results might be due to some non-behavioural factors such as a certain pathological condition pertaining to the patient. For example, it is quite possible that there are still fluctuations in the responses at certain stimulus locations, even if the observer has fully concentrated on the test. The nature of the disease has dictated the responses. These fluctuations might be large enough to enable the stability analyser to judge the test as unstable. If it is the case, the second test results would bear remarkable resemblance to those of the first. In other words, the test can be stopped if the *visual function patterns* (see Equation III) of these two tests show close agreement with each other, supposing the repeatable pattern represents a certain pathological status of the observer.

There may exist many possibilities if the second test is still unstable and the agreement between the two tests is not found. For example, it might be the case that the observer is not in the right "frame of mind" to conduct the test. Equally, it might be the case that the observer experienced the test for the very first time and had absolutely no knowledge of computers or the test itself. During the second test, the observer had begun to get used to the test. In this case, it might be beneficial to subject the observer to another test in order to see whether the patterns from the second and third tests are repeatable.

There must be some exceptional cases that still remain outside of the scope: the third test is still unstable and no repeatable visual function patterns can be found. This might be caused by some unknown pathological conditions, or the individual suffered behavioural problems during the three tests, or was troubled by the testing environment. In this case, the test should be stopped and perhaps advise the individual to come back some other time for another complete CCVP test. It is unlikely that a reliable visual function measurement will be obtained by continuing to test, as he or she could become very fatiguing.

The above strategy can be expressed as the following decision rules. Let $\mathbf{T}_i$ be the ith test of the same observer, and let $P(\mathbf{T}_i)$ be the visual function pattern calculated from the observer's ith test (defined by Equation III). Let $P(\mathbf{T}_{i-1}) \cong P(\mathbf{T}_i)$ denote the agreement between the patterns of the two tests $\mathbf{T}_{i-1}$ and $\mathbf{T}_i$. This agreement is measured by the Euclidean distance between the two patterns. Let $S(\mathbf{T}_i)$ be the stability indicator of $\mathbf{T}_i$ and δ be a threshold to determine the data set is stable.

if $S(\mathbf{T}_i) \geq \delta$ **then** `stop; output` $P(\mathbf{T}_i)$

otherwise

> **if** $i = 1$ **then** `proceed with another follow-up test`
>
> **otherwise**
>
> > **if** $P(\mathbf{T}_{i-1}) \cong P(\mathbf{T}_i)$ **then** `stop; output` $P(\mathbf{T}_i)$
> >
> > **otherwise**
> >
> > > **if** $i < 3$ **then** `proceed with another follow-up test`
> > >
> > > **otherwise**
> > >
> > > > `stop; no output`

6.4 Experimental Results

There have been many experiments conducted for aspects of quality improvement in data collection process. This section will present two such efforts as being within the scope of work of this dissertation.

6.4.1 Number of Repeating Cycles

As the stability analysis of the CCVP depends on the data collected from repeated measurements, it is important to determine an *appropriate* number of repeating cycles so that the number of unstable tests can be minimised. This is accomplished by hypothesising a different number of repeated measurements on existing test records and calculating the corresponding number of unstable tests assessed by the stability analyser. I found that the stability of the CCVP tests would normally improve when the number of repeated measurements increased from 6 to 10, indicated by the percentage of stable tests (Table 3). This finding is consistent with the previous study, in which the detection power (judging from both the detection rate and the false alarm rate) increased as a function of the number of repeated measurements from 2 to 10 [182]. Even though a larger number of repeats are more likely to obtain more stable test results, arbitrarily increasing the number of repeats would be less efficient and would increase fatigue effects. Further results were obtained by Cho [34], with dynamically deciding the optimal number of measurements for each individual.

Number of measurements	***Percentage of stable tests***
6	74.0%
7	72.4%
8	80.0%
9	81.1%
10	86.9%

Table 3. Number of repeated measurements and stable tests

6.4.2 Follow-Up Control Strategy

The effect of the proposed control unit over a large number of the CCVP observers should also be evaluated. For example, the strategy mainly relies on one follow-up test to understand the visual function of the observer and only under certain special circumstances is a third test advised. Does the observer find the follow-up test easier to follow than the initial one? How often is an observer advised to have the second or even the third test? In order to answer these questions, one should examine test records not only from those who *need* to conduct the follow-up test but also from those who are supposed to need only a single test. These results shall be reported as follows.

A follow-up test was carried out randomly on 532 different eyes amongst 3182 test records collected from 2388 different eyes in Nigeria. I have found that 371 (70%) eyes were stable and the remaining 161 eyes were unstable in their first tests. Among those unstable observers, 95 became stable in their second test. In the rest of 66 observers, 51, though indicated unstable, have similar patterns to those from the first test. In other words, only 15 eyes, or 2.8% are still left undecided and need to conduct the third test.

The above observations have shown that one test is sufficient to obtain stable CCVP test results for 70% of eyes, while *one* follow-up test is normally adequate for the rest of the population. Together with the data reliability evaluation [29], it has indicated that the follow-up control strategy is an efficient way of obtaining reliable test results for the CCVP.

6.5 Concluding Remarks

In this chapter, I have investigated a number of issues regarding how measurement noise may be reduced while data are collected. It is an important first step in managing noise, before the data are cleaned by any outlier analysis processes. First, the data collection process should be made less prone to human errors. Next, it is desirable to provide a more reliable measuring facility – external events causing noise such as unstable behavioural problems should be minimised. An intelligent user interface was systematically studied in the context of METRO, an application that helps improve data quality for CCVP. Several important techniques have been developed for this purpose.

While the above measures have helped to improve data quality for the CCVP, a more generic approach is to measure the data quality, and to re-do the measurements if indicated to be necessary. This is achieved by a stability analyser and a follow-up control strategy. The stability analyser provides an indication of the percentage of outliers involved in the data without any *a priori* knowledge about the subject or the population being measured. Though there is no guarantee for the number of outliers approximating the amount of measurement noise, there is generally close correlation between the two. Furthermore, the pieces of data having a significant amount of outliers are usually a superset of those data having a significant amount of noise. The strategy is therefore considered a conservative estimation of the need for a follow-up measurement. I have also pointed out that, although follow-up measurements are usually an inexpensive way to improve data quality, they are still restricted by resources and other limitations. In the example given above, the subject may not tolerate too many follow-ups, and operators may be limited by the resources available. As a result, a strategy of determining whether a follow-up is still needed after the stability analyser plays its role in indicating an unstable test.

Chapter 7 Outlier Management and Knowledge Discovery from Noisy Data

In Chapter 4 to Chapter 6, I have demonstrated how the proposed outlier analysis strategies have improved data quality for the CCVP. Since the ultimate goal of outlier management is to improve the problem-solving tasks in application domains, the outlier analysis strategies shall be applied to problem-solving tasks in the CCVP in this chapter. As the CCVP was a newly proposed visual screening system, one of the most crucial tasks would be to validate its diagnostic power. In other words, how accurately the CCVP can reflect patients' visual functions and how sensitive it is to detecting early optic nerve defects. In addition, to make the best use of CCVP test results for detection, diagnosis, and monitoring of optic nerve diseases such as glaucoma and optic neuritis, it is important to understand how these diseases manifest themselves on the corresponding test data.

In the following sections I shall focus on how outlier management strategies helped to validate the CCVP as a diagnostic tool and to acquire relevant knowledge from the noisy clinical data. I shall first introduce an interactive-heuristic knowledge discovery process, which is used to discover disease patterns from a large quantity of clinical data. I shall then discuss the knowledge acquired, which has been found to be useful either for validating the functions of the CCVP itself, or for better understanding the manifestation of a particular disease on the CCVP data.

7.1 Interactive-Heuristic Knowledge Discovery

In this section, I propose a method for knowledge discovery from a large quantity of noisy data. It involves several steps with a number of IDA techniques and interaction with the knowledge analysts. First, noise in the data is located and eliminated using the proposed outlier analysis strategies. The resultant data are then used to discover the hidden features and form useful knowledge. The SOM is applied for abstracting features; a visualisation method is introduced for displaying those features; and knowledge analysts are invited to guide the knowledge acquisition process. This interactive procedure is employed for all these steps, and the knowledge is discovered in a heuristic manner. At the end, I propose a strategy using the concept of *reproducibility* to verify the discovered knowledge.

The general concept of data cleaning in this method is useful to those applications where the presence of noise in data significantly impacts knowledge discovery. The interactive knowledge discovery process may be applied to situations where data properties are abstracted and visualized by the SOM. More specifically, the method may be applied to multivariate data sets where analysing correlation among variables may reveal interesting properties pertaining in the data, but cannot be directly explained without extensive interaction with human experts.

Figure 21 schematically illustrates this procedure in four steps – filtering noise, abstracting features, visualising features and analysing knowledge. I shall discuss these four steps below.

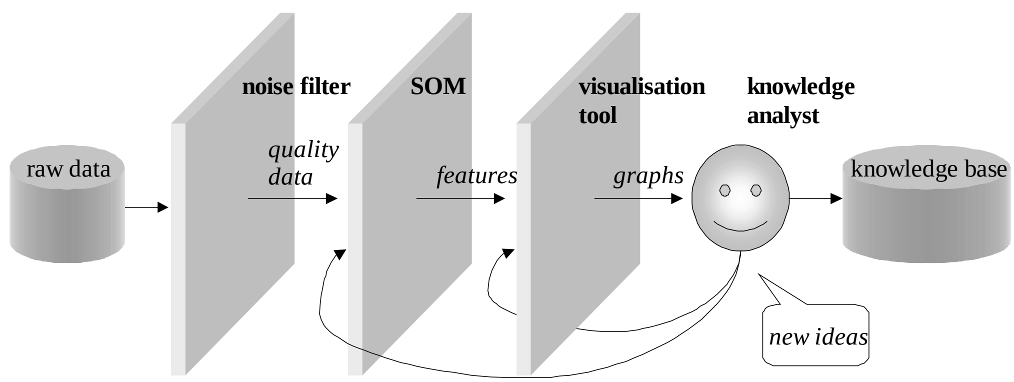

Figure 21. Interactive knowledge discovery

7.1.1 Filtering Noise

As discussed previously, measurement noise is typically involved in most data obtained from real-world applications. Any knowledge discovered from these noisy data, therefore, would certainly be questioned over its validity. Consequently, the first step of knowledge discovery is to filter out noise using outlier analysing strategies described in Chapter 4 and Chapter 5.

7.1.2 Feature Abstraction with the Self-Organising Maps

Having managed noise in the test data, I shall discuss how the SOM can be employed to uncover the hidden features among the resultant data.

The SOM is organised in ways that reflect the physical characteristics of the input data. Using Kohonen's self-supervised competitive learning algorithm [95], output nodes can learn to specialise in detecting intrinsic features existing in the input data. For that reason, the SOM is sometimes known as "self-organising feature maps [93]." I have observed two important characteristics of the learning algorithm from my experiment.

1) *Feature Clustering and Representing*: output nodes can be used to represent different clusters of input data. In other words, each cluster in input data is represented by an output node (although there can be many non-representative output nodes on the map). This is the effect created by the learning algorithm, which always maps a data point onto an output node and updates the weight parameters of that winner node accordingly. As a result, the weight vector of each representative node constitutes the centre of the cluster (bear in mind that the weight vector of an output node has the same dimension as the input vector). After the SOM has learned from the input data, therefore, the features in the input data would be abstracted and represented by the output nodes on the maps.

2) *Neighbourhood Smoothing*: representation of clusters or output nodes could be laterally affected by their neighbours. This is because the learning algorithm updates neighbouring nodes for a given learning node, though to a lesser extent (see Equation VI in Section 3.2). As a consequence, these unusual patterns, though they actually appear in the input data, may be *smoothed* by their neighbours. Therefore the features represented by the SOM would be formed by those more typical clusters.

The first characteristic is obviously required in this application, as we want the features represented by the SOM to reflect the nature of input data. Vector quantisation (VQ) error is a standard measurement for this characteristic as in Equation XIV, where v_i $(i = 1, \ldots, ND)$ is the input vector and c_j is the winner node of the given input vector.

Equation XIV. $$VQ = \frac{1}{ND}\sum_{i=1}^{ND}\left\| v_i - c_j \right\|$$

The second characteristic is also appropriate because noise in the data, though significantly reduced using the techniques discussed previously, may still remain and interfere with extracting features. The “smoothing effect” of the learning algorithm would help reduce the interference. Consequently, it would be reasonable to assume the weight parameters to be better representations of the motion sensitivity measurements for its corresponding cluster than the input data themselves. I have found that using the SOM “abstracted features” with weight parameters instead of the raw data has improved feature visualisation. As we will see in the following sections, the features discovered by this method have helped this application in significant ways.

The topographical product (Equation IX in Section 3.2) is considered a good candidate for measuring the second characteristic. Optimising both characteristics therefore, has been the convergence criterion in my implementation, which is done by minimising both of the measurements.

