Statistical Analysis of High Content Screening Data

Richard Matthew Jacques

Thesis submitted to the University of Sheffield
for the degree of Doctor of Philosophy

April 2009

Department of Probability and Statistics
School of Mathematics and Statistics
University of Sheffield
Sheffield, U.K.

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Nick Fieller for his advice and guidance throughout the preparation of this thesis.

Secondly, I would also like to thank Dr. Edward Ainscow and Chris Harbron from AstraZeneca for making this project possible, providing the data and for their help throughout.

I would like to thank all my friends in the department who have given me advice and made my time so enjoyable. There are too many people to mention everyone's name but I would particularly like to thank Sammy, Emma, Lynsey, Lindsay, Lucy and John. In addition, thanks to all those Friday night regulars from past and present for all the good laughs and helping me escape work.

Finally, I would like to thank my family for supporting me throughout the eight years I have been studying at university.

Abstract

High throughput screening experiments are typically used within the pharmaceutical industry for the identification and evaluation of candidate drugs. Using a high throughput screen with automated imaging platform allows a large number of compounds to be tested in a biological assay in order to identify any activity inhibiting or activating a biological process. High throughput fluorescent images contain information that can be used to define fully the effects of a compound on cells. It is for this reason that florescent imaging assays have been termed high content screening (Clemons, 2004).

The studies analysed in this thesis involve the use of an automated robotic system to administer compounds to cellular assays and take high content images. These images are then analysed and quantified using imaging algorithms to produce a set of variables. Each high content screen may extend to a million or more individual assays.

Supervised classification methods have important applications in high content screening experiments where they are used to predict which compounds have the potential to be developed into new drugs. The use of supervised classification for high content screening data is investigated and a new classification method is proposed for batches of compounds where the rule is updated sequentially using information from the classification of previous batches. This methodology accounts for the possibility that the training data are not a representative sample of the test data and that the underlying group distributions may change as new compounds are analysed.

Unsupervised classification methods are used in the analysis of high content screening experiments to evaluate potential new drugs. The study in this thesis considers clustering compounds based on their toxicological effect on the liver. Drug induced liver injury is the most common cause for non-approval and

withdrawal by the Food and Drug Administration (Ainscow, 2007a) and therefore this is an important stage in drug development.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

High throughput screening experiments are typically used within the pharmaceutical industry for the identification and evaluation of candidate drugs. Using a high throughput screen with automated imaging platform allows a large number of chemical compounds to be tested in a biological assay in order to identify any activity inhibiting or activating a biological process. High throughput fluorescent imaging platforms have several advantages over conventional screening techniques that rely on *in vitro* techniques. The most important of these advantages is that the images contain a wealth of information that can be used to define fully the effects of a chemical compound on cells. It is for this reason that florescent imaging assays have been termed high content screening (Clemons, 2004).

The studies analysed in this thesis involve the use of an automated robotic system to administer compounds to cellular assays and take high content images. These images are then analysed and quantified using advanced imaging algorithms to produce a set of variables. In order to sample as diverse a chemical space as possible, each high content screen may extend to a million or more individual assays (Kenny *et al.*, 1998). This, and the fact that only a small number of compounds in a screen (<1%) are expected to have a desired biological effect means that a number of statistical challenges arise when analysing data from such experiments. In particular, the true classifications of all compounds in a screen are never known because this would involve an expert classifying each of the

million or more compound images by eye. Therefore conventional methods of comparing classification rules cannot be used. In addition, due to the large numbers of compounds the training data for supervised classification is chosen because of the known properties of the compounds and not as a random sample of all compounds. This means that the training data may not be representative of the compounds in future batches.

The motivation for this research comes from AstraZeneca, the industrial sponsor, who provided the data and high content images from a number of screening experiments. Their current approaches for analysing data from this type of screening experiment involve using a single parameter because multi-parametric approaches for compound selection and evaluation are still in their infancy within such industries. However, it is believed that the proper exploitation of the information contained within each high content image will enable more refined compound selection. This results in the statistical problem of developing multi-parametric approaches for classifying and selecting compounds with the desired biological effect on cells from the data generated from high content screening experiments.

Chapter 2 offers a review of the classification literature. The first half of the chapter provides a taxonomy of classifiers and classification methods. In particular, the difference between supervised and unsupervised classification is defined before further subclasses of methods are discussed. The second half of the chapter gives details of some of the statistical methodologies that are used in this thesis such as discriminant analysis, random forests and principal component analysis. This section is designed to be a point of reference for the subsequent chapters. Throughout the chapter reference is made to the suitability of the methodologies to the analysis of high content screening data.

The motivating case studies for the work in this thesis are introduced in Chapter 3. The chapter is divided into two main sections, the first concentrates on the selection of ‘hit’ compounds with the data forming a supervised classification

problem and the second focuses on the clustering of dose response compounds with the data forming an unsupervised classification problem. Each of these sections gives an outline of the problem, a description of the data and defines the objectives before some exploratory analyses are conducted.

Chapter 4 concentrates on the selection of compounds using supervised classification. In particular, it looks at using unlabelled data (i.e. data with unknown group membership) to update classification rules. The first half of the chapter describes the methodologies of model based discriminant analysis and the classification expectation maximization algorithm before they are applied to the high content screening data in the second half. Further analyses adapt the technique of updating using unlabelled data by incorporating the classification reject option and the use of robust estimation of multivariate location and scale.

Chapter 5 introduces a new updating algorithm for classifying high content screening data. This algorithm addresses a number of issues associated with the data; namely, that the training data is not representative of the test data and the underlying group distributions change as new batches of compounds are analysed. This algorithm is applied to the data using a number of different classifiers before comparing the results with the current single parameter approach and classical multivariate classifiers.

The study of the new classification updating algorithm is continued in Chapter 6. The methodology introduced in Chapter 5 is applied to a new high content screening case study with a different biological assay to that considered previously. The data that is analysed comes in two forms. The first uses the same imaging algorithms as were used to produce the variables for the data described in Chapter 3. The second uses some new imaging algorithms which are expected to be more accurate in their measurements of the biological features. The results of classifying the two forms of data are compared and the overall results of using the updating algorithm assessed.

Chapter 7 focuses on the problem of clustering dose response compounds. The chapter begins by reviewing some of the existing methodologies that can be found in the literature. The approach by Perlman *et al*. (2004a, 2004b) is applied to the high content screening data and possible alterations are discussed.

Chapter 8 discusses some areas in which analysis of previous chapters can be extended. It also outlines some areas where further work may be appropriate. The final chapter, Chapter 9, gives conclusions and comments from the analysis carried out in this thesis.

Chapter 2

Review of Multivariate Classifiers

2.1 Introduction

Classification has been used for many hundreds of years, certainly since the Greek philosopher Aristotle used a system to classify animals and plants in the fourth century BC. These early examples of classification were not numerical but were based on the characteristics of the objects of interest (Aristotle's work grouped species of animals into those with red blood and those without). Numerical techniques for classification originated in the natural sciences as an attempt to rid taxonomy of its subjective nature. In particular, the use of single characteristics to classify objects was replaced by the use of multiple characteristics. Techniques were used that produced consistent classifications regardless of whether more objects were added to the study or the analysis repeated. Classification has played an important role in many different scientific fields and will continue to do so into the future (Everitt *et al*., 2001). This chapter outlines the structure of different methods of classification and relates the methodologies to the research in this thesis, that of analysing data from high content screening experiments.

First an introduction to classifiers is provided and the differences between supervised and unsupervised classification are outlined. In particular, Section 2.2, takes the form of a taxonomy with a comparison of the structures of different types of classifiers and the data that they can be used to analyse. Section 2.3 concentrates on the relevant methodology and literature for the later chapters; this

includes descriptions of discriminant analysis, random forests, the k-nearest neighbour classifier, principal component analysis, principal coordinate analysis and the k-means clustering algorithm. Throughout the chapter comments are made on the application of classifiers to high content screening data but specific reviews of existing methodologies are given in later chapters. Section 2.4 discusses methods of assessing classification rules and Section 2.5 outlines the statistical packages used for the analyses conducted in this thesis. Section 2.6 is the final section of this chapter and it contains a summary and some discussion.

2.2 Classifier Taxonomy

This section gives a general introduction to the different methods of classification. The aim is to provide an overview of the structure of different groups of classifiers before Section 2.3 goes into specific details of the classifiers used in the remainder of the thesis. Figure 2.1 is a graphical representation of the taxonomy of classification and will be referred to throughout the remainder of this section.

The main division between classifiers is that of supervised (also referred to as discrimination or supervised pattern recognition) and unsupervised (also referred to as cluster analysis or unsupervised pattern recognition). These two groups of classifiers can then be subdivided further. The supervised classifiers can be split into methods which approximate classification boundaries and methods which approximate class conditional boundaries (see Section 2.2.1 for details). Unsupervised classifiers can be subdivided into three main groups, those which are hierarchical, data analytic and optimized (see Section 2.2.2 for details).

2.2.1 Supervised Classification

In supervised classification the class structure is known a priori. The aim is to use a sample of objects (described in terms of vectors of features) with known class to construct a rule which allows new objects to be assigned to one of the pre-specified classes based only on their measurement vectors. The sample of objects

with known class are called the design set, training set, learning set or labelled data and the new objects with unknown classes are called the test set or unlabelled data. It is assumed that the objects in the training set are randomly sampled from the same distribution as the objects that are in the test set, however, this is not always the case (see Section 5.1 for further discussion). The features that are used to describe the objects together span a multivariate space known as the measurement space or feature space (Hand, 1997 and 2006).

Supervised classification in this thesis will focus on classifying compounds into those that activate a biological process and those that do not. The training data set is made up of compounds with known biological effect (i.e. both compounds that are known to activate and not to activate the biological process) and the test data contains compounds with unknown properties. A full description of the high content screening data set used for supervised classification is given in Section 3.2. Chapters 4, 5 and 6 will concentrate on analysis using supervised classification.

The main distinction between supervised methods of classification is based on whether they approximate classification boundaries (or discriminant functions) or they approximate class conditional densities. The aim of those methods which approximate classification boundaries is to segment the measurement space into regions belonging to different classes. The segmentation of the measurement space can be done in two different ways. Techniques such as the combination of classifiers and tree classifiers are termed structural, techniques such as Fisher's linear discriminant analysis and generalized linear discriminators are termed functional. The functional classifiers estimate discriminant functions using combinations of variables (either linear or non-linear). The structural classifiers estimate classification boundaries by a process of segmenting the measurement space over a number of steps using single variables or combinations of small numbers of variables. A description of the different methods of combining classifiers is given in Section 2.2.3 and the method of Fisher's linear discriminant analysis is given in Section 2.3.1.

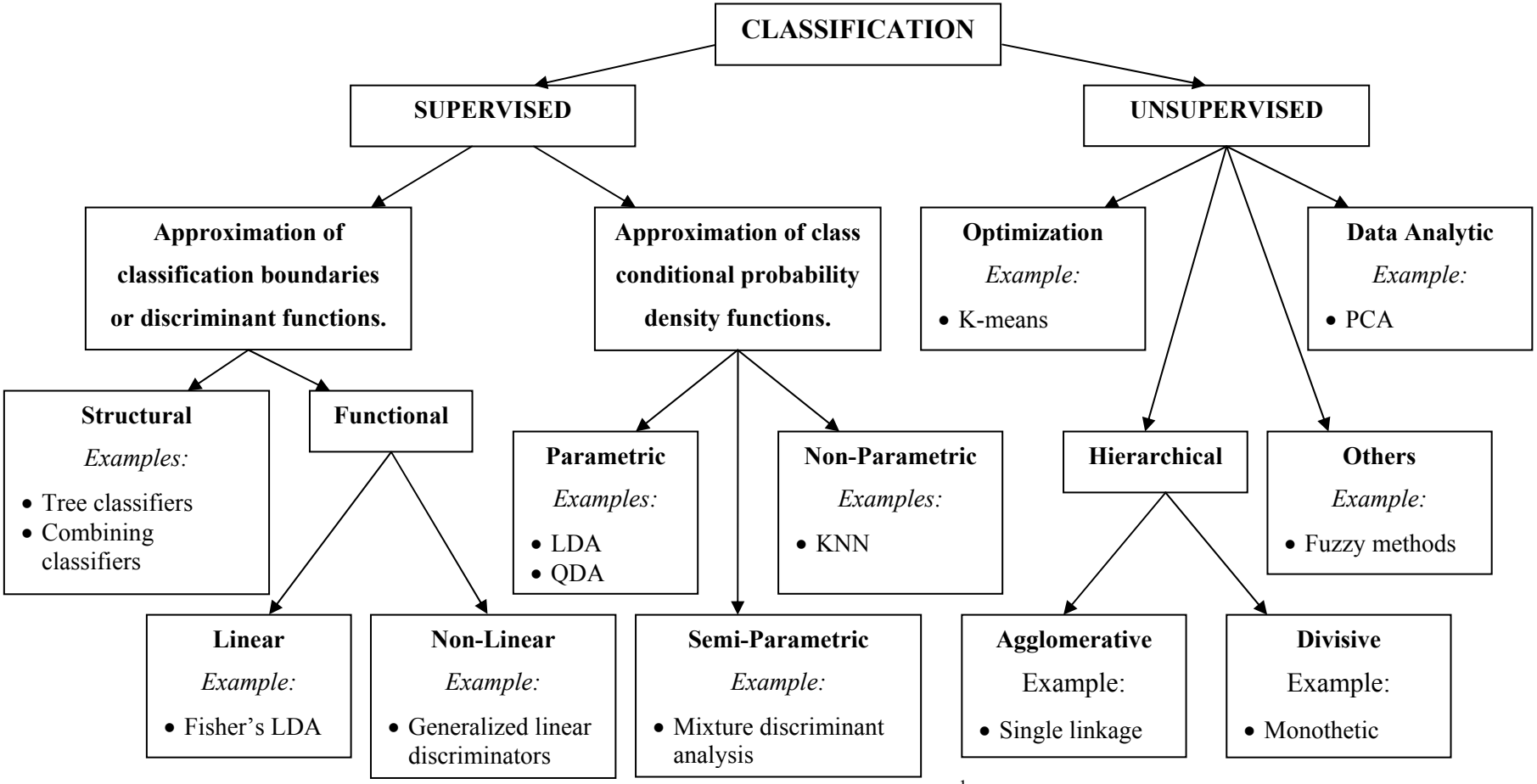


Figure 2.1: Taxonomy of Classifiers[1]

[1] Adapted from Kuncheva (2004)

The approximation of class conditional densities can be done either parametrically, semi-parametrically or non-parametrically. In other words, the differences between the techniques in this group are based on the method used for approximating the class densities. The most common methods used are K-nearest neighbours (see Section 2.3.4) and maximum likelihood linear and quadratic discriminant analysis (see Section 2.3.1). Although, examples of the different classifiers have been given in Figure 2.1, the distinction between the groups is not always clear-cut. For example, the k-nearest neighbour is grouped under the non-parametric estimation of densities but also produces an estimate of the classification boundaries so it could be grouped under structural methods for estimating classification boundaries (Kuncheva, 2004).