At the end of the process, the weight parameters in the SOM are shown to represent “features” in the input data. For a given SOM with N input nodes and M output nodes, there will be M weight vectors, each with N dimensions. Instead of analysing all the ND input data vectors, I shall take the M (usually $M << ND$) weight vectors to visualise the features abstracted, as it is shown in the next section.

7.1.3 Visualising Features

By visualising the weight parameters, features in the large quantity of data can be discovered more easily. For a given set of weight parameters, there can be a number of ways to visualise them. An obvious choice would be to examine the “physical meanings” of the maps, as described in Section 3.3. But instead of averaging all input vectors, here it averages the weight parameters on connections to all input nodes for a given output node.

Suppose $C = \{c_1, c_2, \ldots, c_M\}$ is a set of output nodes with weight parameters $W = \{w_{11}, w_{12}, \ldots, w_{ij}, \ldots, w_{NM}\}$ where N is the number of input dimensions, $P = \{p_1, p_2, \ldots, p_M\}$. A set of $M \times N$ "physical-meaning maps" is then calculated as in Equation XII. For the application of CCVP data, the maps mean the average motion sensitivity for all levels of stimuli and over all testing locations on the retina. An example is given in Figure 8 (in Section 3.4), except it was calibrated with input data instead of weight parameters.

Subsequently, one might find that the physical meanings of the output nodes vary greatly if only looking at the maps at each individual input dimension. In this case, the physical meaning P would be represented as $p_j = w_{ij}$ ($1 \leq i \leq N$, it corresponds to the input dimension). As a result, N physical-meaning maps were produced for all input dimensions and they were visually studied by the knowledge analyst. In the case of CCVP data, the map of an individual input dimension corresponds to the motion sensitivity value at a single stimulus level and a single testing location on the retina. An example of the physical-meaning maps for a set of six dimensional data is illustrated by Figure 22, where the six dimensions correspond to the six testing locations (see Figure 2). Instead of plotting discrete output nodes with density values in Figure 8, I have used a continuous contour plot for easier visualisation of the density distributions of the weight parameters. The contour graphs use relative grey scale range ($min(p_i)$, $max(p_i)$) instead of (0, 1), so that non-linear correlation among dimensions may be exposed more easily.

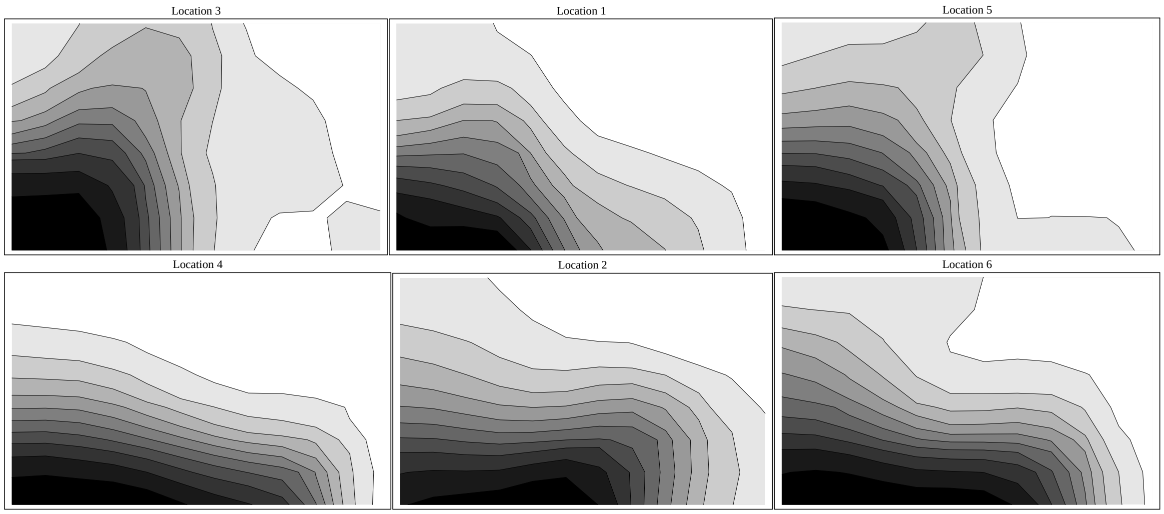

Figure 22. Physical-meaning maps for individual input dimensions

Like the case shown in Figure 22, one might discover that the maps of certain input dimensions are more similar to each other than to others. This discovery leads to an effort to visualise correlation among all input dimensions. First, the weight vectors associated with output nodes are used to form a new vector $P_i = (p_{i1}, p_{i2}, \ldots, p_{iM})^t$ for each input dimension i, where $p_{ij} = w_{ij}$ $(1 \leq i \leq N)$. Vectors associated with any two input dimensions, P_x and P_y, can then be used to draw their correlative graphs. Figure 23 is one way of drawing such a graph, which is done by taking each of the pairs (p_{xj}, p_{yj}) $(1 \leq j \leq M)$ from the two vectors, P_x and P_y, to plot in a two-dimensional graph.

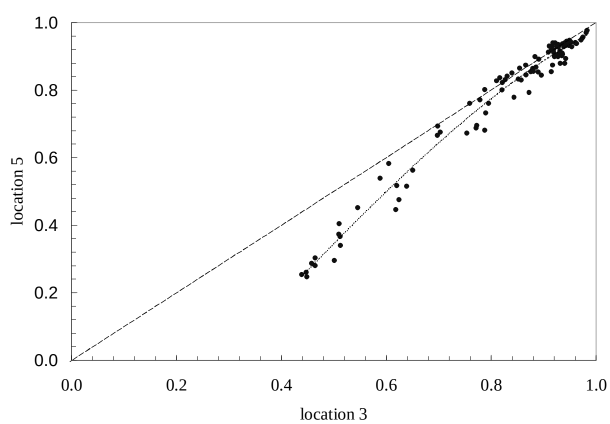

Figure 23. A correlative graph

There are several properties associated with this type of graph. First, a trend line can be made by a regression algorithm (such as an order-3 polynomial curve in Figure 23) to demonstrate the relationship between any two dimensions. For instance, a linear trend line shows strong correlation between two dimensions. Second, there can be a “line of equivalence” drawn for each of the two-dimensional graphs, $f(x) = y$. This means that the corresponding measurements (from input data) of dimensions X and Y are always equal. If a dot is found below this line, it indicates the input dimension represented by the X-axis is measured less than the input dimension represented by the Y-axis.

Figure 23 is a correlative graph for locations 3 and 5 for glaucoma patients or suspects, where the dots are found around a near-linear trend line, and mostly below the dashed “line of equivalence.” Two interesting features can be observed from this plot. First, the motion sensitivity measurements from location 3 and location 5 are strongly correlated. This feature is also visible by comparing the similarity between the physical-meaning maps drawn on location 3 and 5 in Figure 22. It means that the two testing locations are more likely to be affected by the disease at the same time, though to different levels. Second, the sensitivity of location 5 is lower than that of location 3 when sensitivity values of both locations are lower (e.g., both are below 0.7). This indicates that location 5 is mostly worse than location 3 for this disease. These features will be discussed in further detail in Section 7.2.

Using these visualisation techniques, correlation among all testing locations is studied. In general, the features demonstrated by neural networks are considered *low level* [139]. A basic proposition is that these low level features inherent in the data, when combined with the knowledge and judgement of human experts, would be promoted to *high level* knowledge useful for problem solving or decision making. Therefore, visualisation methods for presenting features in a human-oriented form are vital in assisting the interactive knowledge discovery [139]. In Section 7.2 I will discuss how knowledge is inducted and conclusions are drawn based on the analysis.

7.1.4 Knowledge Analyst

The knowledge analyst, who plays a central role in interactive knowledge discovery, needs to perform several important tasks. First, initial hypotheses regarding what to explore in the data need to be formed using the domain knowledge. Second, the appropriate input data to the SOM need to be categorised and grouped, based on age, clinical information, etc. Third, the features displayed by the visualisation techniques have to be organised, observed, and analysed. Fourth, the knowledge analyst needs to decide whether more specific data selection or other alternatives should be made to train the maps again, or whether something interesting has been discovered. This interactive procedure is carried out until a full picture of the test and its associated diagnostic knowledge has been drawn, and validated if possible.

One of the designers of the CCVP test has fulfilled the role of knowledge analyst. In collaboration with other researchers, he effectively performed those tasks using his experience and fundamental knowledge about ophthalmological anatomy. As a result, much useful knowledge regarding the CCVP test characteristics and its relevance for screening optic nerve diseases such as glaucoma and onchocerciasis has been established, which is described in the following sections.

7.1.5 Evaluation Strategy and Results

In this section, the proposed knowledge discovery method is evaluated. Again, I use the concept of *reproducibility* to verify the knowledge discovered, i.e. to see whether the features produced from one group of data are reproduced by another group. Here it assumes that true features are more reproducible than false ones. I shall also use this concept to measure the significance of noise elimination.

I have chosen 431 CCVP data (263 from right eyes and 168 from left eyes) from patients with early glaucoma and 1585 CCVP data (837 from right eyes and 748 from left eyes) from onchocercal populations to experiment. The evaluation strategy is that the features found from the left eye group are verified using those from the right eye group. This is assumed based on the symmetry between the two eyes suggested by ophthalmic anatomy. The results, as in Table 4, show the percentage of agreement between features discovered from both eyes with respect to the two patient groups.

Noise deletion	*Before*	*After*
Glaucoma	20%	66%
Onchocerciasis	53%	100%

Table 4. Rates of feature agreement before and after noise deletion

One can apparently observe that features discovered are scarcely reproducible before the deletion of noise, and the agreement is considerably improved after the noise is deleted. Therefore, the proposed knowledge discovery process with outlier management capabilities does seem to provide a good way of detecting useful features from noisy data.

7.2 Problem-Solving Applications

Using the above-proposed method, different kinds of previously unknown knowledge has been discovered, though some were suspected to be the case. Here I will present results on how certain pieces of knowledge are obtained using the proposed method and discuss their implications.

7.2.1 Discovering Known Knowledge for CCVP Validation

This section describes how the process of knowledge discovery from noisy data has helped to validate the CCVP as an optic nerve disease screening tool. The validation strategy is to compare the disease patterns with those from well-known conventional instruments.

I have experimented with the 431 clinical CCVP test records from patients with glaucoma. Using the methodology described in Section 7.1, I have focused on the relationship for motion sensitivity measurements among all testing locations.

Figure 24 is another correlative graph between testing location 3 and 4, which shows little correlation between the two locations. By comparing it with Figure 23, two of the features are evident: 1) the motion sensitivity measurements from location 3 are much more correlated to location 5 than to location 4; 2) the motion sensitivity measurements from location 3 are higher than those from location 5, when the motion sensitivity is low (e.g., below 0.7). I have carried out a systematic study for the correlative graphs between two of all testing locations in such a way [32]. Three discoveries were confirmed: 1) There is strong correlation among locations 1, 3, 5 (the "upper hemifield"). 2) Locations 2, 4, 6 (the "lower hemifield") are also strongly correlated. 3) The lower hemifield tends to be affected more severely than the upper hemifield.

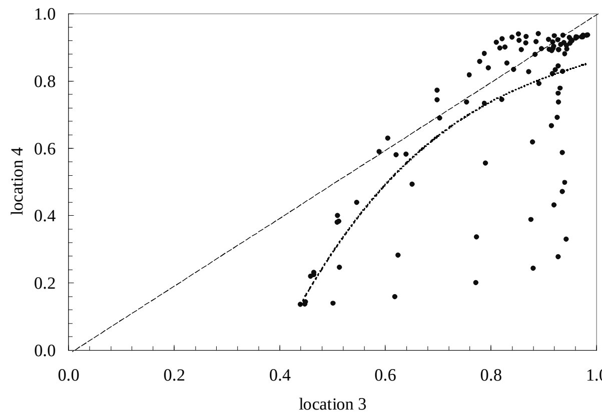

Figure 24. A correlative graph between testing location 3 and 4

The three discoveries make a strong argument that the CCVP is effective at detecting early glaucoma, since the discoveries are compatible with existing knowledge about glaucoma from past studies. In fact, studies carried out based on conventional perimetry have shown the unbalanced horizontal hemifields (retinal locations 1, 3, and 5 versus 2, 4, and 6, refer to Figure 2) as an early sign of glaucoma [73]. Additionally, anatomical [66] and clinical [155] evidence also supports the phenomenon of unbalanced hemifields with the more progressive lower hemifield. As a result, this evidence has not only validated the CCVP as an effective glaucoma screening tool, but more importantly in my research, it has validated the proposed knowledge discovery process.

7.2.2 Discovering New Knowledge from CCVP Data

This section will discuss how the knowledge discovery process has led to learning about lesser-known diseases from the CCVP data. Though onchocerciasis has been a severe threat for blindness, little knowledge has been established, compared to glaucoma, with regard to how the disease affects visual functions. With the help of the WHO project, ophthalmologic specialists were able to collect a large quantity of the CCVP data from onchocercal populations. Based on these data, the proposed knowledge discovery process was able to reveal new knowledge about this disease.