In terms of data from high content screening experiments there is no particular reason why methods that approximate classification boundaries or discriminant functions should be used over methods that approximate class conditional densities. However, from a general statistical view point those methods that require assumptions to be made about distributions (for example, multivariate normality) are less flexible than those that do not. A review of the current methodologies being used for supervised problems in high content screening experiments is given in Section 3.2.3.

2.2.2 Unsupervised Classification

The difference between unsupervised and supervised classification is that the class structure is not known a priori in the former. Given a collection of objects, the aim is to determine a natural class structure; that is, to decide on the number of classes and to assign each object to one of these classes. Unsupervised classification can be separated into two types, depending on the objective of the analysis. The first type seeks to identify naturally occurring structure. The research in this thesis focuses on this type as chemical compounds are classified based on their effect on cells. The second type of unsupervised classification

simply seeks convenient division of objects, examples of this can be found in marketing and finance (Hand, 1997).

Unsupervised classification in this thesis will focus on clustering compounds with respect to their toxicological effect on cells. A detailed description of the high content data set that will be analysed using unsupervised classification is given in Section 3.3. Chapter 7 concentrates on the analysis of this data.

Hierarchical clustering differs from other methods of unsupervised classification as data is not partitioned into groups in a single step. Instead the division occurs over a number of steps where each one can be viewed as the most efficient partition for the progressive subdivision of the population. Hierarchical clustering techniques may be subdivided into agglomerative methods and divisive methods. Agglomerative techniques start with each individual being its own cluster. At each step of the analysis the clusters are then combined until all individuals belong to a single cluster. Two examples of agglomerative hierarchical clustering techniques are single linkage clustering and complete linkage clustering. The techniques differ by the way in which distance (or similarity) is defined between individuals or groups of individuals. Divisive methods work in the opposite way to agglomerative methods, starting with one large cluster of all individuals and successively splitting into smaller clusters. These methods can be further divided in to monothetic techniques which are based on a single specified attribute and polythetic techniques which are based on values taken from all attributes. For both types of hierarchical clustering the results can be displayed on a dendrogram. Further details on hierarchical clustering techniques and methods of measuring distance and similarity can be found in Everitt *et al*. (2001).

Optimization clustering works by their minimizing or maximizing a numerical criterion to produce a partition of the data into a specified number of groups. The clustering methods in this class differ from each other in the criteria that is optimized and the optimization algorithm used. These methods also differ from those of hierarchical clustering in that the number of clusters has to be decided

beforehand. Details of different clustering criteria and methods for optimizing these criteria can be found in Chapter 5 of Everitt *et al*. (2001). A description of the K-means algorithm is given in Section 2.3.7.

Hierarchical and optimization clustering both use algorithms to determine which objects should be grouped into the same clusters. The groups of techniques that are called data analytic differ in that they do not directly determine clusters for individual objects but instead allow the data points to be plotted in low dimensions so the user can visualise clusters of objects. This has a number of benefits which include giving an insight into the structure of the data before a more formal clustering algorithm is used and aiding the selection of the number of groups in the data for use in other analyses. The two data analytic methods used in this thesis are principal component analysis and principal coordinate analysis (see Sections 2.3.5 and 2.3.6 respectively for details).

The three categories of methods for supervised classification (hierarchical, optimization and data analytic) that have already been discussed contain the most widely used clustering techniques. However, there are still a considerable number of other methods that do not fall clearly into these categories. There are far too many techniques left to give a comprehensive review of them all but some examples are: fuzzy methods (objects are assigned a membership function indicating the strength of membership to each cluster), methods which allow overlapping clusters and density search analysis (clusters are assumed to be concentrated in dense patches in a metric space). Details of these and many other methods are given in Chapter 7 of Everitt *et al*. (2001).

2.2.3 Combining Classifiers

The previous sections of this classifier taxonomy have all concentrated on methods for discrimination using a single classifier. However, methodologies are available for combining more than one classifier to use in one classification

technique. When combining classifiers the two main strategies available are classifier fusion and classifier selection. These two strategies are outlined below.

Dasarathy and Sheela (1978) were the first to suggest the idea of using different classifiers for different inputs when they combined a linear classifier with k-nearest neighbours. Classifier selection rules base their classification on a choice from a set of constituent rules. The idea behind this method is to partition the measurement space into regions, with each region being associated with the constituent rule that best suits it (Hand, 1997).

In contrast to classifier selection, each classifier in a classifier fusion ensemble has knowledge of the whole feature space. Hence, methods of combining classifiers such as average and weighted vote are used. The two most popular methods of classifier fusion are bagging (Breiman, 1994) and boosting. A detailed description of a Random Forest (a bagging algorithm for tree classifiers) is given in Section 2.3.3.

2.3 Methodology

2.3.1 Discriminant Analysis

Discriminant analysis can take many different forms but the overall aim of each of the different methods is the same; to predict the membership of objects in the classes of a categorical dependent variable from their measurements on one or more predictor variables. Fisher (1936, 1938) first suggested an approach to discrimination for two groups that does not make any assumptions about the parametric form of the distributions of populations. Fisher's suggestion was to look for a linear function that maximised the ratio of the between-groups sum of squares to the within-groups sum of squares.

The methodology of Fisher can be extended from the situation with two populations to situations with three or more populations. Given k populations or groups Π_i, $i = 1,2,...,k$; each with n_i observations and a data matrix $\mathbf{X}_i$ (so data are $\{\mathbf{x}_{ij}; i = 1,...,k; j = 1,...,n_i\}$) look for the linear function $\mathbf{a}^T\mathbf{x}$ that maximises the ratio of the between-groups sum of squares to the within-groups sum of squares. The linear function $\mathbf{a}^T\mathbf{x}$ is called Fisher's linear discriminant function or the first crimcoords. The vector $\mathbf{a}$ in this function is defined to be the eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ corresponding to the largest eigenvalue, where $\mathbf{B}$ is the between-groups sum of squares given by

$$\mathbf{B}=\frac{1}{k-1}\sum_{i=1}^{k} n_i \left(\bar{\mathbf{x}}_i-\bar{\mathbf{x}}\right)\left(\bar{\mathbf{x}}_i-\bar{\mathbf{x}}\right)^T$$ ($\bar{\mathbf{x}}$ is the overall mean and $\bar{\mathbf{x}}_i$ the mean of group Π_i),

and $\mathbf{W}$ is the within-groups sum of squares given by

$$\mathbf{W}=\frac{1}{n-k}\sum_{i=1}^{k}\left(n_i-1\right)\mathbf{S}_i ,$$

with

$$\mathbf{S}_i=\frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i\right)\left(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i\right)^T .$$

Once the linear discriminant functions have been calculated, an observation x can be classified to one of the populations on the basis of $\mathbf{a}^T\mathbf{x}$; its linear discriminant score. The discriminant rule in the two population case is given by:

"allocate x to population Π_1 if $\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)^T\mathbf{W}^{-1}\left\{\mathbf{x}-\frac{1}{2}\left(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2\right)\right\}>0$ and to population Π_2 otherwise ".

When applying to a situation with three groups, a second linear function is defined as the eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ corresponding to the second largest eigenvalue. Hence, an allocation rule based on two discriminant functions can be calculated.

When the distributions of the group populations are known a maximum likelihood alternative to Fisher's linear discriminant analysis may be used. The maximum likelihood approach described below gives optimal results if each group is normally distributed with a common with-in group covariance matrix. However, if the groups are non-normal but the distribution is known then an alternative optimal rule can be constructed. For a problem with two or more groups the rule for classification is straight forward. Let π_j be the prior probability for class j and then a new point $\mathbf{x}$ is assigned to the class j which has the largest value of

$$f(j|\mathbf{x}) \propto \frac{\pi_j}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right].$$

Fisher's linear discriminant analysis has the advantage over other linear discriminant techniques in that multivariate normal distributions (or any other distribution) are not assumed. This makes the method more flexible than a parametric approach and hence can be applied to a wider range of discriminant problems. Fisher's criterion is also intuitively attractive because it is easier to tell the groups apart if the between-groups sum of squares is large relative to the with-in group sum of squares (Mardia, et al., 1979).

The discriminant analysis decision surfaces that have been discussed thus far have all been linear. An alternative to these linear discriminators can be produced by including transformations of the measurement variables. In other words, by including extra, 'derived' variables, which are functions of the original variables it is possible to produce more flexible classification rules. One possible quadratic classification rule can be derived by relaxing the restriction that the covariance matrices of the groups be assumed equal. A new point x can then be assigned to the class j that has the largest value of

$$f(j|\mathbf{x}) \propto \frac{\pi_j}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right].$$

Although quadratic discriminant analysis provides more flexible classification rules, Hand (1997) suggests that it should be implemented cautiously. This caution is required because the quadratic functions are more complicated than the linear ones; hence, this method can easily overfit the data and produce unreliable results. There is some evidence to support this in Section 3.2.6 where classification using linear discriminant analysis produced a lower overall misclassification rate than when using quadratic discriminant analysis.

In addition to the methods of discriminant analysis that have already been described, there are further ways in which these methods can be generalized. The two techniques that are considered in this thesis are mixture discriminant analysis and model based discriminant analysis. Mixture discriminant analysis allows each observed class to be a mixture of unobserved normally distributed subclasses. A description of the methodology is given in Section 2.3.2. Model based discriminant analysis is an extension of the maximum likelihood approach to discriminant analysis and allows restrictions to be placed on the volume, shape and orientation of the groups. Section 4.2.1 contains a description of the model based discriminant technique and Section 4.3 describes the application of this methodology in the context of using unlabelled data to update classification rules.

2.3.2 Mixture Discriminant Analysis

Linear discriminant analysis can be derived using maximum likelihood for normal populations with different means and common covariance matrix. Mixture discriminant analysis generalizes this method by assuming that each observed class is a mixture of unobserved normally distributed subclasses (Hastie and Tibshirani, 1996). The mixture discriminant analysis model used in this thesis assumes that each subclass has a multivariate normal distribution with mean

vector $\boldsymbol{\mu}_{jr}$ and common covariance matrix $\boldsymbol{\Sigma}$ (other models are possible, for example, each subclass may be allowed to have a different covariance matrix).

Given k classes $G_j, j = 1,...,k$ each with $r = 1,..,R_j$ subclasses, let Π_j be the prior probability for class j, and within class j let π_{jr} be the mixing probability for the r^{th} subclass,

$$\sum_{r=1}^{R_j} \pi_{jr} = 1.$$

The mixture density for class j is

$$m_j(x) = P\left(X=x \middle| G=j\right) = \frac{1}{(2\pi)^{½} \left|\boldsymbol{\Sigma}\right|^{½}} \sum_{r=1}^{R_j} \pi_{jr} \exp\left\{-\frac{1}{2} D\left(\mathbf{x}, \boldsymbol{\mu}_{jr}\right)\right\}$$

where

$$D\left(\mathbf{x}, \boldsymbol{\mu}\right) = \left(\mathbf{x}-\boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}-\boldsymbol{\mu}\right),$$

the Mahalanobis distance between **x** and **μ**. The expectation maximization (EM) algorithm is used to maximize the conditional log-likelihood for the data (for details of the EM steps see Hastie and Tibshirani (1996)). The posterior class probabilities are given by

$$P\left(G=j \middle| X=\mathbf{x}\right) \sim \Pi_j P\left(\mathbf{x} \middle| j\right) \sim \Pi_j \sum_{r=1}^{R_j} \pi_{jr} \exp\left\{-\frac{1}{2} D\left(\mathbf{x}, \boldsymbol{\mu}_{jr}\right)\right\}$$

normalized so that

$$\sum_{j=1}^{J} P\left(G=j \middle| X=\mathbf{x}\right) = 1.$$

An observation is then classified by choosing j to maximize $P(j|\mathbf{x})$.

2.3.3 Classification Trees and Random Forests

The automatic construction of classification trees originates from work in the social sciences (Venables & Ripley, 2002). However, Breiman *et al*. (1984) were the first to bring the work to the attention of statisticians and to propose new algorithms for constructing trees.

Tree-based models are fitted by binary recursive partitioning. A collection of rules, each based on a single variable, are used to split the training data into increasingly similar subsets. The splitting of the data stops when the subsets are either homogeneous or consist of not enough observations. Every time the classification tree is split a new node is formed. The next split is chosen by taking the maximum reduction in deviance over all allowed splits of all nodes. The deviance in this case is a measure of node heterogeneity. At each of the splits the variables have to be considered in order to decide which should be used. The tree is a probability model and therefore at each of the nodes there is a probability distribution over the classes. The endpoints or terminal nodes of a tree have the observations assigned to them, where the class with the highest probability is chosen (Chambers and Hastie, 1993, Ripley, 1996, and Venables and Ripley, 2002).

Breiman (2001) proposed a method of combining many classification trees; this classifier is called a random forest. This method combines the idea of bagging (Breiman, 1994) with random features. (Bagging is a method of combining classifiers using a majority vote. N bootstrap samples are taken from a dataset and a classifier is constructed for each sample. The final classification for each observation is the one most often predicted by the N classifiers). Each classification tree within the forest is constructed in the following way. Let the

number of observations in the training data be N and the number of variables be M. For each tree a training set of size N is selected by taking a bootstrap sample of the data. When growing each tree a number $m \leq M$ is specified such that at each node of the tree, m variables are selected at random out of the M and the best split on these m variables is used to split the node. In addition, there is no pruning of the individual tree classifiers. The random forest classifies observations by choosing the most frequently occurring of the classes as determined by the individual trees in the forest.

In tests conducted by Breiman (2001) the combining of random features with bagging in the random forest classifier indicated that significantly lower error rates were possible for larger data sets. However, less improvement was found for smaller data sets. Nevertheless, this paper does conclude that different injections of randomness can produce better results than classical methodologies. The results of the analyses conducted in Section 3.2.6 support these findings with the random forest producing a smaller overall misclassification rate than a single classification tree. Further comparisons of classifiers in Section 5.4 also show that the random forest outperforms linear discriminant analysis and the k-nearest neighbour classifier.

2.3.4 K-Nearest Neighbour Classifier

The K-Nearest Neighbour (KNN) classifier is based on a non-parametric estimation of class densities (see Figure 2.1). Given an object with measurement vector **x** the objective is to estimate the conditional probability that the object belongs to class j. An estimate of this probability is given by the proportion of training points in class j amongst the k nearest to **x**. In other words, given the k training points that are closest to **x**, the object is classified using a majority vote amongst these neighbours (ties broken at random). Despite the simplicity, KNN has been successful in a large number of classification problems; in particular, it is

often successful when decision boundaries are very irregular (Ripley, 1996 and Hastie *et al*., 2001).

When implementing the k-nearest neighbour methodology, decisions have to be made about the value of the parameter k (determining the size of the neighbourhood) and the metric by which to measure nearness (determining the shape of the neighbourhood). With regards to the parameter k, a large value means that there is less variance in the probability estimates, but there is also likely to be increased bias. Conversely, a small value of k, means that there will be increased variance and decreased bias. The method for determining the value of k in this thesis is the leave-one-out cross-validatory approach. Details of this are given in Section 5.3.3 along with an example of the application. With regards to the distance metric, the Euclidean distance is used for measuring distance during all applications. Further details on choice of metric and the selection of k can be found in Hand (1997).

The k-nearest neighbour classifier is used to classify high content screening data in Chapter 5. The results of these analyses (see Section 5.4) show that it does not perform as well as the current single parameter approach in terms of selecting the compounds of interest but it does reduce the number of false positives. However, the analyses were only conducted on one data set and therefore the results cannot be generalised to all high content screening data.

2.3.5 Principal Component Analysis

Principal component analysis is included in this review of multivariate classifiers because it is a useful tool for visualising data. Using this method the multivariate data can be plotted on a small number of principal components which allows a visual search for any structure or clustering. With regards to high content screening data, principal component analysis will be used for both data that is supervised and unsupervised in nature. This will take the form of visualizing data

in a low number of dimensions as a method of preliminary analysis of both data sets in Chapter 3 and as a method of looking at changing group distributions in Chapter 5.

The aim of principal component analysis is to find a set of orthogonal coordinates such that the sample variances of the data are in descending order of magnitude with respect to the coordinates. Let $\mathbf{X} = (\mathbf{x_1}, ..., \mathbf{x_n})^T$ be a sample data matrix and let $\mathbf{a}$ be a standardized vector. Then $\mathbf{Xa}$ gives *n* observations on a new variable defined as a weighted sum of the columns of $\mathbf{X}$. The sample variance of this new variable is given by $\mathbf{a^TSa}$, where $\mathbf{S}$ is the sample covariance matrix of $\mathbf{X}$. The vector $\mathbf{a}$ such that the projection of the data $\mathbf{X}$ onto $\mathbf{a}$ has maximal variance is known as the first principal component. This is equivalent to maximising $\mathbf{a^TSa}$ subject to the normalizing constraint $\mathbf{a^Ta} = \mathbf{1}$ and is solved by finding λ_1, the largest eigenvalues of $\mathbf{S}$, and $\mathbf{a}$, the corresponding eigenvector. The maximum value of this variance is then λ_1. This is the solution to the eigenequation $\mathbf{Sa} = \lambda_1\mathbf{a}$.

The remaining principal components are found by maximising the variance of the projection of the data $\mathbf{X}$ onto $\mathbf{a_j}$ subject to the additional constraint of being orthogonal to all preceding components $\mathbf{a_1}, ..., \mathbf{a_{j-1}}$. It can be shown that the p principal components of data $\mathbf{X}$ are the p eigenvectors $\mathbf{a_1}, ..., \mathbf{a_p}$ corresponding to the p ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ of the variance of $\mathbf{X}^T$ (Mardia *et al.*, 1979). Typically the majority of the variation in the data is accounted for by the first few principal components and therefore these are the most interesting. Later principal components explain decreasing amounts of variation and little information is lost if they are discarded. A scree plot of cumulative relative proportion of variance explained is a good graphical method of deciding how many principal components need to be examined. Typically, a scree plot will increase steeply for the first few principal components and then begin to level off. The point where it starts to level off indicates that using more principal

components brings less return in terms of variance explained. Examination of the first few principal components can reveal the major types of variation and by displaying the original data referred to these coordinates divisions between cases can be shown that can be described by the differences in the components. Further details of this method can be found in Mardia *et al.* (1979).

2.3.6 Principal Coordinate Analysis

Principal coordinate analysis or multidimensional scaling is a method for visualizing clusters of data. The aim is to construct a configuration of n points in Euclidean space using information about the distances between the n objects. These distances do not need to be Euclidean, and can be based on both dissimilarities and similarities. This method differs from that of principal component analysis in the previous section as the data points are not directly observable as n points in p-space (Cox and Cox, 2001 and Mardia *et al.*, 1979). Principal coordinate analysis is applied in Chapter 7 as a method of visualizing clusters of dose response data.

Given an (n×n) distance matrix $\mathbf{D} = \left(d_{ij}\right)$, the objective is to represent the inter object distances in a Euclidean space of low dimension k. The matrix D does not need to be Euclidean. The classical solution is to choose a configuration of points whose coordinates are determined by the first eigenvectors of the matrix B defined below. The calculations of this classical solution are as follows. Firstly, construct the matrix $\mathbf{A}=(-\frac{1}{2}d_{ij}^2)$. Then obtain the matrix $\mathbf{B} = \mathbf{HAH}$ where $\mathbf{H}$ is the centring matrix $\mathbf{H} = \mathbf{I_n} - \frac{1}{n}\mathbf{J_n}$ ($\mathbf{I_n}$ is the n×n identity matrix and $\mathbf{J_n}$ is the n×n matrix with all entries equal to 1). From the matrix $\mathbf{B}$ find the k largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ and their corresponding eigenvectors. The principal coordinates are then the rows of the matrix of ordered eigenvectors of $\mathbf{B}$ (Mardia *et al.*, 1979).

When applying the principal coordinate methodology, the number of dimensions that are required for representing the dissimilarities needs to be decided upon. The maximum dimensions of the space required are determined by the eigenvalues of the matrix **B**. If **B** is positive semi-definite then the number of non-zero eigenvalue determines the number of dimensions required. If **B** is not positive semi-definite then the required dimension is the number of positive eigenvalues. However, in practical applications the number of dimensions should be smaller than the maximum space. As with principal component analysis, a scree plot of cumulative relative proportion of variance explained is a good graphical method of deciding how many principal coordinates need to be examined. Further details of this method can be found in Cox and Cox (2001).

In the preceding paragraphs the classical solution has been described. Implicit in this solution is the assumption that there is a configuration in k dimensions with inter point distances δ_{ij}. This configuration is constructed using an observed distance matrix **D** with elements of the form $d_{ij} = \delta_{ij} + e_{ij}$ (e_{ij} represent errors of measurement). However, it is possible to have a non-metric solution where there is a less rigid relationship between d_{ij} and δ_{ij}; namely $d_{ij}=f\left(\delta_{ij}+e_{ij}\right)$, where f is an unknown monotone increasing function. For this approach D is not thought of as a distance matrix but as a dissimilarity matrix and the only information that can be used to reconstruct the δ_{ij} is the rank order of the d_{ij}. For more information on non-metric methods see Cox and Cox (2001) and Mardia *et al.* (1979).

2.3.7 K-Means Clustering

The k-means clustering algorithm is an optimization method of unsupervised classification (see Figure 2.1). It seeks to partition the data into a specified number of groups, k. It is not possible to consider every partition (a problem of clustering 100 objects into 5 groups gives a total of approximately 6.6×10^{67}

different partitions) so the algorithm searches for minimum values of the clustering criterion by rearranging existing partitions, keeping the new one if it makes an improvement. However, this does not guarantee that the global minimum is found. Algorithms of this type are called hill-climbing algorithms and the essential steps are as follows:

1. Find an initial partition of the observations into the required number of groups.
2. Calculate the change in the clustering criterion produced by 'moving' each individual from its own to another cluster.
3. Make the change that leads to the greatest improvement in the value of the clustering criterion.
4. Repeat steps (2) and (3) until no move of an individual causes the clustering criterion to improve.

One possible hill-climbing algorithm is to iteratively update the partitions by simultaneously relocating each object to the group whose mean is closest and then relocating the group means. It can be shown that this is equivalent to finding the partition that minimizes the within-group sum of squares. Such algorithms, involving the calculation of the mean (centroid) of each cluster, are referred to as k-means algorithms (Everitt *et al*., 2001 and Everitt, 2005).

In order to apply the k-means clustering algorithm (and for most optimization methods), the number of clusters in the data has to be estimated. There are a variety of methods that can be used for this estimation. In this thesis an informal method of plotting the value of the clustering algorithm against the number of groups will be applied. In other words, plotting the within-group sum of squares associated with a range of values of k. Large changes of levels in the plot are taken as suggestive of a particular number of groups. Examples of more formal methods for selecting the number of groups can be found in Everitt *et al*. (2001).

2.4 Assessment of Classification Rules

The objective of building a classification rule is to classify correctly as many future objects as possible. It is therefore important to identify methods which enable different classification rules to be evaluated and compared.

The most popular measure of performance for classification rules in the misclassification rate. The misclassification rate is the proportion of objects that are misclassified by a rule. However, it is suggested by Hand (1997) to avoid testing a rule on the data used for its construction because this would lead to a misclassification rate that is optimistically biased as an estimate of future performance. The estimate of error rate obtained by reusing the data used to train the classification rule is termed the resubstitution or apparent error rate.

There are several different methods in which an estimate of the misclassification rate can be obtained without incurring resubstitution bias. These methods are the independent test set approach, cross validation and jackknife methods. For the purpose of the evaluation of classification rules in Section 3.2.6 and Section 4.3 the independent test set approach will be used as the data is most suited to this. This method counts the proportion of objects which the rule misclassifies in a set of test data. This method is suitable for the high content screening data that will be analysed in these sections because the data comes from a pre-screen experiment with a randomly selected set of training data that will be used to build the classifiers and a set of test data that shall be used for evaluation. However, when evaluating classification rules in Section 4.4, Chapter 5 and Chapter 6 this method approach is not viable. In these cases, due to the large number of compounds in each of the test batches it was only possible to check the classification of those compounds that were predicted to be true hits. This was considered to be sufficient because the numbers of true hits and false positives could be compared for each classifier.

In addition to using the estimated classification rate to compare classifiers it is also possible to calculate other measures such as inaccuracy (how effective the rule is in assigning an object to the correct class) and imprecision (how different the estimated class probabilities are from the true probabilities). Details of these measures along with other aspects of evaluation can be found in Chapters 6 and 7 of Hand (1997).

2.5 Computational Notes

All of the statistical analyses in this thesis have been conducted in the R statistical computing package (R Development Core Team, 2008). A number of R packages were used when implementing methodologies in Chapters 3-7, details of these packages are as follows. Classification trees in Chapter 3 were applied using the package tree (Ripley, 2007). The random forest classifier used in Chapters 3, 5 and 6 was implemented using the randomForest package (Liaw and Wienner, 2002). The mclust package (Fraley and Raftery, 2007) for model based clustering was used in Chapter 4 for applying the methodology of updating classification rules using unlabelled data. Finally, Chapter 5 used the packages mda (Hastie and Tibshirani, 2006) and class (Venables and Ripley, 2002) for mixture discriminant analysis and the k-nearest neighbour classifier respectively.

2.6 Summary and Discussion

This chapter has provided an overview of multivariate classifiers and comments have been made about general applications to data from high content screening experiments. This thesis is concerned with problems that are both supervised and unsupervised from a classification point of view. The differences between these two types of classification problem were discussed in the form of a taxonomy of classifiers at the beginning of the chapter and the idea of combining different

classifiers was introduced. The remainder of the chapter gave specific details of the methodologies that are applied in the remaining chapters.

Although this research is motivated by high content screening experiments, specific examples of the current methods for analysis of data from these experiments has not been giving in this chapter. Instead a review of the methodologies for the supervised hit selection problem (as described in Section 3.2) is given in Section 3.2.3 and existing techniques for the unsupervised dose response clustering problem (as described in Section 3.3) is given in Section 7.2.

Chapter 3

Data Description

3.1 Introduction

This chapter introduces the datasets whose analyses have provided the motivation for much of the work in this thesis. Various features of the datasets required extensions and developments of statistical methodology, especially for classification. There are two main sections, the first (Section 3.2) concentrates on compound hit selection with the data forming a supervised classification problem and the second (Section 3.3) focuses on clustering dose response compounds with the data forming an unsupervised classification problem. In each of these sections a description of the data, a brief overview of the biological background and the general statistical objectives are given. Key features of the data sets are highlighted through exploratory analyses and lead onto main analyses and new methodologies in subsequent chapters. The chapter concludes by emphasising the key problems for each case study.

3.2 Compound Hit Selection

The data in this section relate to a high content screening experiment designed to discriminate between chemical compounds that may have the potential to be developed into future drugs and those that do not. Hits are compounds that are

selected as having a beneficial effect on the cells. These hits are currently selected using a single parameter approach (see Section 3.2.3) before a manual image inspection is carried out by a screening expert to classify them as either true hits or false hits (false positives). Those compounds that are not selected as hits are denoted as non-hits and no manual image inspection is carried out on these compounds.

The remainder of this section is organised as follows. Section 3.2.1 gives a brief description of the biological background to the dataset before Section 3.2.2 describes the data and method of collection. A review of the current methodology for hit selection in high content screening experiments is given in Section 3.2.3 before the objectives of analysis are described in Section 3.2.4. Finally, Section 3.2.5 describes some exploratory analysis of the data set.

3.2.1 Biological Background

The data is derived from a high throughput screen to identify antagonists for a G-Protein Coupled Receptor (GPCR). The GPCR class of proteins represent a major class of drug targets. The assay used here is derived from a generic assay for GPCR activation. A cell line was generated that expressed the receptor of interest and fluorescently tagged protein β-arrestin. Upon activation of the receptor, β-arrestin will associate with the receptor at the cell membrane and then drive the internalisation of the receptor into intracellular vesicles. The appearance of the fluorescent label thus appears as a punctate distribution. In the presence of an antagonist of the receptor, the receptor does not associate with β-arrestin. Under these conditions, β-arrestin is uniformly distributed throughout the cell's cytoplasm. The assay uses an automated imaging platform to visualise the fluorescence distribution within the cells in response to the test compounds. Image analysis algorithms are then used to quantify the distribution of fluorescence as to the degree to which the fluorescence appears punctate to identify active compounds within those screened.

The cells are also counterstained with a nuclear die to identify their location. Using additional image analysis algorithms, it is possible to quantify features of the cells not related to antagonism of the receptor. These include changes in nuclear morphology, fluorescent label intensity and cell health. In combination, the features potentially report the ability of a test compound to specifically inhibit the receptor of interest, versus non-specific effects such as toxicity.

3.2.2 Data Description

Compounds are processed through the screen on plates consisting of arrays of 384 wells. One compound is added to each well with approximately 250 cells. Images of the wells are taken before advanced imaging algorithms take measurements and produce a set of sixteen variables consisting of identifiers and quantified variables. The identifiers give information on the run or cycle of the experiment, plate number and well grid reference for each compound. This allows the quantified variables to be matched to their compounds and corresponding images. Descriptions of the quantified variables are given in Table 3.1. The first variable listed, Npos, is a count of the cells in the image, while the remaining fifteen variables are means. Note that the imaging algorithms produce raw variables based on the individual cells in each image but only the means and standard deviations of these variables are available.

All wells that are selected as being hits have their images manually inspected by eye. During this process, in addition to the hits being classified as true hits or false hits, each are categorized further. The true hits are classified into two groups depending on the level of inhibition. Those compounds which show the greatest levels of inhibition are classified as ‘potent hits’ (also called ‘good hits’) and the remaining hits are classified as just a ‘hit’. On the other hand, the false hits are

sorted into nine different groups according to the reason for the false hit. Figure 3.1 shows typical images for a non-hit, hit and potent hit.