I have applied similar analysis discussed above to onchocerciasis data – 1585 test records collected in West Africa. A strong correlation between certain pairs of retinal locations has been found, but this time crossing the two horizontal hemifields, i.e. between locations 3 and 4, and locations 1 and 2 (refer to Figure 2). This discovery is in conflict with the belief that onchocerciasis and glaucoma are similar in terms of relationship among damage at retinal locations [16], since a strong correlation among horizontal hemifields is known for glaucoma. Therefore, this new discovery has drawn ophthalmologists' attention. To check the validity of this finding, they have conducted a topographic analysis of clinical chorioretinal changes related to the visual sensitivity at these testing locations. This is done by ophthalmologists examining the pathological condition of the retina of each selected patient. The study has confirmed the finding [187] and the encouraging result has further shown the effectiveness of the proposed method.

In addition, since the reactions of stimuli reflect the physiological condition of the eye, it is possible to find out how the retina and optic nerves are damaged in these patients by looking at relationships between testing locations in the retina. The results of my study have provided clinicians with anther way of finding out the cause and mechanisms of these optic nerve diseases.

7.3 Concluding Remarks

In this chapter, I have demonstrated how the proposed noisy management methods are applied to real world application for practical problem-solving tasks. An interactive method is proposed for knowledge discovery from clinical data. It involves managing outliers (filtering noise), abstracting and visualising features, and the knowledge analyst, who guides the heuristic process in an interactive manner. The process has been applied to large quantities of raw visual field test data. As a result, much useful knowledge regarding the CCVP test and its relevance for screening optic nerve diseases such as glaucoma and onchocerciasis has been established.

The knowledge discovered from the CCVP glaucomatous data is consistent with what was known from conventional perimetry, and therefore the finding has helped to validate the CCVP as an effective visual screening tool. The other type of discovery involving a lesser-known disease has improved the understanding of the disease. In particular, I use the concept of repeatability to see whether the features produced from one group of data are reproduced from another group, based on the assumption that true features are more repeatable than false ones. The results have shown that noise deletion (outlier analysis) is a crucial first step in discovering useful knowledge from noisy clinical data.

Much of the work on knowledge acquisition from noisy data has been on the use of statistical techniques to acquire classificatory knowledge. Notable examples include the TDIDT (Top Down Induction of Decision Trees) family of learning systems using the chi-square test with various pruning techniques [141, 20]. There have been also attempts to use neural networks to acquire knowledge from noisy data. Some of the most relevant work is Ultsch [165, 166], which uses unsupervised neural networks to find regularities from data. An ID3-like induction method is used to transform those regularities into classificatory knowledge.

There are two key differences in proposed approach. First, an effort is made to eliminate noise before data are used to discover features. I have shown that explicit treatment of noise is a necessary step to ensure that discovered features truly reflect data relationships. As discussed in Section 7.1.5, the features found from the raw glaucoma data can rarely be reproduced and this suggests that they have noise. Second, an interactive method, rather than an automatic process as in Ultsch's approach, has been used. Ultsch correctly pointed out that the "emergence" ability of the SOM is critical in extracting underlying knowledge in data [164]. The "emergent features," however, still rely on visual examination and interpretation of human beings, unless the type of knowledge is pre-determined. Ultsch's method uses neural networks to learn from the data, and classificatory knowledge is automatically generated from the weight parameters using an inductive algorithm. The role of knowledge analyst can therefore be taken by machine learning strategies. In this application, the type of knowledge to be discovered is unknown, and we are interested in finding out any information related to a particular group of subjects. As a result, the method attempts to combine the best capabilities of human and machine: using human knowledge and judgement to guide the knowledge acquisition process, but relying on the SOM to search in the data, and abstract information in a more representative way.

Chapter 8 Intelligent Data Analysis for Public Health: Self-Screening for Eye Diseases

This chapter will present my IDA research in the entire application context – self-screening for eye diseases. The application, as part of the METRO project (see Chapter 6), is to establish a software system that enables self-examination for public use [107, 189]. The CCVP visual test was first developed by Dr Wu in his initial study [182]. Since then, a number of developments have been made to address the data quality issue, which have been described from Chapter 3 to Chapter 7. I have integrated those improvements into an application with Dr Wu's CCVP program, to make it a more reliable and more acceptable self-screening tool for eye diseases. The integrated system was put to several trial runs and has received encouraging feedback. At the end of the chapter, the system with field application results will be evaluated. As well, I have systematically studied the key issues involved in evaluating software quality and carried out the evaluations with different strategies.

8.1 Self-Screening for Eye Diseases

Medical diagnostic tests give doctors important clinical information. Such tests normally require a specially designed device and a medical operator's attention; a doctor then analyses the test results. These testing devices are often very expensive and not widely available for mass screening – for example, testing in large-scale epidemic investigations, general practitioner clinics, and public halls. To make these tests easily accessible to the community, thus improving recognition of early signs of disease, one must seek an economical way of providing the tests without compromising reliability.

Developing a test capable of mass screening involves three important issues: *delivery*, *interface*, and *interpretation* [107]. First, the test should be easy to *deliver* without too much cost or compromise in the test reliability. Second, the test *interface* should adequately support the subject undertaking the test, preferably providing a self-testing environment where no instructions from a medical operator are necessary. Third, the test system should be able to interpret the test results, to give the subject a general warning of possible problems, without a doctor's involvement.

For many years, psychophysical researchers, influenced by the pioneer work of M. Flocks and his colleagues [50], have been trying to apply CRT and LCD technology as stimulus display devices for visual-function testing. Flocks' group and his followers partially addressed the test delivery issue – the use of television for eye testing lets many ordinary people perform the test without additional cost. However, this idea was not properly developed and realised, largely because human behavioural variants during the testing made obtaining reliable test results difficult, especially in a self-testing environment. The collected data contained much noise, making the interpretation of these results particularly difficult. Also, such an approach did not address the issue of the interface between the subject and the machine; communication was essentially one-way. Finally, the machine did not record the test results, and interpretation of the data recorded on paper was tedious.

More recently, the CCVP has demonstrated early success in detecting damage in optic nerves [182]. Like Flocks' work, the CCVP partially addresses the delivery issue by using PCs as testing machines, and the data collected from individual subjects also contain much measurement noise. Unlike Flocks' work however, the software-based approach addresses the interface issue: communication between the subject and the PC is two-way. Moreover, on-site interpretation of test results becomes possible.

In the last decade, a self-screening test has been developed based on CCPV, in which PCs without specialised hardware can examine visual functions. To address the above three issues, the test system has three main intelligent components: machine-learning programs, an intelligent user interface, and an interactive knowledge-discovery process. Operating on portable or desktop computers, this system has been used in several different public environments. Managing data quality is one of the major developments in this effort, as we shall see from the following discussions. A number of IDA methods have been integrated into the system, which contributed to the success of this application. This demonstrates how the integrated IDA techniques help in real world applications.

8.2 AI for Self-Screening Eye Diseases

To develop an effective self-screening system, various AI methods have been integrated into the CCVP. Three major software components have made it possible for the transformation from a standard perimetry system to a self-screening system. First, the self-screening system uses software-controlled perimetry that operates on PCs, where the proposed IDA methods are incorporated to help clean measurement noise and maintain reliability of the tests. Second, instead of having an experienced medical operator constantly monitoring the subject during the test, the system incorporates an intelligent user interface (see Chapter 6), which allows helpful interaction between the subject and the test. Third, rather than asking a doctor to make sense of the test results, the system attempts to provide the subject with important information regarding their visual functions. This allows pre-screening of the suspects with potential eye defects so that they can be referred to ophthalmologists for further diagnosis.

8.2.1 Reliability in Software-Based Perimetry

The development of the software-based perimetry based on the CCVP has been discussed in the previous chapters. The most significant problem with software-based visual field tests is that there is no longer dedicated hardware capable of monitoring the behaviour of the subject and providing reliability indicators. In addition, there is no longer the luxury of a standard test environment where important factors such as light and viewing distance can be strictly controlled. Therefore, software-based tests will naturally be less reliable in that the data collected from subjects will contain more measurement noise, which has been a key issue in the project. Computational methods have been developed based on a variety of IDA techniques, among which outlier management plays a central role. It essentially involves three phases.

1) Collecting data more reliably and reducing noise at source
2) Modelling test behaviour by visualising data and detecting outliers
3) Separating noise from real measurements among outliers

Phase 1) has been discussed in Chapter 6; phase 2) have been detailed in Chapter 3. Phase 3), described in Chapter 4 and Chapter 5, is achieved by modelling either noise or real measurements, depending on the type of information and domain knowledge available.

8.2.2 Interactive Knowledge Discovery for Pre-Screening

At the end of a self-testing, subjects are naturally interested in knowing whether their vision is normal. The average motion sensitivity values for different testing locations give an overall assessment of how well the subject has done. However, an indication of the likely diagnosis based on these sensitivity values and the relationships between different test locations would be desirable. For example, might the subject have a certain disease, if the sensitivity values for a subject are low for testing locations 3 and 4 (see Figure 2) but high for other testing locations?

To determine which test results indicate possible eye diseases, several thousand clinical test records were compiled from patients with vision loss from glaucoma and optic neuritis. The records were then analysed them using interactive data exploration where a data analyst steered the discovery process [30].

The analyst performed three important tasks. First, the analyst used relevant domain knowledge to form initial hypotheses regarding what to explore in the data. Second, the analyst organised different data sets according to various criteria such as disease types and sampling considerations before using those sets to extract the behavioural relationships between test locations. Third, after the relevant features were extracted by neural networks and were displayed by visualisation techniques, the analyst organised, observed and analysed them. This was very much an interactive and iterative process where the analyst made a number of decisions, e.g., whether to select more specific data or to take alternative action, whether any patterns detected had any significant meanings, and whether they could be validated.

The detailed findings of the above analysis were discussed in Chapter 7. The discovered disease patterns are then used to provide warning messages to the subject regarding the possible danger.

8.2.3 The Integrated IDA System

As pointed out previously, there are many areas where efforts have been made to improve data quality for the CCVP. Therefore it would be not only desirable, but also natural to facilitate the efforts by incorporating these efforts into an integrated system. A fully automated system would have not only a least resource requirement, but also a minimum chance of human errors. Such a system, with improved reliability, self-screening capability and better user acceptance, would be then put to the test for public health screening.

The architecture of such an integrated system that I have implemented is illustrated in Figure 25. The integrated system comprises a number of components:

- The computer controlled perimetry, taking advantage of processing power facilitating advanced AI and statistical techniques, and the mobile and inexpensive features from notebook computers
- The stability analyser that estimates data quality for each test as described in Section 6.3.1
- The control unit that decides whether the observer needs to have a follow-up test using the strategy described in Section 6.3.2
- An intelligent user interface that facilitates self-screening capabilities and reduces undesirable behavioural effects
- A database containing patients' personal information as well as test results, which makes it particularly easy for pre-test data entry and historic testing data retrieval (Figure 26)
- A visual function reporting facility that automatically generates a test result after each test, or presents past results retrieved from the database

- The knowledge discovery process, though not fully automated, that can provide disease patterns, therefore facilitating automated pre-screening on a large quantity of test data