Table 3.1: Description of variables

Variable	Description
Npos	Number of positive cells in the image
Ngrains	Mean number of punctate granuli per cell
Agrains	Mean fractional area of the cytoplasm of the cells containing granuli
Fgrains	Mean fractional fluorescence within granuli compared to total cell fluorescence
Ipos	Mean Green Fluorescent Protein fluorescence intensity per cell
Ixpr	Nuclear intensity of green fluorescence within the nucleus
Itail	Mean cellular green fluorescence at the nuclear end of a series of radial spokes originating from the nucleus
Ipeak	Mean cellular peak intensity of green fluorescence found along the radial spokes
Rpk-tl	Mean cellular Ipeak / Itail
Rpk-xp	Mean cellular Ipeak / Ixpr
Dpeak	Mean cellular distance along the radial spoke to the position of peak intensity of green fluorescence
Iline	Mean cellular intensity of green fluorescence averaged across the length of the radial spokes
Dwght	Mean of the intensity weighted distribution along the radial spokes
Imrk	Mean intensity of the red nuclear marker
Amrk	Mean area of the red nuclear marker

Figure 3.1: Examples of images

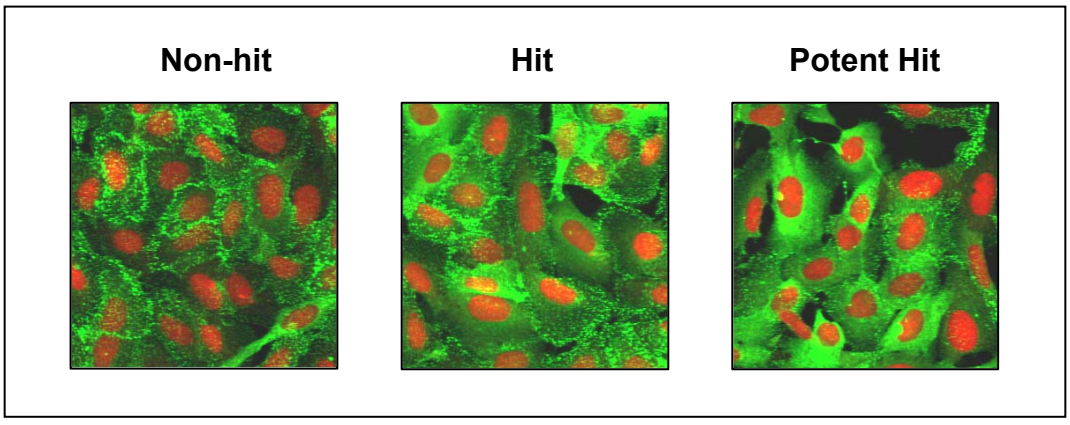


Figure 3.2 shows images for four types of false hit: 'toxicity', 'high background', 'overconfluent' and 'focus error'. The high background error is caused by a compound being fluorescent. If a compound is fluorescent then it causes problems with the granularity algorithm because it is looking for punctuate regions of fluorescence, but the image has a higher overall level of fluorescence than is usual. If there are too few or too many cells in the well then the error is termed 'underconfluent' and 'overconfluent' respectively. The cells have to be grown before the screening takes place; if they grow on top of each other then this leads to them being overconfluent. The nuclei of the cells are dyed red (the red dye used binds to DNA which is contained in the nuclei) during the automated assay procedure. If there is a problem with this procedure then this leads to a 'low Draq5' error. Table 3.2 shows the nine different types of false hit with an explanation for each.

Figure 3.2: Images of false positives

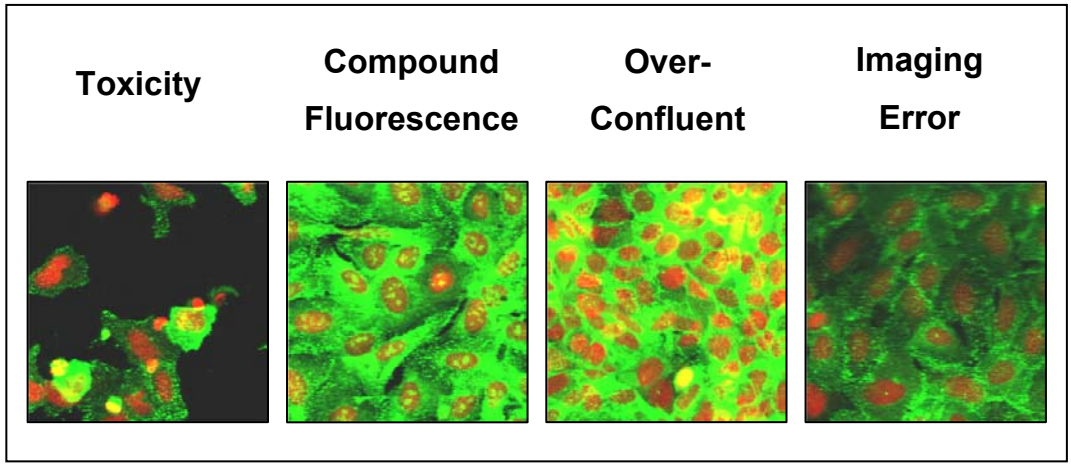
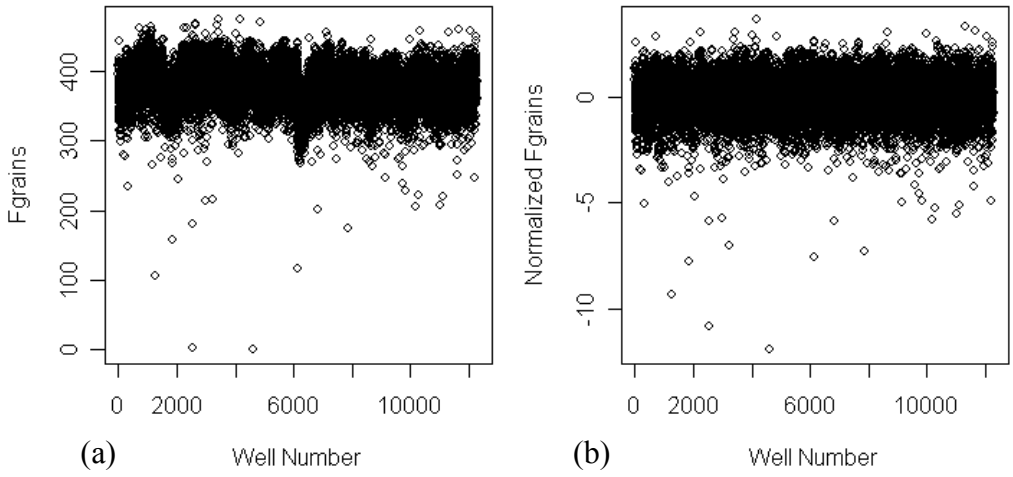


Table 3.2: Causes of false hits

Cause	Explanation
Toxicity	Possibly due to plate or compound
Focus Error	Cells in focus not inhibited
No visible image	No discernable cells
Underconfluent	Unreliable score, no clear inhibition
Overconfluent	Resulting in unreliable underestimated score
Low Draq5	Poor nuclear stain but no apparent inhibition
High background	Not a hit but possibly fluorescent compound
Foreign particle	But no apparent inhibition
Well dry	Poor focus but no evidence of inhibition

The data available from the screening experiment were collected in three batches. The first of 12,288 compounds were selected because of their known properties and were used in a pre-screen to validate the experimental procedures. The data from this batch form the training data. The remaining two batches of 33,941 and 33,408 compounds (labelled A and B respectively) yielded the test data for classifying each of these 67,349 compounds.

Figure 3.3: Example of data scaling:
(a) before scaling and (b) after scaling.



All data that are produced from the experiment are normalized to the median to account for the possibility of 'plate slip'. A plate is considered to have 'slipped' when all the values for all the variables are found to be lower than on all the other plates. This is exemplified in Figure 3.3. The left hand plot shows the Fgrains parameter plotted against the well number for the raw data. It can be seen that the compounds on the plate with well numbers of approximately 6000 have lower Fgrains values than the other plates and may therefore have an effect on any analysis which is carried out on the data. Using the scaling

$$Fgrain\ norm = \frac{x - median_p}{\sigma_p},$$

where $median_p$ and σ_p are respectively the median and standard deviation corresponding to plate p, the slip effect can be removed. The results of this scaling are shown in the right hand plot in Figure 3.3, which shows the normalized Fgrains parameter plotted against the well number. It can be seen that there is no longer any evidence of plate 'slip'.

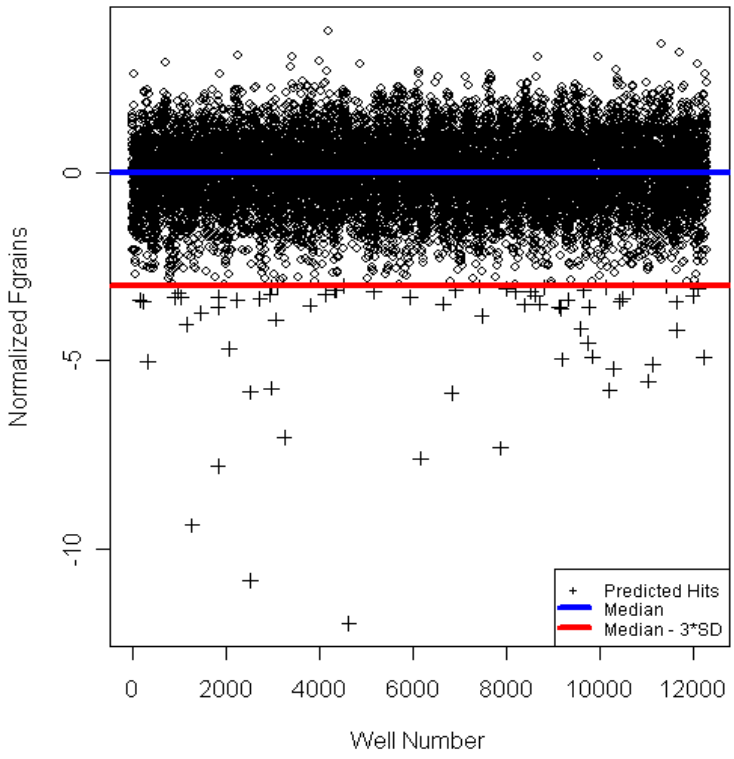
3.2.3 Review of Current Methodologies

Early approaches for identifying hits from high content screening data involved the use of a single parameter selected as the most sensitive during assay development. Hits are identified as those compounds whose measurements deviate from the majority of measurements on the same plate. The current practice is to select compounds that differ from the median by c standard deviations, where c is a preliminary chosen constant (Gagarin *et al*., 2006). For the data set analysed here the Fgrain parameter is used for filtering as it was found to be the most instructive in identifying activity of compounds. In this case, an observation is considered to be a statistical outlier if it is more than three standard deviations away from the median and since the objective is to find an antagonist a 'hit' is expected to have a low value for this parameter. However, a low Fgrain value can also occur when there are false positives so all wells selected as hits have to be manually checked by eye so that these wells can be excluded (Cooke *et al*., 2003). In addition, the images from a small random sample of non-hits are also checked by eye to ensure that the experiment and method of selection is working correctly. Figure 3.4 shows the process of selecting hits using this approach.

Recent developments in the analysis of high content screening data have focused on investigating the implementation of multivariate classifiers. Huang and Murphy (2004) and Zhou *et al.* (2007) compare classification using K-nearest neighbours, neural networks, support vector machines, Gaussian mixture models and decision trees with high content screening data from location proteomics and

time-lapse fluorescence microscopy respectively. Svetnik *et al.* (2003) made a comparison of the random forest classifier (see Section 2.3.3 for details) with other classifiers for predicting the activities of a compound based on a quantitative description of its molecular structure. The random forest was found to have the highest accuracy amongst all of the classifiers compared. A general review of classifiers and statistical modelling of high content screening data can be found in Zouh and Wong (2006) and Ainscow (2007b).

Figure 3.4: Hit selection using a single parameter



To show the potential of multi-parameter analysis of high content screening data a statistical pilot study (Mills, 2004) considered a refined selection of compounds from a data set previously analysed using the one-parameter approach. This refined analysis enabled the removal of 'false positives', arising from compounds that were, for example, intrinsically fluorescent or toxic. In this way, the number of selected compounds was reduced and therefore enabled rapid progression of the most likely candidate drugs (Cooke *et al*., 2003).

3.2.4 Objectives

The aim of this high content screening experiment is to discriminate between compounds that inhibit a biological process and those that do not (see Section 3.2.1). The main objective of the data analysis is to develop a multi-parametric approach to hit selection that uses the full potential of the information that is extracted from the high content images. AstraZeneca would like a reliable automated method of hit selection so that manual image inspection, which is slow and subjective, can be minimized. It is therefore important to reduce the number of false positives (i.e. false hits and non-hits that have been misclassified as true hits) as these generate unnecessary additional costs through manual image inspection. At the same time it is also important not to have many false negatives (i.e. true hits that have been misclassified as non-hits) because this may result in compounds with the potential to be developed further being ignored. However, as there are only a limited number of compounds that can be taken forward to the next stage of screening the most important factor is to reduce the amount of manual image inspection.

3.2.5 Exploratory Analyses

The exploratory analysis for this case study involves the 12,288 compounds used in the pre-screen validation experiment and takes several forms. The structure of the data is investigated to determine whether it is possible to discriminate between the pre-defined groups. The remaining analysis focuses on applying existing classifiers to determine which produces the least numbers of misclassifications.

The existing single parameter approach described in Section 3.2.3 relies on hits being outlying from non-hits for correct classification to be possible. It is therefore important when considering multi-parametric classification to investigate the structure of the variables to see if it is possible to discriminate

between the different groups. Visual inspection of scatterplots and kernel density estimates for the different groups of each variable reveals some common features of the individual parameters.

Figure 3.5: Scatter plots and kernel density estimates for Agrains variable

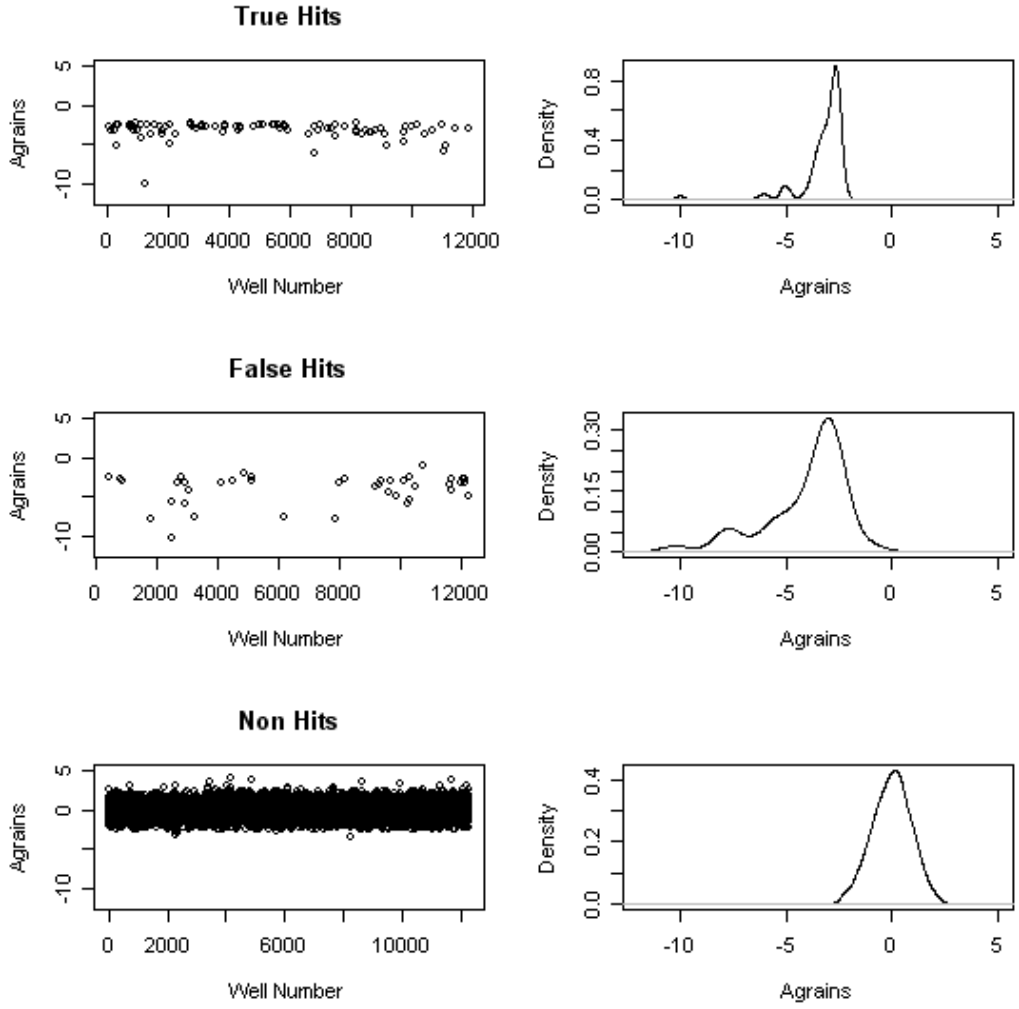
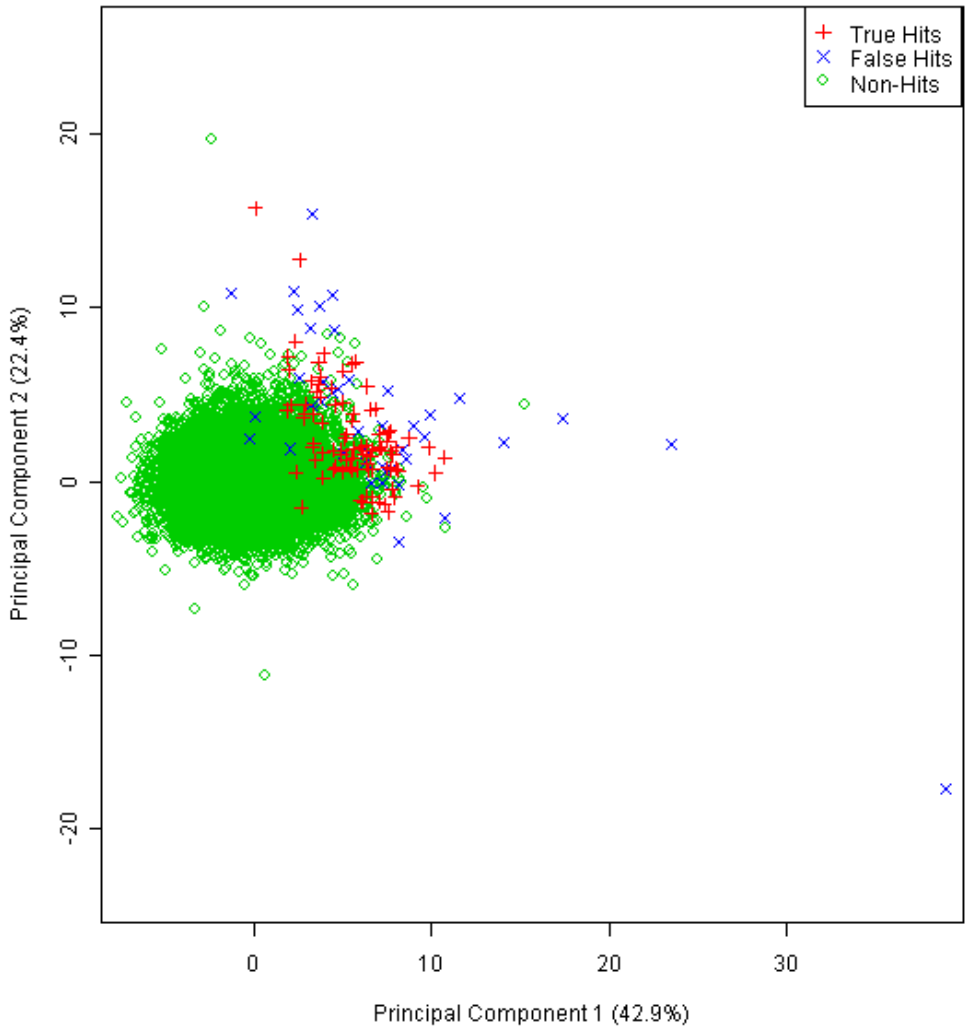


Figure 3.5 shows scatter plots and plots of kernel density estimates of the Agrains parameter for each of the three groups (true hits, false hits and non-hits). Examining the scatter plots it can be seen that the majority of the points for the non-hits have values with range between -2 and 2 whereas the majority of points for the true and false hits have values that are less than -2. This suggests that when using the Agrains parameter there maybe some difficulty in distinguishing between the true hits and false hits. This is further reflected in the kernel density

plots. Here it can see that the densities of the true hits and false hits both have peaks at approximately -4 with a large proportion of the densities overlapping. On the other hand, only the left hand tail of the non-hits overlaps with right hand tails of the true hits and false hits. This suggests that it should be much easier to distinguish between the non-hits and the other two groups using this Agrains parameter. Scatter plots and kernel density estimates of the remaining variables were also examined but are not shown here.

Figure 3.6: Principal component plot of the training data



To further assess the ability to discriminate between the three groups principal component analysis was used to visualize the data. This was performed on the training data with the true hits, false hits and non-hits being determined by the

single parameter classification approach. As the variables are of mixed type the correlation matrix was used for analysis; in addition to the two principal components that are displayed in Figure 3.6, plots of higher components were also examined (not shown here) but found they did not add any extra information about the structure of the data. (The principal component methodology is described in Section 2.3.5).

The plot of principal components in Figure 3.6 reveals similar results to the scatterplots and kernel density estimates in Figure 3.5. In particular, those points that represent true hits and false hits occupy the same space on the plot suggesting that discriminating between these groups maybe difficult. In addition, both the true hit and false hit groups overlap with the edge of the non-hits suggesting that a model may misclassify some of the observations within these groups. However, it is important to consider that the groupings of the observations in both Figures 3.5 and 3.6 have been determined by using the single parameter approach and therefore only the classifications of the true hits and false hits are certain (these compounds have been checked by eye whereas the non-hits have not).

In order to assess some of the existing supervised classification methodologies, an artificial problem was created using the 12,285 compounds used in the pre-screen validation experiment. After classifying these compounds using the single parameter methodology described in Section 3.2.3, the data was used to create a training and test set. A random sample of 1000 non-hits was placed in each of the two sets before the true hits and false hits were randomly split between them. This gave a training set consisting of 1261 compounds (1000 non-hits, 158 false hits and 103 true hits) and a test set consisting of 1260 compounds (1000 non-hits, 157 false hits and 103 true hits).

The results of classification using the existing multi-parametric methodologies are displayed in Table 3.3. The first column indicates the method of classification being used (linear discriminant analysis, quadratic discriminant analysis, classification trees or random forests) while the second column indicates which

data (training or test) is being used for evaluation. The remaining four columns show the overall misclassification rate, the percentage of true hits misclassified, the percentage of false hits misclassified as true hits and the percentage of non-hits misclassified as true hits respectively. When comparing results it is important to remember that the results of classifying the test data give a better indication of performance because they have not been used to train the classifier and are not biased.

Table 3.3: Comparison of multivariate classifiers

Method	Data	Overall Misclassification Rate	Percentage Misclassified		
			True Hits	False Hits As True Hits	Non Hits As True Hits
LDA	Train	7.06%	21.34%	7.59%	1.00%
LDA	Test	9.52%	24.27%	13.38%	1.10%
QDA	Train	9.13%	5.83%	17.72%	3.90%
QDA	Test	12.30%	13.59%	21.66%	4.80%
C. Tree	Train	4.58%	23.30%	9.49%	1.00%
C. Tree	Test	8.57%	29.13%	16.56%	1.40%
R. Forest	Train	0%	0%	0%	0%
R. Forest	Test	6.59%	25.24%	11.52%	1.20%

Comparing the results for the test data in Table 3.3, it can be seen that the smallest overall misclassification rate of 6.59% is produced when using the random forest classifier. The misclassification rates for the true hits are all too high with linear discriminant analysis, classification trees and random forests classifying over 20% of this group incorrectly. A high misclassification rate for the true hit group implies that many compounds that may have the potential to be developed further are being ignored as false negatives or non-hits. Although the misclassification rates for the false hits are smaller than those of the true hits when using linear discriminant analysis, classification trees and random forests, these values still represent many compounds that are incorrectly being classified as hits. This evidence of misclassification of the true hits and false hits suggests that none of the classifiers considered in Table 3.3 are suitable for the classification of this data

in their current form and therefore new methods of classification need to be investigated in order to meet the objectives set out in Section 3.2.5.

3.3 Dose Response Clustering

The data described in this section relate to a high content screening experiment designed to assay chemical compounds for hepatotoxicity (a compound is said to be hepatotoxic if it is toxic to liver cells). Specifically, the data represent dose responses from compounds applied to liver cells to analyse indicators of the altered metabolism of phospholipids (this altered metabolism of phospholipids leads to phospholipidosis). This type of screen is an important step in the evaluation of potential drugs because drug induced liver injury is the most common cause for non-approval, withdrawal, limitation in use, and clinical monitoring by the Food and Drug Administration (Ainscow, 2007a).

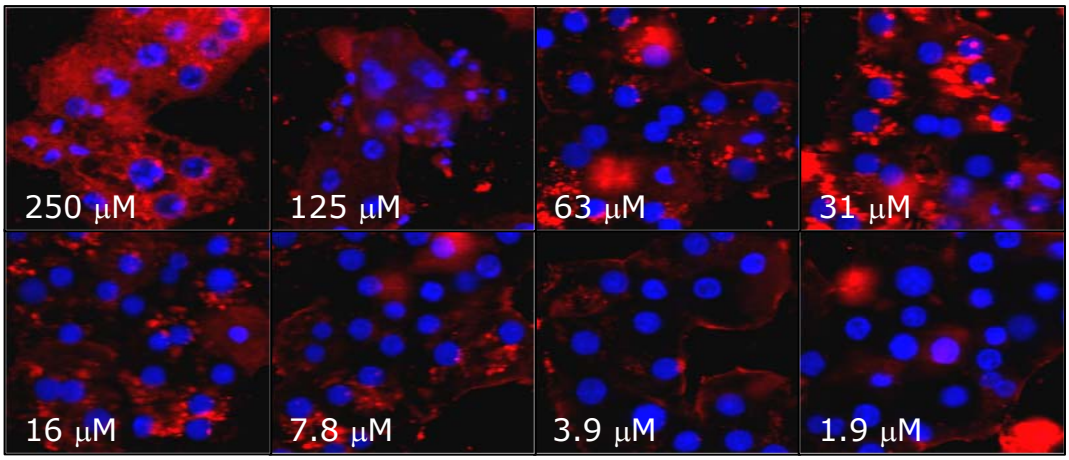
The remainder of this section is organised as follows. Section 3.3.1 describes the data and method of collection before Section 3.3.2 outlines the objectives of analysis. Finally, Section 3.3.3 describes some exploratory analysis of the data set. Note that the review of current methodologies from the literature for this type of data is not described in this chapter but forms a section of Chapter 7.

3.3.1 Data Description

Compounds were processed through the screen on plates consisting of eight rows of wells. Approximately 250 cells were added to each well and then dosed with one of 850 compounds with a range of different doses (8 concentrations). Each row of wells on a plate had one compound at each of the 8 doses and 2 control compounds. The control compounds were the same for each row and consisted of one positive control (i.e. a compound that was toxic and showed strong signs of

phospholipidosis) and one negative compound (i.e. a compound that was non-toxic and showed no signs of phospholipidosis). The 850 compounds were made up of a mixture of compounds with unknown properties and compounds that had known toxicological properties. The compounds with unknown properties had 2× dose responses with concentrations between 1.9μM and 250μM. The reference compounds had 2× dose responses with concentrations between 0.08μM and 1000μM. The majority of the reference compounds were replicated several times.

Figure 3.7: Examples of dose response images



For each well a high content image was taken and advanced imaging algorithms produced a set of quantified variables. An example of images from each of the 8 dose concentrations of one compound is given in Figure 3.6. The imaging expert indicated that the images show a toxic response for doses above 125μM, doses between 16μM and 63μM show signs of phospholipidosis and doses below 7.8μM show no toxicological effect.

There are three imaging algorithms that are used to produce the measured variables from the high content images. The main test algorithm, GRN, analyses the extent of dye accumulation in punctate regions of the cells; analysis of the nuclei is conducted by the NUC algorithm (this can only be carried out on non-

viable cells); and the OBI algorithm analyses the nuclear area, intensity and cytoplasmic label. Table 3.4 gives details of the variables that are produced by each of the algorithms. Note that each of the imaging algorithms produce raw data based on the measurements of the individual cells in each image but only the means and standard deviations of these measurements are available for the full 850 compounds. The raw data values are available for 197 compounds. In addition, values are calculated from two viable subsets. The first subset consists of all cells in the images and a total of 18 variables are produced; the second subset consists of all 'viable' or live cells and a total of 11 variables are produced (this is because the 7 variables measured by the NUC algorithm produce values of 0 for viable cells). Information is available on the cells count in each image along with the viable cell count.

Table 3.4: Description of variables associated with the imaging algorithms

Algorithm	Variable	Description
GRN/NUC	Ngrains	Number of puntate regions found in cells
GRN/NUC	Agrains	Fractional area of the cell defined as being puncta
GRN/NUC	Fgrains	Fractional proportion of cell label found within puncta
GRN/NUC	Dnuc	Diameter of nuclear marker
GRN/NUC	Inuc	Mean pixel intensity of label in nucleus
GRN/NUC	Icyt	Mean pixel intensity of label in area analysed for puncta
GRN/NUC	Igrains	Pixel intensity of label within puncta
OBI	Imrk	Mean pixel intensity of nuclear marker
OBI	Amrk	Area of nuclear marker
OBI	Isig	Mean pixel intensity of label in perinuclear region
OBI	Asig	Area of perinuclear region

Table 3.5 shows the classification of the 11 reference compounds used in the screen. The compounds are classified both for toxicity and phospholipidosis; the positive or negative label for a compound relates to the phospholipidosis and any remaining label relates to toxicity (for example, Amiodarone shows strong signs of phospholipidosis and is toxic). In addition to the classifications, the screening expert also indicated that Amiodarone and Metformin are used as the plate positive and negative controls respectively and that Memantidine is related to

Amantidine in structure but is more potent. Further to the classifications of the 11 compounds, there are also classifications available for 289 of the 850 test compounds. These classifications will allow observations to be made on the effectiveness of any method of clustering.