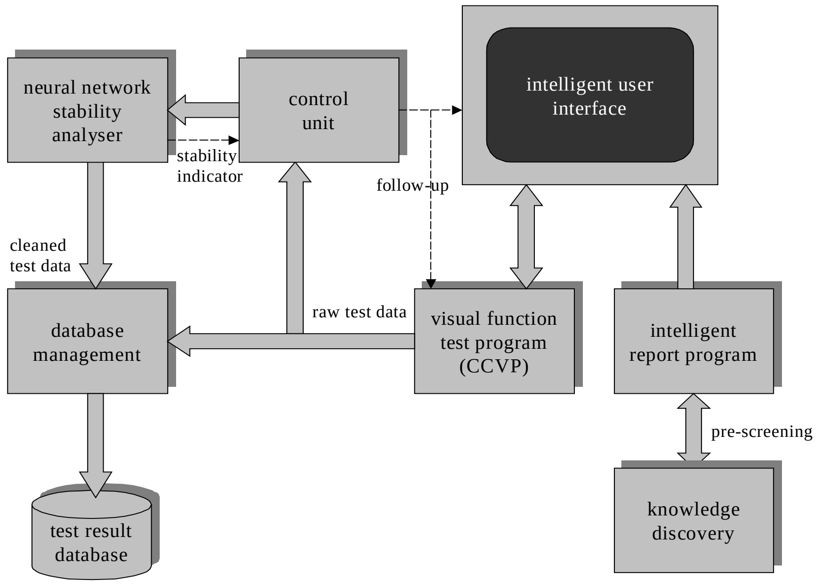

Figure 25. Architecture of the integrated system

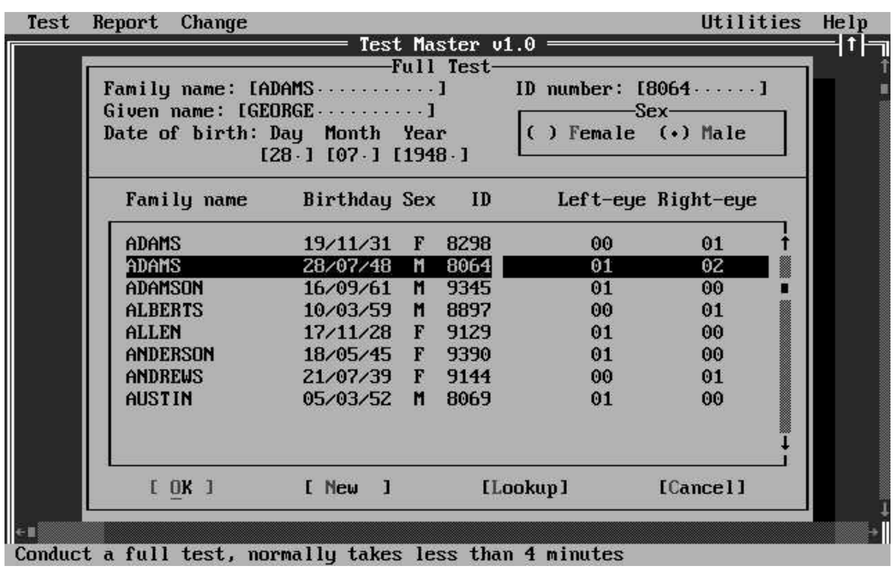

Figure 26. The user interface for pre-test data entry, connected to a database

8.3 Self-Screening in the Community

The self-screening system has been used in a World Health Organisation (WHO) program for preventing optic neuritis [29], in occupational health screening [177], and most recently, in a Medical Research Council (MRC) pilot study to detect people with glaucoma [107, 188]. These studies have been conducted to evaluate the test's performance and acceptance (i.e. how many people agree to take the test) in different primary care settings.

In the WHO study – mass screening for vision loss due to optic neuritis – the subjects were from a farming community in Africa. These subjects were largely computer-illiterate; the test was conducted in farmers' houses, using portable PCs. In the occupational health study – screening for ocular abnormalities and vision defects – the subjects were the employees of a large telecommunications company. These subjects were young (the average age was about thirty) and healthy, and they used computers daily; the test was performed in the company's offices using PCs connected by a local-area network. In the glaucoma study, the subjects were visitors of a general practice, and the test was conducted in the waiting room, using a desktop PC. Each study examined a large number of subjects and reported a good acceptance of the test.

Looking at the third of these studies in more detail, the test was offered during routine attendance at a large urban general practice in North London and was conducted by subjects themselves in the waiting room. For a three-month period during the pilot study, all patients aged forty or older who routinely attended the practice were offered the test. Upon entering the clinic, each patient received an information sheet explaining the purpose of the pilot study, the test and the nature of glaucoma, what to expect during and after the test, and information about whom to contact if they wished to know more about the test in general or were concerned about their own results. Each interested patient then signed a consent form.

Of the 925 people tested during the three-month period, 33 failed the test, i.e. the results indicated an abnormality in their visual field. These 33 people, together with 45 of those who passed the test were chosen as the “control group,” and were later assessed clinically in the practice by an ophthalmologist, who had no knowledge about the CCVP and its test results.

Figure 27 summarises the results. The overall picture is clear: the “failed group” had many more eye problems than the “control group.” For example, none in the “control group” was a confirmed glaucoma case, and 70% in the “control group” had a normal visual field. On the other hand, 82% in the “failed group” had various visual defects, including 34% confirmed glaucoma cases and 9% glaucoma suspects.[7] These findings are particularly encouraging, as an overwhelming majority of the subjects did not consider the possibility of having any eye problems when they visited the clinic. This opportunistic test has shown a high detection rate and allowed for the early detection of eye diseases such as glaucoma.

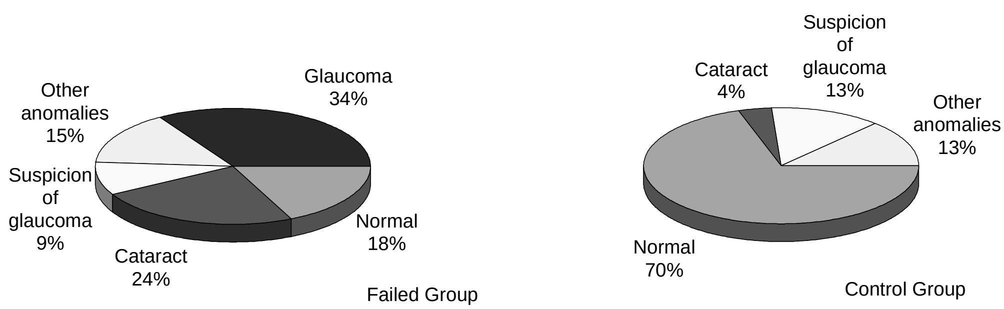

Figure 27. Distributions of clinical diagnosis by an ophthalmologist: (left) subjects who failed the test (the failed group); (right) selected subjects who passed the test (the control group)

Acceptance is also one of the most important issues in evaluating a screening test or an opportunistic test. Although little was done to increase the number of patients participating (no advertising, and little stimulation from the clinic staff), this study showed higher-than-expected acceptance: of the 1,215 people who were offered the test, 925 (76%) accepted. Surveyed by the same study, the acceptance rate was only 58% if a free eye test could be offered by an ophthalmologist in an eye hospital [188]. This is an encouraging finding in such an elderly population during a very short period, as the acceptance of opportunistic tests in city practices generally ranges from 50% to 70% [78].

[7] Those who had an increased intraocular pressure or whose optic disc appeared abnormal, but who had no visual-field loss.

The study has concluded that testing patients in the GP's waiting room offers a good opportunity to screen patients effectively and with minimal additional cost. Moreover, it provides patients with a meaningful alternative to the usual activities in the waiting room – reading newspapers and magazines, or chatting with others.

8.4 Software System Evaluation

A sensible evaluation of AI systems or software systems in general has always been a challenging research issue [10, 90, 27]. A careful assessment of such systems in laboratory environments is important but is no substitute for testing them in the real-world environments that they are developed for. This is especially important in medical informatics applications where their use in clinical environments is vital [190]. In the above section, I give examples of using the IDA-based self-screening system in different clinical environments. In this section, I systematically study various characteristics affecting software quality. Clinical data collected from laboratory-based and field-based investigations in different communities were used to analyse those aspects.

There has been much research carried out on how to evaluate software systems. In particular, there is an ISO document, which specifies various software evaluation characteristics, including functionality, reliability, usability, and efficiency, etc. [83]. I shall focus on strategies in evaluating functionality, reliability, and efficiency, as they are the most important characteristics for screening applications.

8.4.1 Functionality

This refers to a set of functions that satisfy the stated needs for an application. For the screening application, it means the discriminating power – maximising the detection rate while minimising false alarms. To serve this purpose, ROC analysis [65] is considered the most direct assessment method. ROC curves are drawn to assess a test's diagnostic performance by displaying pairs of detection rates and false alarm rates throughout the whole range of a test's measurements [74]. I have applied ROC analysis to the data collected by the screening system [29, 188], and evaluated the functionality of the self-screening system [27]. Figure 28 is an example of such a ROC analysis result from the WHO study.

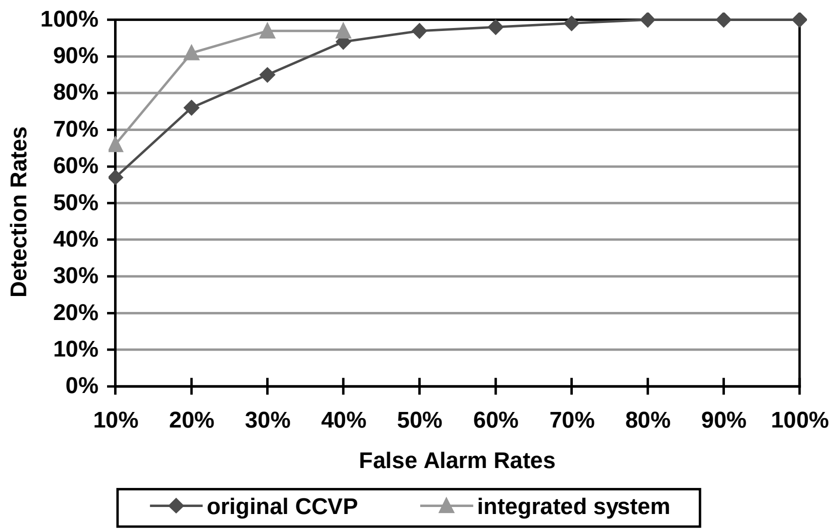

Figure 28. Detection rates versus false alarm rates for the self-screening system: the WHO study

8.4.2 Reliability

This is defined as the capability of software to maintain its level of performance under the stated conditions for a stated period. One of the most important criteria for screening applications is how reliably the data collected by the system reflect a subject's visual functions or impairment. I have proposed the following two criteria for measuring such reliability.

The consistency between the follow-up test results from the same subject. Since typical optic nerve diseases such as glaucoma and optic neuritis are generally long-term progressive diseases, patients' visual function should change little during a few weeks or even several months. Therefore test results from within a short time frame should be close, though under real clinical situations this is not always true since measurement noise is normally involved in each test. Naturally, the agreement between two tests repeated on the same individual within a short period can be a good indicator for the reliability of the test. As an example, Chapter 4 (see Figure 16) provides one such agreement analysis.

The agreement between disease patterns discovered from this application and those from other established visual functional tests. A reasonable assumption is that the detected patterns of impaired vision should be similar to those from conventional testing instruments, or can be verified by clinical or anatomical analysis. With this assumption, it has been able to discover compatible patterns for glaucomatous impairment found by conventional perimetry [30]. The findings in the WHO study were also successfully verified by a clinical analysis [187].

8.4.3 Efficiency

Efficiency is concerned with the relationship between the level of performance of the software and the amount of resources used. In the screening context, this is about how to minimise the amount of time a subject has to spend on a visit, while maintaining the quality of the test results. The system uses the following two ways to determine the testing duration.

Controlling the number of follow-ups if the previous test was unstable. I have suggested a strategy to decide whether an immediate follow-up test is required in order to obtain more reliable measurements, as described in Chapter 6. Two criteria were used to evaluate how efficient this strategy is: first, how many tests are so unstable that outlier management strategies will not work (the class of "totally unreliable" described in Chapter 3), and among them how many become more reliable in their follow-up tests; second, how the overall quality of test results improves when following the recommendations made by the system.

Using a dynamic number of repeated testing cycles during a single test. A test with a dynamic number of repeating cycles, or "dynamic testing" in short, has been devised by Cho [34] to find out the minimal number of repeating measurement cycles for an individual test, while maintaining the quality of the test results. The measurements obtained by the tests with a fixed number of cycles were used as the "golden standard" to evaluate how dynamic testing performs.

8.5 Concluding Remarks

Public health care is the collective action taken by society to protect and promote the health of entire populations [12]. Medical screening, with its focus on the prevention of disease at the population level, is one of the most important tools contributing to public health care.

A mass screening test should be easy to deliver at low cost while maintaining reliability; it should be easy to use while providing a supportive and interesting test interface; and it should offer early, appropriate warning messages for those who fail the test. It has explored the use of the computer as an affordable test machine and examined several challenging issues involved in a software-based test, such as the handling of data quality problems, the provision of intelligent assistance to the subject under examination and the understanding of patterns in the test data. It is found that AI has played a crucial role in managing these issues. Machine learning programs, an intelligent user interface and a knowledge discovery process are the major AI components of a self-screening system, which has found successful applications in different primary care settings with high acceptability.

As computers become ever more accessible, software-based tests are an obvious choice for mass screening in the community without additional hardware costs. Although most general practices in the UK have PCs, no solid evidence indicates that this has led to significant clinical improvement [191]. One of the problems is the lack of clinically orientated software for general practitioners. By making such software available in GP clinics, we could efficiently use existing computing resources to provide additional medical care to the community at minimal capital cost.

The integrated system, which incorporates most of the developments, has been used under diversifying health-care environments. It has so far demonstrated that reliable data from volatile test environments can be obtained, and the acceptability of the integrated system is also encouraging.

I have found it very important to carefully analyse various factors affecting the system evaluation, including evaluation objectives, target populations, related operating constraints, and what and how data should be collected. The feedback from early system trials had contributed in no small part to the continuing refinement of the screening system. I have adopted a systematic way of evaluating different aspects of the screening system.

Finally, this is a truly interdisciplinary project in which community health experts, computer scientists, epidemiologists, eye specialists, general practitioners, and ophthalmic nurses have worked together in identifying the system requirements, designing and implementing the system, testing the system in different operating environments, analysing the data collected, and continuously refining the software. Collaboration from the trial communities is also vital as this is directly related to the quality of data collected. Currently the test is being used in more health-care environments and the data collected will allow further improvement of the system.