Table 3.5: Classification of reference compounds

Compound	**Class**
Amantidine	Weak Positive
Amiodarone	Strong Positive, Toxic
CCCP	Negative, Toxic
Fluoxetine	Strong Positive
HT0026	Strong Positive
HT0027	Strong Positive
Maprotilene	Strong Positive
Memantidine	Positive
Metformin	Negative, Non-Toxic
Rimantidine	Positive
Tacrine	Weak Positive

3.3.2 Objectives

The primary aim of the analysis of this dose response high content screening experiment is to investigate how to cluster compounds in terms of their toxicological effect on cellular assays. In particular, it is of interest to compare the toxicological effect of clusters of compounds with the 11 reference compounds that were introduced in Table 3.5. This would allow the screening experts to evaluate compounds (which have unknown toxicological effects) in terms of phospholipidosis and other forms of toxicity. As discussed at the beginning of Section 3.3, drug induced liver injury is the most common cause for non-approval, withdrawal, limitation in use, and clinical monitoring by the Food and Drug Administration and therefore investigating toxicity forms an important step in the evaluation of potential drugs (Ainscow, 2007a).

The primary aim of the analysis for this experiment leads to a key problem: how to compare compounds with different dose response ranges. This problem is important to the investigation because the dose responses of the compounds with unknown properties have been measured at different concentrations to the dose responses of the 11 reference compounds (see Section 3.3.1). Therefore a method is required that will allow compounds to be compared at different ranges of dose so that the screening expert can be shown evidence to suggest that, for example, compound A has the same response over doses of 1.9μM to 15.6μM as compound B over doses of 31.25μM to 250μM.

3.3.3 Exploratory Analyses

The exploratory analysis for this data set takes a number of different forms. Firstly the reference compounds were investigated to see how the response changes as dose increases and to compare the effects of compounds with different properties (i.e. those which are negative positive and toxic). The second part of the analysis concentrates on visualising both the reference compounds and the compounds with unknown properties using principal component analysis. Throughout the exploratory analyses, comparisons between results using measurements from all cells and using measurements from only live cells shall be made.

Figure 3.8 shows dose response values of the granularity (GRN) Ngrains parameter for 8 of the reference compounds. The mean values (calculated using the measurements from all the individual cells on the image) have been plotted for the eight doses and the curves between these points have been calculated using cubic spline interpolation (see Venables and Ripley (2002) for details). It can be seen that for some of these reference compounds there is more than one curve; in these cases the compound has been replicated on a number of different plates. Examining these replicates shows that there is some variability between the values found on different plates but the overall shapes of the curves are the same.

Comparing reference compounds with different classifications (see Table 3.5) it can be seen that some of the compounds with different classifications have similar responses for this Ngrains parameter. For example, Metformin which is classified as being a negative compound has a similar response to Amantidine which is classified as being a weak positive compound. In addition, the two compounds that are structurally related to each other (Amantidine and Memantidine) have very different responses with Amantidine showing little difference in response over the dose range and Memantidine showing a large drop off in response for the highest dose.

Figure 3.8: Dose response plots of reference compounds

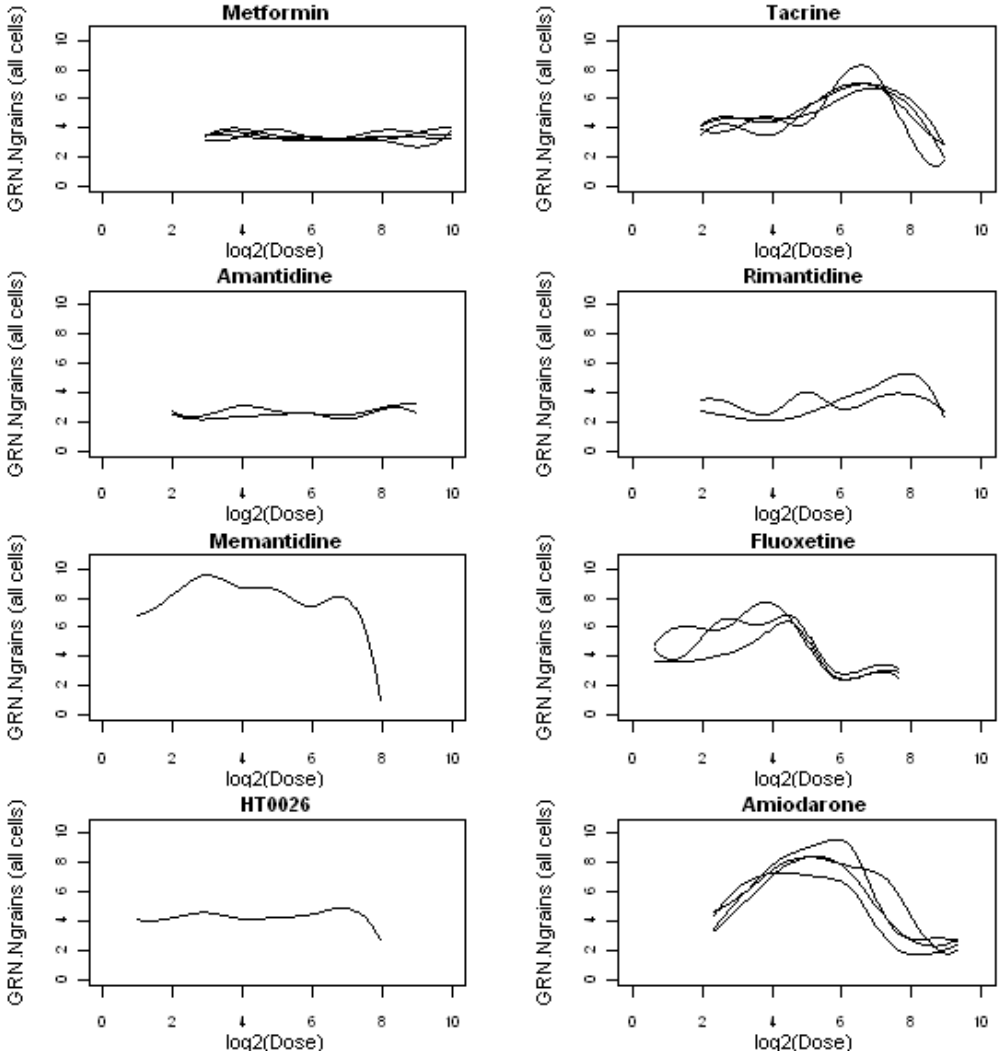


Figure 3.9: Kernel density plots of the Tacrine compound

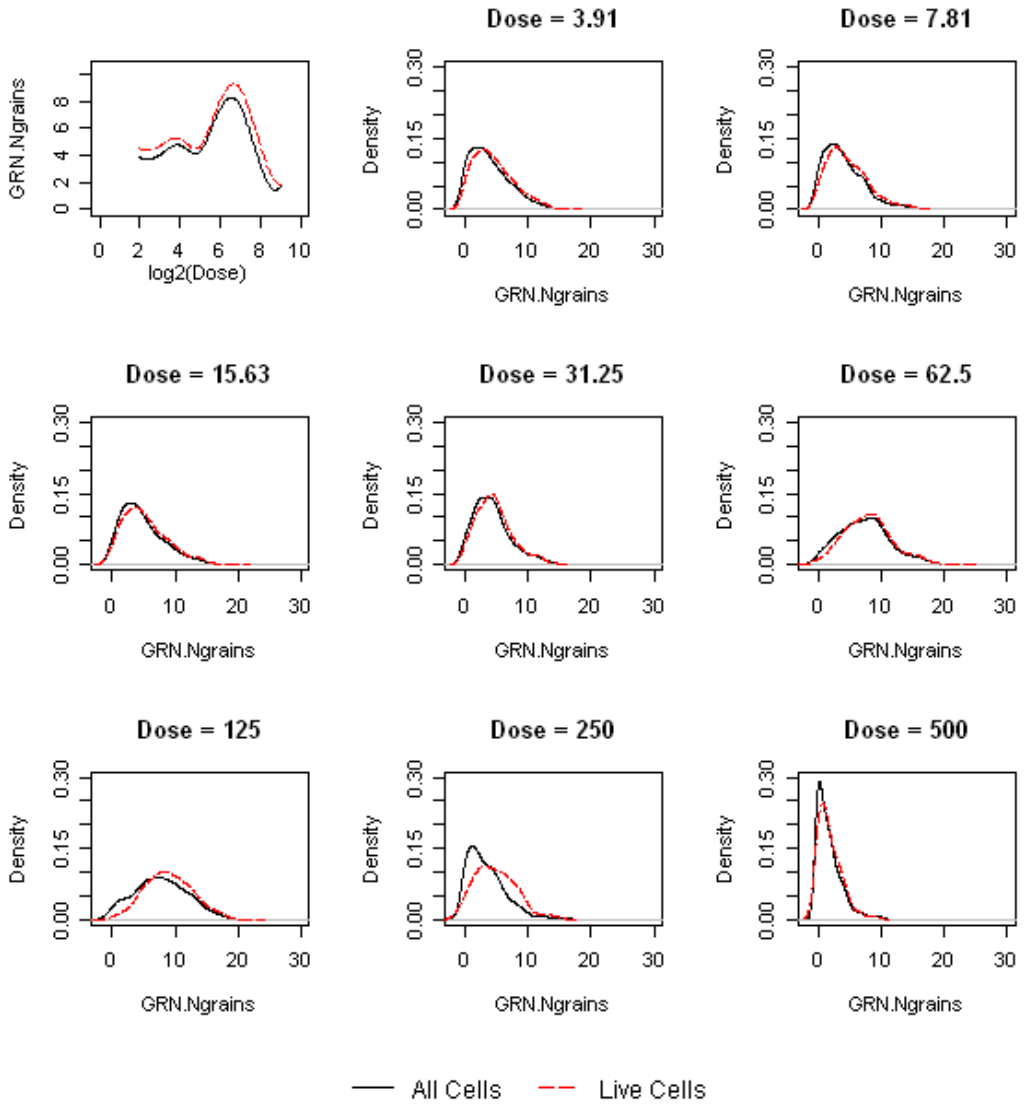
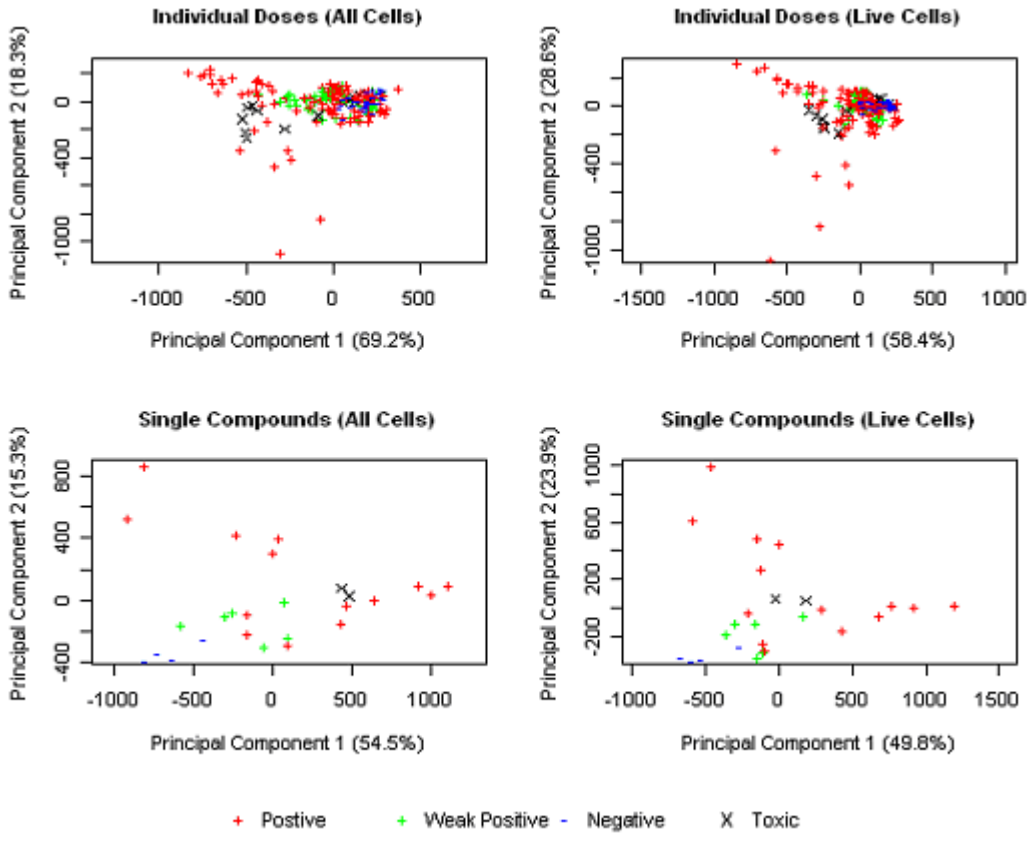


Figure 3.9 shows the distributions of the dose response values of the granularity (GRN) Ngrains parameter for one of the Tacrine reference compounds. The top left-hand plot shows the log dose plotted against the parameter values calculated from the individual cells in the high content image (as in Figure 3.8). However, unlike the plots in Figure 3.8, values have now been plotted for both the case when all cells are used for analysis and for when only the live cells are used. The remaining eight plots show kernel density estimates of the data corresponding to the eight mean parameter values. By examining these plots it can be seen how the distributions of the single cell values change as the dose increases. It will also be

possible to compare the different densities that are produced when using all the cells and when using only the live cells.

From the plots in Figure 3.9 it can be seen that for the majority of doses there is little difference between the densities when all cells are used and when only live cells are used. However, concentrating on the plots for doses 125µM and 250µM shows that the distributions when using all cells are positively skewed while the distributions when using the live cells are much more symmetric. This suggests that there are more cells dying at these doses; this is to be expected as they are the larger doses of the Tacrine compound. There is further evidence of this in the 500µM dose with both densities being positively skewed but the density for all cells having a larger peak.

Figure 3.10: Principal component plots of the reference compounds



Principal component analysis (see Section 2.3.5) was used for exploratory visualisation of the multivariate data. This analysis was conducted in several stages with the first stage using all 850 compounds with each of the eight doses displayed as different observations. The principal component plot produced (not shown here) was not informative due to the large number (6800) of observations that were plotted. Therefore the next stages of analysis concentrated on visualising the reference compounds.

Figure 3.10 shows four principal component plots of the reference compounds (note that plots are displayed only for the first two principal components but plots of larger components have also been examined). The top two plots show individual doses of compounds plotted as separate observations with the left hand plot using data from all cells and the right hand plot using data from those cells which were determined to be live. The bottom two plots show individual compounds (i.e. the data from each dose are combined in a single observation). Again, the left hand plot uses data from all cells and the right hand plot uses data from those cells which were determined to be live.

Examining the two plots in Figure 3.10 that are based on individual doses it can be seen that in both cases (using all cells and using only live cells) there is no separation between compounds that are classified as being negative and those that are classified as being positive. This suggests that when using the individual doses of compounds as observations it is not possible to distinguish between clusters of reference compounds with different classifications; hence, classification of future compounds with unknown properties would be difficult. Examination of the plots of individual compounds shows that there is more separation of the groups than when using individual doses. In particular, the negative compounds are in a separate cluster to the positive compounds. However, there are no clearly defined clusters for the positive and weak positive compounds.

The results of comparing analyses (see Figures 3.9 and 3.10) that have been conducted using measurements from all cells and using measurements from only live cells have shown that there is little difference between the two. However, it is suggested that there may be extra information contained in those cells which are considered to be dead or non-viable and will therefore conduct future analyses using measurements from all cells.

3.4 Summary

This chapter has provided an introduction to the case studies that are motivating the work in this thesis. In each case a description of the high content screen in question has been given and the relevant study aims highlighted. Although the data sets described are from specific high content screens it is envisaged that the methodologies developed in the later chapters will be widely applicable to other areas both within and outside of screening experiments.

Exploratory analysis for the compound hit selection problem in Section 3.2 has identified a number of key features and problems. Visualisation of the pre-screen data using principal component analysis has shown that it may be difficult to distinguish between the three groups of interest (especially the true hits and false hits) and therefore any method of classification would have to take this into account. This was further reflected in the number of true hits that were misclassified as false hits when examining the results of classifying this data using a number of existing multivariate classifiers. It is important to overcome this problem because reducing the number of false positives and false negatives was one of the main objectives set out in Section 3.2.5.

The classifications that were carried out during exploratory analysis in Section 3.2.6 showed that the multi-parameter classifiers (linear discriminant analysis, quadratic discriminant analysis, classification tress and random forests) did not meet the objectives that were set out in Section 3.2.5. This suggests that none of

these classifiers are suitable for analysing the data in their current form. Chapters 4, 5 and 6 will continue the analysis of this data by looking at different methods of updating classification rules. Chapter 4 will begin the investigation by using unlabelled data to update classification rules.

The exploratory analyses that were conducted for the dose response clustering problem in Section 3.3.3 all concentrated on methods of visualising the compounds and their responses. However, none of methods (with the exception of principal component analysis for individual doses) have taken into account the problem of how to compare responses over different dose ranges. It is this problem that will be key to any methodology that is designed to cluster the compounds. Chapter 7 concentrates on this problem further.

Chapter 4

Using Unlabelled Data to Update Classification Rules

4.1 Introduction

This chapter is concerned with the updating of classification rules using unlabelled data. This method of classification stems from the idea that the unlabelled data may contain useful information even though the group membership is unknown. The data from a full high content screening experiment contains many fewer observations in the labelled training data (approximately ten thousand) than in the batches of unlabelled test data (approximately one million) and therefore large amounts of potential extra information is being ignored if classification rules are not updated.

In Chapter 3, four multivariate classifiers (linear discriminant analysis, quadratic discriminant analysis, classification trees and random forests) were investigated to analyse the high content screening data. The results of these analyses showed that in their current form the classifiers did not meet the objectives that were set out in Section 3.2.5. Hence, updating the rules using the methodologies described in this chapter is the next step in trying to find a suitable method of classifying high content screening data.

The idea of using unlabelled data originated with Hartley and Rao (1968) who used a combination of labelled and unlabelled data for estimating maximum likelihood and classification. It has been used in various forms since then and in particular an iterative procedure for classification of two groups based on both labelled and unlabelled data was introduced by McLachlan (1975) and was found to be asymptotically optimal (i.e. minimising the risk of misclassifying a randomly chosen observation) when the number of unlabelled observations tends to infinity and the number of labelled observations from the groups is moderately large. More recent applications have considered problems with more than two groups with Dean *et al.* (2006) applying the idea to data from food authenticity studies. It is the outline of the methodology in this paper that forms the basis of this chapter.

The remainder of this chapter is structured in the following way. The first section outlines the methodology that is used in the remainder of the chapter. Sections 4.3 and 4.4 describe the application of the methodology to data in two different forms. Firstly, Section 4.3 describes the application of updating using unlabelled data when using the artificial data set that was created from the pre-screen. It also extends the updating methodology by introducing and applying the ideas of robust estimation of multivariate location and scale, and the reject classification option. Section 4.4 then applies the methodology to a version of the full data set with a training set and two batches of test data. Section 4.5 concludes the chapter with a summary and some discussion.

4.2 Methodology

This section outlines the methodology of using unlabelled data to update classification rules. In particular, Sections 4.2.1 and 4.2.2 introduce model based discriminant analysis and model selection respectively before Section 4.2.3 outlines the Classification Expectation Maximisation (CEM) algorithm. Finally, Section 4.3.4 introduces some methods for the robustly estimating multivariate

location and scale which shall be applied as an extension to the current methodology.

4.2.1 Model Based Discriminant Analysis

Model based discriminant analysis is an extension of the maximum likelihood approach to linear discriminant analysis that was described in Section 2.3.1. The basis of this model-based approach is to use constraints to impose restrictions on the volume, shape and orientation of the groups in the data. These imposed restrictions are implemented through the eigenvalue decomposition of the covariance matrix Σ_g for each group. These covariance matrices can be written in the form

$$\Sigma_g = \lambda_g D_g A_g D_g^T$$

where D_g is an orthogonal matrix of eigenvectors of Σ_g, the A_g is a diagonal matrix with entries proportional to the eigenvalues of Σ_g (and first eigenvalue equal to 1) and λ_g is a proportional constant. Each component of the eigenvalue decomposition has a different morphological interpretation in terms of the data in the group. The matrix D_g governs the orientation of the group, A_g controls the shape and λ_g controls the volume. By imposing restrictions on some or all of the different components of the eigenvalue decomposition different models are formed (Dean *et al*., 2006).

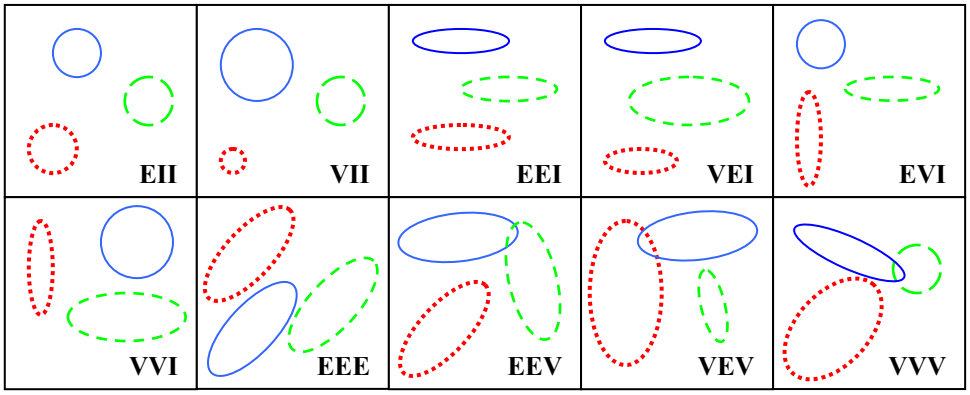
Table 4.1 contains details of the ten possible discriminant models that are available by imposing different restrictions on the components of the eigenvalue decomposition of the covariance matrices. The first column of the table contains the names of the different models; and the second, third and fourth columns show details of the restrictions on volume, shape and orientation respectively. The final column shows the form that the eigenvalue decomposition takes for each of the models. A visual representation of each of the ten models in Table 4.1 is shown

in Figure 4.1. Note that the model EEE corresponds to classical linear discriminant analysis and VVV corresponds to quadratic discriminant analysis.

Table 4.1: Covariance restrictions

Model	Volume	Shape	Orientation	Decomposition
EII	Equal	Spherical	-	$\Sigma_g = \lambda I$
VII	Variable	Spherical	-	$\Sigma_g = \lambda_g I$
EEI	Equal	Equal	Axis Aligned	$\Sigma_g = \lambda A$
VEI	Variable	Equal	Axis Aligned	$\Sigma_g = \lambda_g A$
EVI	Equal	Variable	Axis Aligned	$\Sigma_g = \lambda A_g$
VVI	Variable	Variable	Axis Aligned	$\Sigma_g = \lambda_g A_g$
EEE	Equal	Equal	Equal	$\Sigma_g = \lambda DAD^T$
EEV	Equal	Equal	Variable	$\Sigma_g = \lambda D_g A D_g^T$
VEV	Variable	Equal	Variable	$\Sigma_g = \lambda_g D_g A D_g^T$
VVV	Variable	Variable	Variable	$\Sigma_g = \lambda_g D_g A_g D_g^T$

Figure 4.1: Cluster shapes allowed by covariance restrictions[1]



[1] Adapted from Dean *et al.* (2006)

4.2.2 Model Selection

In discriminant analysis the grouping of the observations in the training set and the number of groups are known. Therefore, a natural way to choose a model is to select the one that minimizes the sample based estimate of future misclassification by cross-validation. In other words, for the ten available discriminant models a leave-one-out method is employed to calculate a cross validation error. The model which minimizes this error is then applied. However, in many circumstances several models may provide exactly the same cross-validated misclassification rate. In such a case, the most parsimonious or simplest model is selected; that is to say, tied models are ranked using the following criteria: at first, a spherical model is preferred to a diagonal model which is preferred to a non-diagonal model; secondly, a model with different volumes is preferred to a model with different shapes which is preferred to a model with different orientations (Biernacki and Govaert, 1999).

4.2.3 Classification Expectation Maximization (CEM) Algorithm

Let $x_N = (x_1, x_2, ..., x_N)$ denote the labelled data with labels $l_N = (l_1, l_2, ..., l_N)$ where $l_{ng} = 1$ if observation n comes from group g and $l_{ng} = 0$ otherwise. Let the unlabelled data be denoted by $y_M = (y_1, y_2, ..., y_M)$ with unknown labels $z_N = (z_1, z_2, ..., z_M)$ that are defined in the same way as the known labels. The four steps of the CEM algorithm are then defined as:

Step 1

Set $k = 0$. Find the starting values for the algorithm by using the model-based discriminant analysis estimates of the parameters in the model. Call these estimates $\hat{\pi}^{(0)}$ and $\hat{\theta}^{(0)}$.

Step 2

Calculate the expected value of the unknown labels using the formula

$$w_{mg} = \frac{\hat{\pi}_g^{(k)} f\left(y_m \middle| \hat{\theta}_g^{(k)}\right)}{\sum_{g'}^{G} \hat{\pi}_{g'}^{(k)} f\left(y_m \middle| \hat{\theta}_{g'}^{(k)}\right)} \text{ for g = 1,2,…G and m = 1,2,…,M}$$

where

$$f\left(y_m \middle| \hat{\theta}_g^{(k)}\right) = \frac{1}{\sqrt{(2\pi)^n \left|\hat{\Sigma}_g^{(k)}\right|}} \exp\left(-\frac{1}{2}\left(y - \hat{\mu}_g^{(k)}\right)\left(\hat{\Sigma}_g^{(k)}\right)^{-1}\left(y - \hat{\mu}_g^{(k)}\right)^T\right).$$

Then the expected class of the unknown labels is calculated by

$$\hat{z}_{mg}^{(k+1)} = \begin{cases} 1 & \text{if } w_{mg} > w_{mg'} \text{ for all } g' \neq g \\ 0 & \text{otherwise} \end{cases}.$$

Step 3

Using the data, the known labels and the current estimates of the unknown labels (from step 2) estimate the parameters by

$$\hat{\pi}_g^{(k+1)} = \frac{\sum_{n=1}^{N} l_{ng} + \sum_{m=1}^{M} \hat{z}_{mg}^{(k+1)}}{N + M} \quad \text{for g = 1, 2,…,G}$$

$$\hat{\mu}_g^{(k+1)} = \frac{\sum_{n=1}^{N} l_{ng} x_n + \sum_{m=1}^{M} \hat{z}_{mg}^{(k+1)} y_m}{\sum_{n=1}^{N} l_{ng} + \sum_{m=1}^{M} \hat{z}_{mg}^{(k+1)}} \quad \text{for g = 1, 2,…,G}$$

The estimation of the covariance matrices Σ_g depends on the constraints that are put on the eigenvalue decomposition of the matrix. Details of the different covariance matrices can be found in Bensmail and Celeux (1996).

Step 4

Check that the algorithm has converged. The algorithm should be stopped when the $\hat{z}_{mg}^{(k)}$-values are equal to within specified limit on two consecutive iterations. If the algorithm has not converged, set k = k+1 and repeat steps 2-4.

4.3 Application to Pre-Screen Data

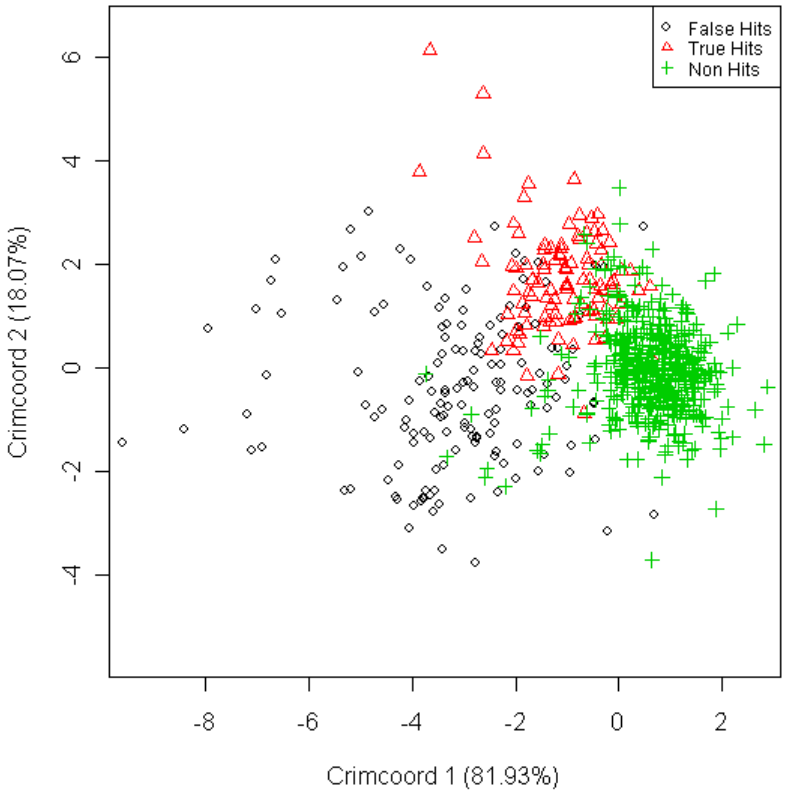
This section describes the application of the methodology to the artificial data set that was created from the pre-screen. The training set (or labelled data) from this data consists of 1000 non-hits, 158 false hits and 103 true hits; and the test set (or unlabelled data) consists of 1000 non-hits, 157 false hits and 103 true hits. The application of the methodology is computationally slow when using large data sets and therefore the aim of this first application was to use a small data set to assess the potential of the methodology.

Applying the cross validatory method of model selection (described in Section 4.2.2) it was found that the model with equal volume, shape and orientation had the smallest cross validation error. This model is equivalent to classical linear discriminant analysis and the observations from the labelled training data are plotted on crimcoords in Figure 4.2.

Examining the linear discriminant plot in Figure 4.2 it can be seen that there are a number of important features. Firstly, the three groups are not clearly defined; in particular, there is a large overlap between the cluster of true hits and non-hits, and the cluster of true hits and false hits. This feature is consistent with visualising the data using principal component analysis in Section 3.2.4 and

suggests that misclassifications of observations in all of the groups are expected. The second feature of the visualised data is the number of outlying observations from each group. All three groups appear to have some points that are outlying from their respective main cluster. This feature is considered further in Section 4.3.1.