Chapter 9 General Conclusions and Further Research

One of the most difficult and costly tasks in intelligent data analysis is trying to extract clean and reliable data, and many have estimated that as much as 50% to 70% of a project's effort is typically spent on this part of the process. It is the general objective of this dissertation to conduct a systematic investigation into the handling of outliers, a difficult but important topic for the following reasons. First, outliers can have a considerable influence on the results of an analysis [4, 8]. Second, although outliers are often measurement or recording errors, some of them can represent phenomena of interest, something significant from the viewpoint of the application domain. Third, for many applications, exceptions identified may lead to the discovery of unexpected knowledge.

The statistical literature is full of methods for detecting outliers in different data sets, but is short on methods for automatically distinguishing between different types of outliers. Such a method would require the use of relevant domain knowledge. In this study, I have developed novel ways of using such knowledge to reason about outliers, and to improve data quality.

In this dissertation, the different types of outliers in problem-solving contexts have been discussed, and the statistical treatment of outliers has been reviewed. The need for reasoning with outliers using relevant domain knowledge has been described in the context of intelligent data analysis. For this purpose, I have introduced a large-scale medical application for screening eye diseases, involving thousands of subjects in different field studies. It is ideal for the study of outliers since the effective management of outliers in test data is essential to the success of the application. I have shown how different outlier management strategies have been effectively applied to a variety of data sets collected from this application. I have also explored interactive methods for knowledge discovery from noisy data, which has demonstrated how the "cleaned data" have helped real-world problem-solving tasks. As a result, previous unknown knowledge has been discovered from clinical data, which has helped ophthalmologists validate their tests as well as further their understanding of diseases such as onchocerciasis.

My research on outlier management was initially motivated by a challenging medical application. However, the proposed methodology is sufficiently general to be applied to other types of applications, as discussed in Section 2.4. I have found that AI modelling techniques, when properly integrated, have great potential in automating the knowledge-based outlier analysis process.

9.1 An IDA Approach to Outlier Management

As summarised above, I have developed a systematic approach to the related issues in the entire process of outlier management:

1) Reducing noise at source
2) Detecting outliers using the SOM
3) Reasoning about outliers using domain-specific knowledge
4) Using the better quality data to solve domain problems

Although much of the approach is generic and can be applied to different applications (see Section 2.4), this approach has been carefully applied to a class of medical applications where the management of outliers is important to the success of the applications.

In the second step, Kohonen's SOM has been suggested as an effective way of visualising and learning characteristics among data, and has been found suitable in detecting outliers in visual field data. In particular, individual patients' behaviour during the test can be visualised, which offers an interesting way of understanding the test data. Early analysis results suggest that the SOM appears to be capable of showing when, where, and to a large extent, why outliers occur in the data. However, this method alone is not able to help distinguish the outliers.

Further research has been conducted on how to analyse outliers using domain knowledge for the third step. Two general strategies have been proposed to distinguish between "good" and “bad” outliers. The first strategy attempted to identify “good” outliers by modelling how “normal data” should be distributed in a domain of interest, and rejected outliers that did not fall within the model. In this context, domain experts are involved in providing examples of different patterns representing different pathological conditions, and machine learning methods such as back-propagation are used to generalise over these examples of how patients manifest themselves on the test data. The positive results of applying this strategy to the visual screening data have been reported using the notion of “reproducibility.”

The next general strategy for the third step uses knowledge regarding the understanding of “bad” outliers, or measurement noise in this particular application. Here it attempts to model undesirable outliers instead, and accept data outside of the norms if it is not accounted for by this model. A system that uses this strategy has been found to perform significantly better at a diagnostic task than an equivalent approach that either utilises all data, or attempts to reject all non-normal values. This strategy is better suited to those applications where relevant knowledge about undesirable outliers can be obtained.

Realising that much of the noise may be reduced at the data collection phase, a considerable amount of effort has been made to address the issues in the first step above. An intelligent user interface has been developed to obtain measurements with better quality. The interface has directly targeted a number of root causes of the measurement noise. Additionally, the outlier detection method is used to assess the stability of the test results on-line. A rule-based control strategy has been derived from the domain expertise and the results of data exploration, so that follow-up tests can be conducted immediately afterwards, based on the stability indicator and the rules. Consequently, I have shown that higher quality data may be collected on-line, instead of having to work with very noisy data off-line.

One of the key objectives in reasoning about outliers is the discovery of useful knowledge in the relevant application area. As a result, I have proposed an interactive process for knowledge discovery from noisy clinical data, as in the fourth step. This process has been applied to several thousand test records and many interesting patterns were uncovered. For example, the test patterns corresponding to onchocerciasis discovered from the data were in conflict with the existing knowledge that onchocerciasis and glaucoma should have similar patterns. This has attracted the attention of ophthalmologists, who have subsequently confirmed this finding by examining the pathological condition of the retina of patients involved.

Finally, I have presented my research in the application context – self-screening for eye diseases. In this IDA system, I have integrated virtually all the IDA methods that have been developed: noise reduction techniques in data collection, outlier detection and analysis, noise filtering, and interactive knowledge discovery from noisy data. It is a demonstration of how a variety of IDA techniques can be integrated to address various requirements from an application. Based on the application, I have also presented the evaluation methods from two aspects: the results of the field application and the quality of the IDA software system in terms of functionality, reliability, and efficiency.

9.2 Summary of Contributions

The main contributions of this dissertation are the following.

- Realisation that IDA techniques are sufficiently mature for explicit outlier treatment. I have argued that the explicit treatment of outliers are necessary for some types of applications where different types of outliers contribute to the data analysis results in significantly different ways. Using IDA techniques, specially machine learning methods, it is able to show that different types of outliers can be distinguished with domain knowledge. It is hoped that this will help stimulate the work on outlier management to go beyond traditional outlier detection to outlier explanation.
- First application of the self-organising maps in visualising and detecting outliers (Chapter 3). I have demonstrated the special features of the SOM in visualising high dimensional data and detect outliers based on both of their mathematical distances and their statistical frequencies. I have also shown that the SOM is capable of exposing temporal behaviour of individuals. These two key features have made the

SOM-based outlier detection method a novel addition to the existing outlier detection methods.

- First automated knowledge-based approaches for distinguishing different types of outliers (Chapter 4 and Chapter 5). It has been long acknowledged that outliers may be of different nature and decisions of retaining or rejecting them should be made carefully. However, the existing work, almost without exception, has been suggesting analysts to perform such a task. In the thesis I have developed two methods to learn knowledge from domain experts and construct data models to analyse the outliers automatically.
- Introducing IDA techniques for outlier reduction at source as part of outlier management process (Chapter 6). I have shown that with various IDA techniques, the quality of data may be improved before given to data analysts.
- Knowledge-discovery from noisy data (Chapter 7). I have helped the clinicians to discover new knowledge about some previously lesser-known disease. I have also helped verify the knowledge about some better-known disease by comparing it with traditional perimetry, so that it validates CCVP's diagnosis power. In particular, the study shows that the cleaned data can provide better quality knowledge, and this has provided a supporting example that outlier analysis is needed in this application.
- Successful real world applications in public health (Chapter 8). I have shown how the integrated IDA system, in which outlier management plays an important role, has helped to solve a challenging medical screening problem.
- Systematic evaluation of IDA systems (Section 8.4). The issue of how IDA systems should be evaluated has not received sufficient attention. This thesis has introduced a systematic way of evaluating an IDA system from the software quality perspective.

9.3 Further Work

Although much has been achieved as discussed above, further work has also been considered. The proposed outlier management strategies should be investigated more completely by using different AI and statistical methods, and by conducting careful comparative studies to establish the most appropriate methods connected to a particular type of application.

For example, the SOM was used as a basis for detecting outliers and for visualising behavioural data. However, the Generative Topographic Mapping (GTM) approach has recently been proposed and shown to have certain advantages over the SOM [18]. It would be interesting to find out whether GTM can help build a better outlier detection method.

Second, it would be desirable to explore different knowledge acquisition and representation techniques based on various situations in the application domain. My study has largely explored the instance-based machine learning methods to establish data models for outlier distinction, but there may be different form of knowledge available for other types of applications. For example, if the knowledge can be acquired as a set of rules from domain experts, then reasoning techniques associated with knowledge-based systems, such as Bayesian networks, may be a better candidate for investigation.

Finally, yet importantly, it is of great interest to have further analytical or experimental work on testing the proposed outlier strategies on other data sets in different fields, such as in science, engineering and economics. More extensive study should be carried out to address additional issues that might come up with applications of an entirely different nature, although my study presented here has shown itself capable of addressing most of the challenging issues of outlier management in this type of medical applications.

Publications Related to this Dissertation

A list of publications having the author of the thesis as a co-author and published during the studies on subjects discussed in this thesis.

1. Cheng, G., Cho, K., Liu, X., Loizou, G. and Wu, J. X., Evaluating an eye screening test, ***Advances in Intelligent Data Analysis*** (eds. D. Hand, J. Kok and M. Berthold), 461-471, Springer-Verlag (1999)
2. Cheng, G., Liu, X. and Wu, J., Interactive knowledge discovery through self-organising feature maps, ***Proceedings of the World Congress on Neural Networks***, 4:430-434, San Diego, California (1994)
3. Cheng, G., Liu, X., Wu, J. and Jones, B., Establishing a reliable visual function test and applying it to screening optic nerve disease in onchocercal communities, ***International Journal of Bio-Medical Computing***, 41:47-53 (1996)
4. Cheng, G., Liu, X., Wu, J., Jones, B. and Hitchings, R., Discovering knowledge from visual field data, ***Medical Informatics Europe '96***, 629-633, IOS Press (1996)
5. Cheng, G., Liu, X., Wu, J., Jones, B. and Hitchings, R., Knowledge discovery from noisy clinical data, ***Computational Medicine, Public Health, and Biotechnology***, 3:1217-1225, World Scientific (1995)
6. Cheng, G., Wu, J. X., Liu, X., Fitzke, F. and Hitchings, R., Pattern discovery of motion sensitivity loss in glaucoma using unsupervised neural networks, ***Investigative Ophthalmology & Visual Science***, 35(4):1305 (1994)
7. Cheng, G., Wu, J., Hitchings, R. and Poinooswamy, D., Long-term learning factor in Humphrey visual field test, ***Proceedings of the 11th International Perimetric Society Meeting***, Wasington DC (1994)
8. Liu X., Cheng, G. and Wu, J., AI for public health: self-screening for eye diseases, ***IEEE Intelligent Systems***, 13(5):28-35 (1998)
9. Liu, X. and Cheng, G., Analysing outliers by searching for plausible hypotheses, In ***Knowledge Discovery and Data Mining: Theory and Practice*** (ed. M. Bramer), The Institution of Electrical Engineers (1999)

10. Liu, X., Cheng G. and Wu, J., Analysing visual field data by self-organising maps, ***Intelligent Techniques and Soft Computing*** (ed. H-J. Zimmerman), 1435-1439, ELITE Foundation (1996)

11. Liu, X., Cheng, G. and Wu, J. X., Identifying the measurement noise in glaucomatous testing: an artificial neural network approach, ***Artificial Intelligence in Medicine***, 6:401-416 (1994)

12. Liu, X., Cheng, G. and Wu, J., A study of behavioural instability in visual function self-testing, ***Medical Informatics Europe '96, 13th International Congress***, IOS Press, 778-82 (1996)

13. Liu, X., Cheng, G. and Wu, J., Managing the noisy glaucomatous test data by self-organizing maps, ***Proceedings of IEEE International Conference on Neural Networks***, 649-652, Orlando, Florida (1994)

14. Liu, X., Cheng, G. and Wu, J., METRO: making eye test reliable for ophthalmologists, ***Proceedings of AAAI Spring Symposium on AI in Medicine***, 99-103 (1996)

15. Liu, X., Cheng, G. and Wu, J., Noise and uncertainty management in intelligent data modelling, ***Proceedings of the AAAI-94***, 263-268 (1994)

16. Liu, X., Johnson, R., Cheng, G., Swift, S. and Tucker, A., Soft computing for intelligent data analysis, ***Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society***, 527-531, IEEE Press, New York (1999)

17. Liu, X., Cheng, G., Wu, J., Analyzing outliers cautiously, ***IEEE Transactions on Knowledge and Data Engineering***, to appear (2001)

18. Wu, J. X., Cheng, G., Liu, X. and Hitchings, R., Assessment of the reliability in a psychophysical test by neural networks, ***Proceedings of the 1994 US Conference on Vision Science***, Optical Society of America, 2:159-162 (1994)

19. Wu, J., Cheng, G. and Liu, X., Reasoning about outliers by modelling noisy data, ***Advances in Intelligent Data Analysis: Reasoning about Data*** (eds. X. Liu, P. Cohen and M. Berthold), 549-558, Springer-Verlag (1997)

20. Wu, J., Cheng, G., Hitchings, R., Liu, X. and Fitzke, F., Identifying the measurement status for psychophysical testing in glaucoma, ***Proceedings of the 14th Annual British Glaucoma Conference***, London, S2:3 (1993)

21. Wu, J., Hitchings, R., Cheng, G. and Liu, X., Improved reliability estimates in the Humphrey visual field test in glaucoma using a knowledge based system, ***Proceedings of the 11th International Perimetric Society Meeting***, Washington DC (1994)

22. Wu, J., Hitchings, R., Jones, B., Cheng, G. and Liu, X., A novel method to use redundant data from visual field tests by neural networks, ***Proceedings of the 1st World Congress on Computational Medicine and Public Health***, Austin, Texas (1994)

23. Wu, J., Wormald, R., Hitchings, R., Liu, X. and Cheng, G., Software-oriented visual field test for glaucoma screening, ***Investigative Ophthalmology & Visual Science***, 39(4):58 (1998)

Bibliography

1. Abiose, A., Jones, B. R., Cousens, S. N., Murdoch, I., Cassels-Brown, A., *et al*, Reduction in incidence of optic nerve disease with annual ivermectin to control onchocerciasis, ***Lancet***, 341:130-134 (1993)
2. Agrawal, R., Imielinski, T. and Swami, A., Mining association rules between sets of items in large databases, ***Proceedings of the ACM-SIGMOD International Conference on Management of Data***, 207-216 (1993)
3. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A., Fast discovery of association rules, in ***Advances in Knowledge Discovery and Data Mining*** (eds. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy), 307-328, AAAI Press / MIT Press (1996)
4. Altman, D., ***Practical Statistics for Medical Research***, Chapman and Hall (1991)
5. Arning, A., Agrawal, R. and Raghavan, P., A linear method for deviation detection in large databases, ***Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*** (eds. E. Simoudis, J. Han and U. Fayyad), 164-169, AAAI Press (1996)
6. Baldi, P. and Brunak, S., ***Bioinformatics: The Machine Learning Approach***, The MIT Press (1998)
7. Bandemer, H. and Naether, W., ***Fuzzy Data Analysis***, Kluwer (1992)
8. Barnet, V. and Lewis, T., ***Outliers in Statistical Data***, John Wiley & Sons (1994)
9. Barnet, V., The ordering of multivariate data (with discussion), ***Journal of the Royal Statistics Society, Series A***, 139:318-354 (1976)
10. Basili, V., Selby, R. and Hutchens, D., Experimentation in software engineering, ***IEEE Transactions on Software Engineering***, SE-12:733-743 (1988)
11. Bauer, H. U. and Pawelzik, K. R., Quantifying the neighborhood preservation of self-organizing feature maps, ***IEEE Transactions on Neural Networks***, 3(4):570-579 (1992)
12. Beaglehole, R. and Bonita R., ***Public Health at the Crossroads***, Cambridge University Press (1997)
13. Bebie, H., Fankhauser, F., Spahr, J., Static perimetry: accuracy and fluctuations, ***Acta Ophthalmology***, 54:339-348 (1976)

14. Becker, S. and Hinton, G. E., Self-organizing neural network that discovers surfaces in random-dot stereograms, ***Nature*** 355:161-163 (1992)

15. Belsley, D. A., Kuh, E., and Welsch, R. E., ***Regression Diagnostics: Identifying Influential Data and Sources of Collinearity***, John Wiley (1980)

16. Bird, A. C., Sheikh, H., Anderson, J. and Fuglsang, H., Changes in visual function and in the posterior segment of the eye during treatment of onchocerciasis with diethylcarbamazine citrate, ***British Journal of Ophthalmology***, 64(3):191-200 (1980)

17. Birt, C. M., Shin, D. H., Samudrala, V., Hughes, B. A., Kim, C. and Lee, D., Analysis of reliability indices from Humphrey visual field tests in an urban glaucoma population, ***Ophthalmology***, 104(7):1126-1130 (1997)

18. Bishop, C. M., Svensén, M. and Williams, C., GTM: the generative topographic mapping, ***Neural Computation***, 10:215-234 (1998)

19. Brachman, R. J. and Anand, T., The process of knowledge discovery in databases: a human-centered approach, in ***Advances in Knowledge Discovery and Data Mining*** (eds. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy), 37-57, AAAI Press / MIT Press (1996)

20. Bratko, I. and Kononenko, I., Learning diagnostic rules from incomplete and noisy data, In ***Interactions in Artificial Intelligence and Statistical Methods*** (ed. B. Phelps), Technical, 142-153 (1987)

21. Breiman, L., Friedman, J., Olshen, R. and Stone, C., ***Classification and Regression Trees***, Wadsworth International (1984)

22. Brodley, C. and Friedl, M., Identifying and eliminating mislabeled training instances, ***Proceedings of AAAI-96***, 799-805 (1996)

23. Bryson, A. and Ho, Y-C., ***Applied Optimal Control***, Blaisdell, New York (1969)

24. Buntine, W., A guide to the literature on learning graphical models, ***IEEE Transactions on Knowledge and Data Engineering***, 8:195-210 (1996)

25. Caroni, C. and Prescott, P., Sequential application of Wilks's multivariate outlier test, ***Applied Statistics***, 41:355-364 (1992)

26. Cheeseman, P. *et al.*, AutoClass: a Bayesian classification system, ***Proceedings of the AAAI-88***, 607-611 (1988)

27. Cheng, G., Cho, K., Liu, X., Loizou, G. and Wu, J. X., Evaluating an eye screening test, ***Advances in Intelligent Data Analysis*** (eds. D. Hand, J. Kok and M. Berthold), 461-471, Springer-Verlag (1999)

28. Cheng, G., Liu, X. and Wu, J., Interactive knowledge discovery through self-organising feature maps, ***Proceedings of the World Congress on Neural Networks***, 4:430-434, San Diego, California (1994)

29. Cheng, G., Liu, X., Wu, J. and Jones, B., Establishing a reliable visual function test and applying it to screening optic nerve disease in onchocercal communities, ***International Journal of Bio-Medical Computing***, 41:47-53 (1996)

30. Cheng, G., Liu, X., Wu, J., Jones, B. and Hitchings, R., Discovering knowledge from visual field data, ***Medical Informatics Europe '96***, 629-633, IOS Press (1996)

31. Cheng, G., Liu, X., Wu, J., Jones, B. and Hitchings, R., Knowledge discovery from noisy clinical data, ***Computational Medicine, Public Health, and Biotechnology***, 3:1217-1225, World Scientific (1995)

32. Cheng, G., Wu, J. X., Liu, X., Fitzke, F. and Hitchings, R., Pattern discovery of motion sensitivity loss in glaucoma using unsupervised neural networks, ***Investigative Ophthalmology & Visual Science***, 35(4):1305 (1994)

33. Cho, K. W., Liu, X. and Loizou, G., Decision making in dynamic visual field testing by backpropagation: C4.5 and multiple regression, ***Proceedings of the 6th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems***, 401-406, Proyecto Sur de Ediciones, Granada, Spain (1996)

34. Cho, K. W., Liu, X., Loizou, G., and Wu, J. X., An AI approach to dynamic visual field testing. ***Computers and Biomedical Research***, 31:143-163 (1998)

35. Codd, E. F., ***Providing OLAP (On-Line Analytic Processing) to User-Analysts: An IT Mandate***, E F Codd and Associates (1994)

36. Cohen, P. R., ***Empirical Methods for Artificial Intelligence***, MIT Press (1995)

37. Collett, D. and Lewis, T., The subjective nature of outlier rejection procedures, ***Applied Statistics***, 25:228-237 (1976)

38. Collins, J., ***EYETEST: A Perimetry Test Developed to Aid the Detection of Glaucoma***, M.Sc. Dissertation, Computer Science, Birkbeck College, London (1993)

39. Cooper, G. and Herskovits, E., A Bayesian method for the induction of probabilistic networks from data, ***Machine Learning***, 9:309-347 (1992)

40. Cruz-Neira, C., Sandin, D. and DeFanti, T., Surround-screen projection-based virtual reality: the design and implementation of CAVE, ***Proceedings of SIGGRAPH '93 Computer Graphics Conference***, 135-142 (1993)

41. Dempster, A., Laird, D. and Rubin, D., Maximum Likelihood from Incomplete Data via the EM Algorithm, ***Journal of the Royal Statistical Society, Series B***, 39:1-38 (1977)

42. Duda, R. and Hart, P., ***Pattern Classification and Scene Analysis***, Wiley (1973)

43. Duda, R., Hart, P. and Nilsson, N., Subjective Bayesian methods for rule-based inference systems, ***AFIPS Conference Proceedings of the 1976 National Computer Conference***, 45:1075-1082 (1976)

44. Erwin, E., Obermayer, K. and Schulten, K., Self-organising maps: stationary states, metastability and convergence rate, ***Biological Cybernetics***, 67:35-45 (1992)

45. Everitt, B.S., ***Cluster Analysis***, Gower Publications (1993)

46. Favata, F. and Walker, R., A study of the application of Kohonen-type neural networks to the travelling salesman problem, ***Biological Cybernetics***, 64:463-468 (1991)

47. Feigenbaum, E., The art of artificial intelligence: I. themes and case studies of knowledge engineering, ***Proceedings of the 5th International Joint Conference on Artificial Intelligence***, 1014-1029 (1977)

48. Fisher, D. H., Pazzani, M. J. and Langley, P., ***Concept Formation: Knowledge and Experience in Unsupervised Learning***, Morgan Kaufmann (1991)

49. Fitzke, F. W., Poinoosawmy, D., Nagasubramanian S. and Hitchings, R. A., Peripheral displacement threshold in glaucoma and ocular hypertension, ***Perimetry Update 1988/89***, 399-405 (1989)

50. Flocks, M., Rosenthal, A. R., and Hopkins, J., Mass visual screening via television, ***Ophthalmology***, 85:1141-1149 (1978)

51. Fogel, L., Owens, A. and Walsh, M., ***Artificial Intelligence Through Simulated Evolution***, Wiley (1966)

52. Fox, J., ***Regression Diagnostics: An Introduction***, Sage Publications, Inc. (1991)

53. Friedman, N. and Goldszmidt, M., Building classifiers using Bayesian networks, ***Proceedings of the AAAI-96***, 1277-1284 (1996)

54. Geman, S. and Geman, D., Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 6:721-741 (1984)

55. Gnanadesikan, R. and Kettenring, J. R., Robust estimates, residuals and outlier detection with multi-response data, ***Biometrics***, 28:81-124 (1972)

56. Goldberg, D., ***Genetic Algorithms in Search, Optimization, and Machine Learning***, Addison-Wesley (1989)

57. Goonatilake, S. and Treleaven, P., ***Intelligent Systems for Finance and Business***, John Wiley & Sons (1995)

58. Grubbs, F. E., Procedures for detecting outlying observations in samples, ***Technometrics***, 11:1-21 (1969)

59. Grubbs, F. E., Sample criteria for testing outlying observations, ***Annuals of Mathematical Statistics***, 21:27-58 (1950)

60. Guyon, I. and Stork, D., Linear discriminant and support vector classifiers, in ***Advances in Large-Margin Classifiers*** (eds. A. Smola, P. J. Bartlett, B. Schölkopf and D. Schuurmans), 147-169, MIT Press (2000)

61. Guyon, I., Matic, N. and Vapnik, V., Discovering informative patterns and data cleaning, ***Proceedings of AAAI-94 Workshop on Knowledge Discovery in Databases***, 143-156 (1994)

62. Hair, J., Anderson, R., Tatham, R. and Black, W., ***Multivariate Data Analysis***, Prentice Hall International (1998)

63. Hand, D. J., ***Construction and Assessment of Classification Rules***, Wiley (1997)

64. Hand, D. J., Intelligent data analysis: issues and opportunities, ***Advances in Intelligent Data Analysis: Reasoning about Data*** (eds. X. Liu, P. Cohen and M. Berthold), 1-14, Springer-Verlag (1997)

65. Hanely, J. X. and McNeil, B. J., The meaning and use of the area under a receiver operator characteristic (ROC) curve. ***Radiology***, 143:29-36 (1982)

66. Harrington, D. O. and Drake, M. V., ***The Visual Fields Text and Atlas of Clinical Perimetry***, The CV Mosby Co. (1990)

67. Hawkins, D. M., ***Identification of Outliers***, Chapman and Hall, London (1980)

68. Hawkins, D. M., The detection of errors in multivariate data using principal components, ***Journal of American Statistics Association***, 69:340-344 (1974)

69. Hecht-Nielsen, R., ***Neurocomputing***, Addison-Wesley (1990)

70. Heckerman, D., Bayesian networks for data mining, ***Data Mining and Knowledge Discovery***, 1:79-119 (1997)

71. Heckerman, D., Geiger, D. and Chickering, M., ***Learning Bayesian Networks: the Combination of Knowledge and Statistical Data***, Technical Report MSR-TR-94-09, Microsoft Research (1994)

72. Heiji, A., Lindgren, G. and Olsson, J., Computer-assisted recognition of glaucomatous field loss, ***Archives of Ophthalmology***, 106:721-722 (1988)

73. Heijl, A., Some characteristics of glaucomatous visual field loss, ***Glaucoma Update IV***, Springer-Verlag (1991)

74. Henkelman, R. M., Kay, I., and Bronskill, M. J., Receiver operator characteristic (ROC) analysis without ground truth, ***Medical Decision Making***, 10:24-29 (1990)

75. Hinton, G. and Sejnowski, T., Learning and relearning in Boltzmann machines, in ***Parallel Distributed Processing*** (eds. D. Rumelhart and J. McClelland), 282-317, The MIT Press (1986)

76. Hitchings, R., Psychophysical testing in glaucoma, ***British Journal of Ophthalmology***, 77(8):471-472 (1993)

77. Hobbs, F. D., Cherry, R. C., Fielding, J. W., Pike, L. and Holder, R., Acceptability of opportunistic screening for occult gastrointestinal blood loss, ***British Medical Journal***, 304(6825):483-486 (1992)

78. Holland, J., ***Adaptation in Natural and Artificial Systems***, The MIT Press (1975)

79. Hopfield, J., Neurons with graded response have collective computational properties like those of two-state neurons, ***Proceedings of the National Academy of Sciences of the USA***, 79:2554-2558 (1982)

80. Hoskins, H. D. and Kass, A. M., ***Diagnosis and Therapy of the Glaucoma***, The CV Mosby Co. (1989)

81. Huber, P. J., ***Robust Statistics***, John Wiley & Sons (1981)

82. Hunter, L. and States, D., Bayesian classification of protein structure, ***IEEE Expert***, 7:67-75 (1992)

83. ***ISO International Standard, Information Technology – Software Product Evaluation – Quality Characteristics and Guidelines for Their Use***, ISO/IEC 9126 (1991)

84. Jensen, F., ***An Introduction to Bayesian Networks***, UCL Press (1996)

85. John, G. H., Robust decision trees: removing outliers from databases, ***Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*** (eds. U. Fayyad and R. Uthurusamy), 174-179, AAAI Press (1995)

86. Kaski, S., Kangas, J. and Kohonen, T., Bibliography of self-organizing map (SOM) papers: 1981–1997, ***Neural Computing Surveys***, 1:1-176 (1998)

87. Katz, J. and Sommer, A., Reliability indexes of automated perimetric tests, ***Archives of Ophthalmology***, 106:1252-1254 (1988)

88. Katz, J. and Sommer, A., Screening for glaucomatous visual field loss: the effect of patient reliability, ***Ophthalmology***, 97(8):1032-1037 (1990)

89. Keating, D., Mutlukan, E., Damato, B. E. and Evans, A. L., A neural network for the classification of visual field data, ***Investigative Ophthalmology & Visual Science***, 33:970 (1992)

90. Kitchenham, B., Towards a constructive quality model: software quality modelling, measurement and prediction, ***Software Engineering Journal***, 2:105-113 (1987)

91. Kleiner, B. and Hartigan, J. A., Representing points in many dimensions by trees and castles (with discussion), ***Journal of American Statistics Association***, 76:260-276 (1981)

92. Knorr, E. and Ng, R. A unified notion of outliers: properties and computation, ***Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining***, 219-222 (1997)

93. Kohonen, T. and Makisara, K., Representation of sensory information in self-organising feature maps, ***AIP Conference Proceedings 151, Neural Networks for Computing***, 271-276 (1986)

94. Kohonen, T., Makisara, K. and Saramaki, T., Phonotopic maps – insightful representation of phonological features for speech recognition, ***Proceedings of the 7th International Conference on Pattern Recognition***, Montreal (1984)

95. Kohonen, T., ***Self-Organization and Associative Memory***, Springer-Verlag (1989)

96. Kohonen, T., ***Self-Organizing Maps***, Springer-Verlag (1995)

97. Kohonen, T., The self-organising map, ***Proceedings of the IEEE***, 78:9:1464-1480 (1990)

98. Kolodner, J., ***Case-Based Reasoning***, Morgan Kauffman (1994)

99. Koza, J., ***Genetic Programming: On the Programming of Computers by Means of Natural Selection***, The MIT Press (1992)

100. Krakau, C. E. T., Artificial intelligence in computerised perimetry, ***Documenta Ophthalmologica Proceedings Series***, 49:169-174 (1987)

101. Krzanowski, W. and Marriott, F., ***Multivariate Analysis***, E. Arnold (1995)

102. Lakshminarayan, K., Samad, T. and Harp, S., ***Causal Modelling for Time Series***, Internal White Paper, Honeywell Technology Center, USA (1996)

103. Lauritzen, S. and Spiegelhalter, D., Local computations with probabilities on graphical structures and their applications to expert systems (with discussion), ***Journal of the Royal Statistical Society, Series B***, 50:157-224 (1988)

104. Lene, M., Martin-Boglind and Wanger, P., Computer-assisted evaluation of the results from high-pass resolution perimetry: a knowledge-based system, ***Perimetry Update 1990/91***, 297-301 (1991)

105. Lieberman, M.F. and Drake, M. V., ***Computerized Perimetry***, Slack Inc. (1992)

106. Little, R. and Rubin D., ***Statistical Analysis with Missing Data***, Wiley (1987)

107. Liu X., Cheng, G. and Wu, J., AI for public health: self-screening for eye diseases, ***IEEE Intelligent Systems***, 13(5):28-35 (1998)

108. Liu, X. and Cheng, G., Analysing outliers by searching for plausible hypotheses, In ***Knowledge Discovery and Data Mining: Theory and Practice*** (ed. M. Bramer), The Institution of Electrical Engineers (1999)

109. Liu, X., Cheng G. and Wu, J., Analysing visual field data by self-organising maps, ***Intelligent Techniques and Soft Computing*** (ed. H-J. Zimmerman), 1435-1439, ELITE Foundation (1996)

110. Liu, X., Cheng, G. and Wu, J. X., Identifying the measurement noise in glaucomatous testing: an artificial neural network approach, ***Artificial Intelligence in Medicine***, 6:401-416 (1994)

111. Liu, X., Cheng, G. and Wu, J., A study of behavioural instability in visual function self-testing, ***Medical Informatics Europe '96, 13th International Congress***, IOS Press, 778-82 (1996)

112. Liu, X., Cheng, G. and Wu, J., Managing the noisy glaucomatous test data by self-organizing maps, ***Proceedings of IEEE International Conference on Neural Networks***, 649-652, Orlando, Florida (1994)

113. Liu, X., Cheng, G. and Wu, J., METRO: making eye test reliable for ophthalmologists, ***Proceedings of AAAI Spring Symposium on AI in Medicine***, 99-103 (1996)

114. Liu, X., Cheng, G. and Wu, J., Noise and uncertainty management in intelligent data modelling, ***Proceedings of the AAAI-94***, 263-268 (1994)

115. Liu, X., Intelligent data analysis: issues and challenges, ***The Knowledge Engineering Review***, 11(4):365-371 (1996)

116. Liu, X., Johnson, R., Cheng, G., Swift, S. and Tucker, A., Soft computing for intelligent data analysis, ***Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society***, 527-531, IEEE Press, New York (1999)

117. Lo, Z-P and Bavarian, B., On the rate of convergence in topology preserving neural networks, ***Biological Cybernetics***, 65:55-63 (1991)

118. Lunn, A. and McNeil, D., ***Computer-Interactive Data Analysis***, Wiley (1991)

119. Madigan, D. and Raftery, A. E., Model selection and accounting for model uncertainty in graphical models using Occam's window, ***Journal of the American Statistical Association***, 89:1535-1546 (1994)

120. Mannila, H. and Raiha, K., On the complexity of inferring functional dependencies, ***Discrete Applied Mathematics***, 237-243 (1992)

121. Masand, B. M. and Piatetsky-Shapiro, G., A comparison of approaches for maximizing business payoff of prediction models, ***Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining***, 195-201 (1997)

122. Matic, N., Guyon, I., Bottou, L., Denker, J. and Vapnik, V., Computer aided cleaning of large databases for character recognition, ***Proceedings of the 11th International Conference on Pattern Recognition***, 2:330-333 (1992)

123. McCulloch, W. and Pitts, W., A logical calculus of the ideas immanent in nervous, ***Bulletin of Mathematical Biophysics***, 5:115-133 (1943)

124. Mills, R. P., Schulzer, M., Hopp, R. H. and Drance, S. M., Estimates of variance in visual field data, ***Perimetry Update 1988/89***, 93-101 (1989)

125. Minsky, M. and Papert, S., ***Perceptrons***, The MIT Press (1969)

126. Mitchell, M., ***An Introduction to Genetic Algorithms***, The MIT Press (1996)

127. Mitchell, T., ***Machine Learning***, McGraw Hill (1997)

128. Mitchison, G. and Durbin, R., Learning from your neighbour, ***Nature*** 355:112-113 (1992)

129. Muggleton, S., King, R. and Sternberg, M., Predicting protein secondary structure using inductive logic programming, ***Protein Engineering***, 5:647-657 (1992)

130. Munoz, A. and Muruzábal, J., Self-organizing maps for outlier detection, ***Neurocomputing***, 18:33-60, (1998)

131. Munzner, T., Exploring large graphs in 3D hyperbolic space, ***IEEE Computer Graphics***, 5:18-23 (1998)

132. Nagata, S., Kani, K. and Sugiyama, A., A computer-assisted visual field diagnosis system using a neural network, ***Perimetry Update 1990/91***, 291-295 (1991)

133. Obermayer, K. H., Ritter, H. and Schulten, K., Large-scale simulations of self-organising neural networks on parallel computers: application to biological modeling, ***Parallel Computing***, 14: 381-404 (1990)

134. Packard, N., A genetic learning algorithm for the analysis of complex data, ***Complex Systems***, 4:543-572 (1990)

135. Pawlak, Z., ***Rough Sets – Theoretical Aspects of Reasoning about Data***, Kluwer (1991)

136. Pearl, J., ***Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference***, Morgan Kaufmann (1988)

137. Pearl, J., Reverend Bayes on inference engines: a distributed hierarchical approach, ***Proceedings of the AAAI-82***, 133-136 (1982)

138. Pfeiffer, K., Bechter, E., Embacher, G. and Erhart, K., Identification of empirically "suspicious" data sets in data quality assurance for a new hospital financing system in Austria, in ***Proceedings of Medical Informatics Europe '96*** (eds. J. Brender, *et al.*), 401-404, IOS Press (1996)

139. Piatetsky-Shapiro, G. and Frawley, W. J., ***Knowledge Discovery in Databases***, AAAI Press / The MIT Press (1991)

140. Quinlan, J. R., ***C4.5: Programs for Machine Learning***, Morgan Kaufmann (1993)

141. Quinlan, J. R., Induction of decision trees, ***Machine Learning***, 1:81-106 (1986)

142. Raedt, L. De, ***Inductive Theory Revision: An Inductive Logic Programming Approach***, Academic Press (1992)

143. Ramoni, M. and Sebastiani, P., The use of exogenous knowledge to learn Bayesian networks from incomplete databases, in ***Advances in Intelligent Data Analysis: Reasoning about Data*** (eds. X. Liu, P. Cohen and M. Berthold), 539-548, Springer-Verlag (1997)

144. Rechenberg, I., ***Cybernetic Solution Path of an Experimental Problem***, Ministry of Aviation, Royal Aircraft Establishment, UK (1965)

145. Redman, T., ***Data Quality for the Information Age***, Artech House (1996)

146. Riesbeck, C. and Schank, R., ***Inside Case Based Reasoning***, Lawrence Erlbaum Associates (1989)

147. Ritter, H. J. and Kohonen, T., Self-organising semantic maps, ***Biological Cybernetics***, 61:241-254 (1989)

148. Ritter, H. J. and Schulten, K., On the stationary state of Kohonen's self-organising sensory mapping, ***Biological Cybernetics***, 54:99-106 (1986)

149. Rumelhart, D. E. and McClelland, J. L., ***Parallel Distributed Processing***, 1 & 2, The MIT Press (1986)

150. Rumelhart, D. E., Hinton, G. E. and Williams, R. J., Learning representations by back-propagating errors, ***Nature***, 323:533-536 (1986)

151. Russell, S. and Norvig, P., ***Artificial Intelligence: A Modern Approach***, Prentice-Hall (1995)

152. Schulze-Kremer, S., Genetic Algorithms for Protein Tertiary Structure Prediction, in ***Parallel Problem Solving from Nature, 2*** (eds. R. Männer and B. Manderick), North-Holland (1992)

153. Schweizer, L. G., Parladori, G. L., Sicuranza, G. L. and Marsi, S., A fully neural approach to image compression, in ***Artificial Neural Networks***, edited by T Kohonen, K Makisara, O Simula and J Kandas, 815-820 (1991)

154. Shortliffe, E. and Buchanan, B., A model of inexact reasoning in medicine, ***Mathematical Bioscience***, 23:351-379 (1975)

155. Sommer, A., Enger, C. and Witt, K., Screening for glaucomatous visual field loss with automated threshold perimetry, ***American Journal of Ophthalmology***, 103(5):681-684 (1987)

156. Spiegelhalter, D., Dawid, P., Lauritzen, S. and Cowell, R., Bayesian analysis in expert systems, ***Statistical Science***, 8:219-282 (1993)

157. Strong, N. P., How optometrists screen for glaucoma: a survey***, Ophthalmic Physiological Optics***, 12:3-7 (1992)

158. Swanson, W. H., Birch, E. E., Effects of extraneous noise on threshold estimates, ***Investigative Ophthalmology & Visual Science***, 31:10, (1990)

159. Swanson, W. H., Birch, E. E., Extracting threshold from noisy psychophysical data, ***Digest of Topical Meeting on Noninvasive Assessment of the Visual System***, Optical Society of America, 3:6-9 (1990)

160. Swets, J. A., Pickett, R. M., Whitehead, S. F., Getty, D. J., Schnur, J. A., Swets, J. B. and Freeman, B. A., Assessment of diagnostic technologies, ***Science***, 205:753-759 (1979)

161. Tayi, G. and Ballou, T., Examining data quality, ***Communications of the ACM***, 41:2 (1998)

162. Thylefors, B., The World Health Organization's programme for the prevention of blindness, ***International Ophthalmology***, 14:211-219 (1990)

163. Tuck, M. W. and Crick, R. P., Efficiency of referral for suspected glaucoma, ***British Medical Journal***, 302:998-1000 (1991)

164. Ultsch A., Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series, in ***Kohonen Maps*** (eds. E. Oja and S. Kaski), 33-46, Elsevier (1999)

165. Ultsch, A., Halmans, G. and Mantyk, R., CONKAT: A connectionist knowledge acquisition tool, ***Proceedings of the 24th Annual Hawaii International Conference on System Sciences***, 1:507-513 (1991)

166. Ultsch, A., Knowledge extraction from self-organizing neural networks, in ***Information and Classification*** (eds. O. Opitz, B. Lausen and R. Klar), 301-306, Springer-Verlag (1993)

167. Vapnik, V., ***The Nature of Statistical Learning Theory***, Springer (1995)

168. Wang, R., Storey, V. and Firth C., A framework for analysis of data quality research, ***IEEE Transactions on Knowledge and Data Engineering***, 7:623-640 (1995)

169. Webb, A., ***Statistical Pattern Recognition***, Arnold Publishers (1999)

170. Weiss, S., ***A System for Model-based Computer-Aided Diagnosis and Therapy***, Ph.D. Dissertation, Rutgers University (1974)

171. Weiss, S., Kulikowski, C. A. and Safir, A., A model-based consultation system for the long-term management of glaucoma, ***Proceedings of IJCAI-77***, 826-831 (1977)

172. Weiss, S.M. and Kulikowski, C.A., ***Computer Systems that Learn***, Morgan Kaufmann (1995)

173. Werbos, P., ***Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences***, Ph. D. Thesis, Harvard University (1974)

174. WHO, ***10 Years of Onchocerciasis Control***, OCP/GVA/85 1B (1985)

175. Widrow, B. and Hoff, M., Adaptive switching circuits, ***IRE WESCON Convention Record***, 4:96-104 (1960)

176. Wilks, S. S., Multivariate statistical outliers, ***Sankhya: The Indian Journal of Statistics***, 25:407-426 (1963)

177. Wright, L.A., Wilcox, R., Wu, X., Fitzke, F. W., Wormald, R. P. L. and Johnson, G. J., Motion sensitivity testing in occupational health screening, ***Perimetry Update 1994/95***, 335-338 (1995)

178. Wright, T., ***Statistical Methods and the Improvement of Data Quality***, Academic Press (1983)

179. Wu, J. X., Cheng, G., Liu, X. and Hitchings, R., Assessment of the reliability in a psychophysical test by neural networks, ***Proceedings of the 1994 US Conference on Vision Science***, Optical Society of America, 2:159-162 (1994)

180. Wu, J. X., Fitzke, F.W., Poinoosawmy, P., Hitchings, R. and Johnson, G., Variability in glaucomatous visual damage measured with motion detection, ***Investigative Ophthalmology & Visual Science***, 34(4):1475 (1993)

181. Wu, J. X., Jones, B. R., Cassels-Brown, A., Murdoch, I., Adeniyi, B., Alexander, N., Minassian, D. and Abiose, A., Preliminary report on the use of laptop computer perimetry with a motion sensitivity screening test to detect optic nerve disease in onchocercal communities of rural Nigeria, ***Perimetry Update 1992/93***, 323-329 (1992)

182. Wu, J. X., ***Visual Screening for Blinding Diseases in the Community Using Computer Controlled Video Perimetry***, Ph.D. Dissertation, University of London (1993)

183. Wu, J. X., Wormald, R., Fitzke, F., Poinoosawmy, D., Nagasubramanian, S. and Hitchings, R., Laptop computer perimetry for glaucoma screening, ***Investigative Ophthalmology & Visual Science***, 32(4):810 (1991)

184. Wu, J., Cheng, G. and Liu, X., Reasoning about outliers by modelling noisy data, ***Advances in Intelligent Data Analysis: Reasoning about Data*** (eds. X. Liu, P. Cohen and M. Berthold), 549-558, Springer-Verlag (1997)

185. Wu, J., Cheng, G., Hitchings, R., Liu, X. and Fitzke, F., Identifying the measurement status for psychophysical testing in glaucoma, ***Proceedings of the 14th Annual British Glaucoma Conference***, London, S2:3 (1993)

186. Wu, J., Hitchings, R., Jones, B., Cheng, G. and Liu, X., A novel method to use redundant data from visual field tests by neural networks, Proceedings of the 1st World Congress on Computational Medicine and Public Health, Austin, Texas (1994)

187. Wu, J., Jones, B. R., Cassels-Brown, A., Adeniyi, B., and Abiose, A., Topographical correlation between motion sensitivity test locations and onchocercal chorioretinal lesions, ***Acta XXVIII Internaltional Congress on Ophthalmology***, Toronto (1994)

188. Wu, J., Wormald, R., Hitchings, R., Fitzke, F., Corcoran, S. and Fraser, S. G., ***Acceptability of Glaucoma Case Finding by Motion Sensitivity Perimetry in a General Practice***, MRC Report G9405999, Moorfields Eye Hospital (1997)

189. Wu, J., Wormald, R., Hitchings, R., Liu, X. and Cheng, G., Software-oriented visual field test for glaucoma screening, ***Investigative Ophthalmology & Visual Science***, 39(4):58 (1998)

190. Wyatt, J. and Spiegelhalter, D., Field trials of medical decision-aids: potential problems and solutions, ***Proceedings of the 15th Symposium on Computer Application in Medical Care***, Washington, 3-7 (1991)

191. Wyatt, J. C., Clinical data systems, part 1: data and medical records, ***Lancet***, 344:1682-1688 (1994)

192. Zadeh, L. A., Fuzzy Sets, ***Information and Control***, 8:338-353 (1965)

GLOSSARY

Blind spot: Also called **optic disk**, it is the place on the retina of the eye from which the optic nerve emerges. This spot is insensitive to light. In binocular vision it is not noticed, because the part of the visual field covered by the blind spot of one eye is covered by a sensitive area in the other. (See also: **visual field**.)

Cataract: A cataract is a clouding of the eye's lens that causes loss of vision.

False negative response: In the context of perimetry, the subject fails to indicate seeing a stimulus, which is supposed to be seen. (See also: **false positive response**, **perimetry**.)

False positive response: In the context of perimetry, the subject indicates seeing a stimulus, which does not exist or is not supposed to be seen. (See also: **false negative response**, **perimetry**.)

Fixation loss: The subject's eyeball moves during the perimetry test. (See also: **fixation point**, **perimetry**.)

Fixation point: A spot in the perimeter to which the subject stares at in order to maintain a fixed direction of the eyeball during the perimetry test. (See also: **perimeter**, **perimetry**.)

Glaucoma: Glaucoma refers to a group of eye symptomatic conditions. These can include raised intra-ocular pressure, visual field loss, enlargement of the blind spot and changes in the appearance of the optic nerve head. (See also: **blind spot**, **optic nerve**, **visual field**.)

Hemifield: Half of the visual field. (See also: **visual field**.)

Motion sensitivity: The capacity of an organ or organism to respond to stimulation in the form of a moving object.

Onchocerciasis: Also known as “river blindness,” is a major public health problem in tropical Africa and Central America. It gives rise to serious visual impairment, including blindness; to intensely itching rashes, wrinkling and depigmentation of the skin; to lymphadenitis, resulting in hanging groins and elephantiasis of the genitals; and to general debilitation. Victims often become blind by the time they have reached their mid-30’s.

Optic disk: See **blind spot**.

Optic nerve: The nerve that carries visual impulses from the retina to the brain. In the eye, the nerve forms from the convergence of visual nerve fibres in the optic disk at the rear of the eyeball. (See also: **optic disk**, **retina**.)

Optic neuritis: An inflammation of the optic nerve. The nerve tissue becomes swollen and red, and the nerve fibres do not work properly. (See also: **optic nerve**.)

Perimeter: A device used in perimetry for quantitating the function of the visual field. (See also: **perimetry**, **visual field**.)

Perimetrist: A specialist who operates, controls, or monitors the perimeter to examine the function of the visual field. (See also: **perimeter**, **visual field**.)

Perimetry: The systematic measurement of visual field function. Colour vision testing, flicker sensitivity, contrast sensitivity, pupillary responses and motion testing are some of the other methods of quantitative vision evaluation. (See also: **visual field**.)

Psychophysical test: A procedure that makes experimental measurements of perception using the relationships between physical stimuli and resulting sensations and mental states of human beings.

Receiver Operator Characteristic curve: It shows the relationship of “probability of false alarm” (x-axis) to “probability of detection” (y-axis) for a certain test. Alternatively in medical terms: the “probability of a positive test, given no disease” to the “probability of a positive test, given disease.” The ROC curve may be used to determine an “optimal” cutoff point for the test.

Retina: The retina is the part of the eye that receives the light and converts it into chemical energy. The chemical energy activates nerves that conduct the messages out of the eye into the higher regions of the brain.

Detection rate: The ability to correctly identify those who *have* the disease. (See also **false alarm rate**.)

$$detection\_rate = \frac{true\_abnormals}{true\_abnormals + false\_normals} \times 100\%$$

False alarm rate: The ability to correctly identify those who *do not* have the disease. (See also **detection rate**.)

$$false\_alarm\_rate = \frac{false\_abnormals}{true\_normals + false\_abnormals} \times 100\%$$

Ideally, a test should have 100% detection rate and 0% false alarm rate, i.e. the test always correctly identifies the disease status of the people tested.

Result of Screening	Disease Status	
	Disease	No Disease
Abnormal	True abnormal	False abnormal
Normal	False normal	True normal

The people who have the disease are the “true abnormals” (also called “true positives”). The people who do not have the disease are the “true normals” (also called “true negatives”).

Visual acuity: The capacity to discriminate the fine details of objects.

Visual field: The portion of space in which objects are visible in the eye at the same moment during steady fixation of gaze in one direction.

Principle Publications Arising from the Research of Gongxian Cheng

Evaluating an eye screening test

Cheng, G., Cho, K., Liu, X., Loizou, G. and Wu, J. X.

Advances in Intelligent Data Analysis (eds. D. Hand, J. Kok and M. Berthold), 461-471, Springer-Verlag (1999)

Interactive knowledge discovery through self-organising feature maps

Cheng, G., Liu, X. and Wu, J.

Proceedings of the World Congress on Neural Networks, 4:430-434, San Diego, California (1994)

Establishing a reliable visual function test and applying it to screening optic nerve disease in onchocercal communities

Cheng, G., Liu, X., Wu, J. and Jones, B.

International Journal of Bio-Medical Computing, 41:47-53 (1996)

Discovering knowledge from visual field data

Cheng, G., Liu, X., Wu, J., Jones, B. and Hitchings, R.

Medical Informatics Europe '96, 629-633, IOS Press (1996)

AI for public health: self-screening for eye diseases

Liu X., Cheng, G. and Wu, J.

IEEE Intelligent Systems, 13(5):28-35 (1998)

Identifying the measurement noise in glaucomatous testing: an artificial neural network approach

Liu, X., Cheng, G. and Wu, J. X.

Artificial Intelligence in Medicine, 6:401-416 (1994)

Noise and uncertainty management in intelligent data modelling

Liu, X., Cheng, G. and Wu, J.

Proceedings of the AAAI-94, 263-268 (1994)

Analyzing outliers cautiously

Liu, X., Cheng, G., Wu, J.

IEEE Transactions on Knowledge and Data Engineering, to appear (2001)