Figure 4.2: Linear discriminant plot of training data



The results of applying the methodology from Section 4.2 to the current data set are shown in Table 4.2. The first column shows the observed classifications of the data and the remaining three columns show the number of observations that were predicted to be false hits, non-hits and true hits respectively. From this table it can be seen that 12.3% of the observations in the test data have been misclassified. This corresponds to 24.3% of the true hits being misclassified, 21% of the false hits being classified as true hits and 0.9% of non-hits being classified as true hits. These results are compared with the remaining analyses on this data set and the results from classifying during exploratory analysis in Section 4.3.3.

Table 4.2: Non-Robust Model

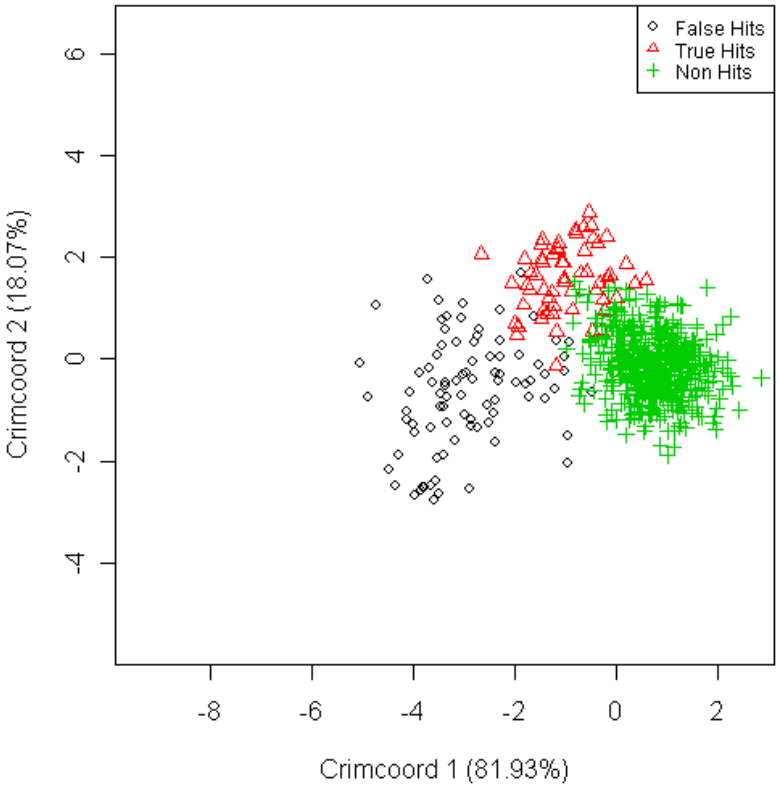
Observed Classification	**Predicted Classification**		
	False Hits	Non-Hits	True Hits
False Hits	81	43	33
Non-Hits	45	946	9
True Hits	0	25	78

4.3.1 Robust Estimation of Multivariate Location and Scale

The estimation of the multivariate location vector μ and the scale matrix Σ is important when applying the methodology described in Section 4.2. In the application to data above, these parameters were estimated using maximum likelihood and are optimal if the data come from a multivariate normal distribution. These estimates are extremely sensitive to the presence of outliers and this may cause a decline in the performance of a procedure based on these estimates. Hence, it is important to consider robust alternatives for the estimation of location and scale.

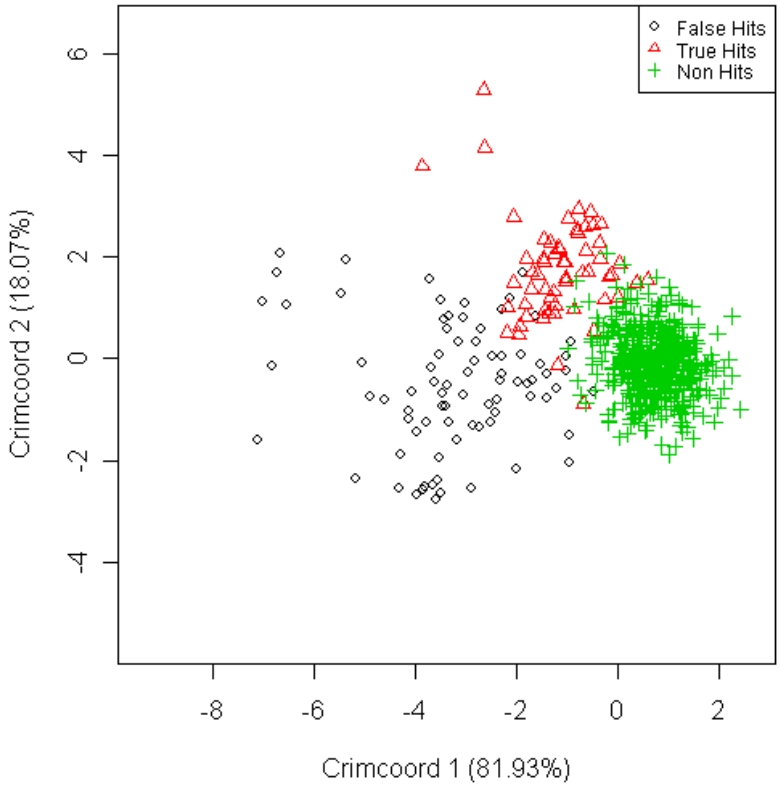
For the purpose of this work two different methods (minimum covariance determination (Rousseeuw, 1984) and minimum volume ellipsoid) were applied for robustly estimating location and scale. Give n observations and p variables, the minimum covariance determination method identifies h observations whose classical covariance matrix has the lowest possible determinant (Hubert *et al*., 2005). The minimum volume ellipsoid method seeks an ellipsoid containing $h = \left[\left(n + p + 1 \right) / 2 \right]$ observations that is of minimum volume. For both methods the estimation of location and scale is the average of the h selected observations and their covariance matrix respectively. The two methods were implemented using cov.rob in R (Venables and Ripley, 2002). A full description for the current algorithm for calculating minimum covariance determination can be found in Rousseeuw and Van Driessen (1999).

Figure 4.3: Linear discriminant plot of training data with outliers removed by minimum covariance determinant method



Figures 4.3 and 4.4 show linear discriminant plots of the training data with the outliers removed respectively by the methods of minimum covariance determination and minimum volume ellipsoid. Comparing these plots with the plot of the full training data in Figure 4.2 it can be seen that there is much less overlap between the observations in the different groups when the outliers have been removed. This suggests that linear discriminant analysis would perform better if it was trained using the data with the outliers removed. A further comparison of the two different methods that were used for the removal of the outliers suggest that the minimum covariance determination method is more strict than the minimum volume ellipsoid method; this is particularly reflected in the groups of true hits and false hits where the groups look much more homogeneous after the removal of the outliers using the minimum covariance determination method.

Figure 4.4: Linear discriminant plot of training data with outliers removed by minimum volume ellipsoid method



When using the classification EM algorithm to update the model with robust estimations of location and scale it was found that after 1000 iterations the algorithm had not converged (i.e. $\hat{z}_{mg}^{(999)} \neq \hat{z}_{mg}^{(1000)}$). Further investigation found that there were inconsistencies in the classifications of two compounds when using the minimum volume ellipsoid method and a further five inconsistencies were found when using the minimum covariance determinant method. Examining the posterior probabilities of these seven compounds revealed that the difference in the probability of being a true hit and the probability of being a non-hit was very small (with both being approximately 0.5). This suggested that the algorithm was not converging because the largest posterior probability for these observations was oscillating between the true hit and non-hit group; and that using a higher number of iterations would not make the algorithm converge. It was decided that in this case the reject classification option should be used.

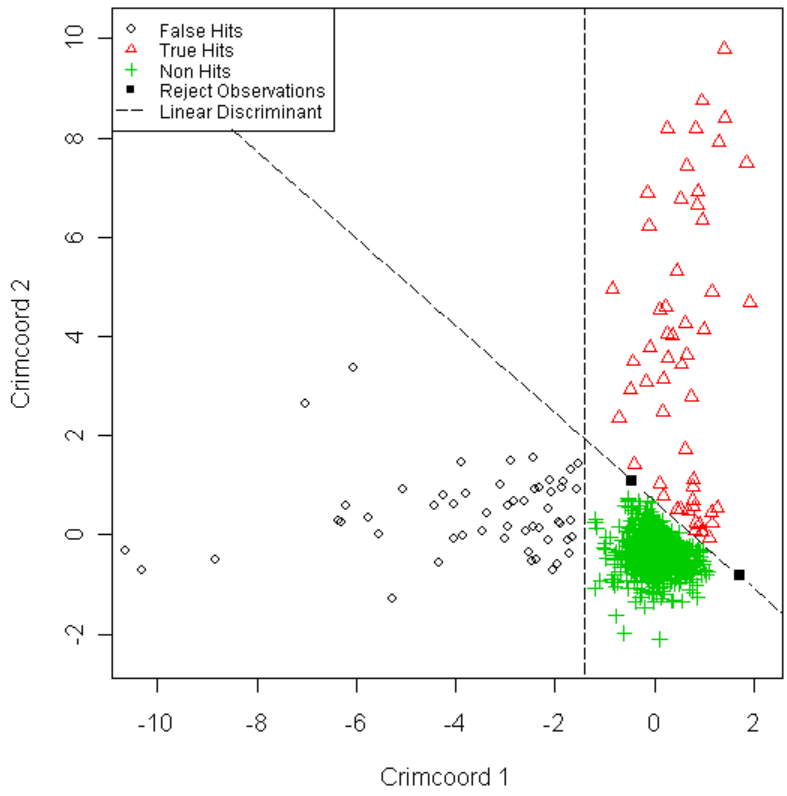
4.3.2 The Reject Option

The reject option can be used within some classification models to reduce the number of misclassifications. This method allows the final classification decision to be left to a human expert when the model is not sufficiently confident in making a decision. In this situation the model rejects the observation without allocating it to a group. Dubuisson and Masson (1993) define two different types of reject options, ambiguity rejection and distance rejection. The ambiguity rejection option relates to a situation when an observation is located between existing classes, that is, near a decision boundary. Alternatively, the distance rejection option relates to a situation when an observation is actually a member of class that is unknown or not a part of the training set.

The previous section identified that the algorithm was not converging because the largest posterior probability for a small number of observations was oscillating between two groups. These oscillations maybe caused by these observations being close to a decision boundary and therefore it is recommended that the ambiguity rejection option is applied. This means that the seven observations from the analyses that did not have consistent classifications for the 999^{th} and the 1000^{th} iteration are classified by eye by the screening expert.

Figure 4.5 is an example of a situation where the ambiguity reject option would be used. In this example it can be seen that the two observations that are rejected lie on the discriminant boundary between the true hits and non-hits and therefore a decision cannot be made by the automated classifier. This plot was created using the two observations that were rejected when using the minimum volume ellipsoid method of robust estimation and a selection of observations from the three groups. The plot is not an accurate representation of the data that were classified and is only included for illustrative purposes.

Figure 4.5: Linear discriminant plot illustrating the ambiguity reject option



The results of applying the ambiguity reject option when classifying using minimum covariance determination and minimum volume ellipsoid robust estimates of location and scale are shown respectively in Tables 4.3 and 4.4. For both tables the first column shows the possible observed classifications of the data and the remaining three columns show the number of observations that where predicted to be false hits, non-hits and true hits respectively. Table 4.3 shows that 17.3% of the observations have been misclassified when using minimum covariance determination for robust estimation. This corresponds to 8.7% of true hits being misclassified and 14.2% of non-hits being misclassified. In addition, 31.9% of false hits were incorrectly classified as true hits and 9.6% of non-hits were misclassified as true hits. The results in Table 4.4 show that 17.6% of observations were incorrectly classified when using the minimum volume ellipsoid approach with 7.8% of true hits and 14.6% of non-hits being misclassified. It can also be seen that 31.9% of false hits and 10% of non-hits were misclassified as true hits. These results are compared further in the

following section with the results from the non-robust model and the results obtained during exploratory analysis.

Table 4.3: Classification using minimum covariance determination estimation

Observed Classification	**Predicted Classification**		
	False Hits	Non-Hits	True Hits
False Hits	90	17	50
Non-Hits	46	858	96
True Hits	5	4	94

Table 4.4: Classification using minimum volume ellipsoid estimation

Observed Classification	**Predicted Classification**		
	False Hits	Non-Hits	True Hits
False Hits	89	18	50
Non-Hits	46	854	100
True Hits	5	3	95

4.3.3 Comparison of Methodologies

To assess how using unlabelled data to update classification rules performs on the current data set the three models (non-robust, minimum volume ellipsoid and minimum covariance determination estimation) were compared with the results of classifying using linear discriminant analysis and random forests from Section 3.2.6. The results for this comparison are displayed in Table 4.5. The first two columns show the method of classification and the corresponding overall misclassification rate. The remaining three columns respectively show the percentage of misclassifications for true hits, false hits classified as true hits and non-hits classified as true hits.

Table 4.5: Comparing methodologies

Method	Overall Misclassification Rate	Percentage Misclassified		
		True Hits	False Hits As True Hits	Non Hits As True Hits
Update	12.30%	24.27%	21.02%	0.90%
MVE	17.62%	7.77%	31.85%	10.00%
MCD	17.30%	8.74%	31.85%	9.60%
LDA	9.52%	24.27%	13.38%	1.10%
Random Forest	6.59%	25.24%	11.52%	1.20%

Comparing the classification results in Tables 4.5 it can be seen that both classical linear discriminant analysis and the random forest classifier have lower overall misclassification rates than the three models that use unlabelled data to update the classification rules. However, the overall misclassification rate may not give the best indication of which classifier is best for high content screening data. As discussed in Section 3.2.5, the aim of the analysis for the supervised classification data is to reduce the number of false positives and false negatives. Hence, a more appropriate comparison may be for the percentage of true hits that were misclassified, the percentage of false hits that were predicted to be true hits and the percentage of non-hits that were predicted to be true hits.

A comparison of the three different models for updating using unlabelled data shows that the two robust methods predict less false negatives than the non-robust method. In other words, the two robust methods are correctly predicting more true hits. However, the non-robust method predicts less false positives than the two robust methods. Overall this shows that although the robust methods are predicting more true hits correctly they are also predicting more false hits and non-hits to be true hits. An overall comparison of the three different models for updating with linear discriminant analysis and random forests also shows that the two robust updating methods correctly classify the most true hits but both linear discriminant analysis and random forests predict less false positives.

The objective of this application section was to assess the potential of the methodologies described in Section 4.2. The results of these analyses have shown

that updating using unlabelled data (especially the robust models) has reduced the number of predicted false negatives when compared to linear discriminant analysis and random forests. However, these models also have the disadvantage of having increased numbers of false positives and are much computationally slower than linear discriminant analysis. As the results indicated that there is no clear 'best' method at this stage it was decided to continue the investigation of these methods by applying them to a larger set of data which has the same form as data from future high content screening experiments. A description of this larger data set and the analyses completed on it are given in Section 4.4.

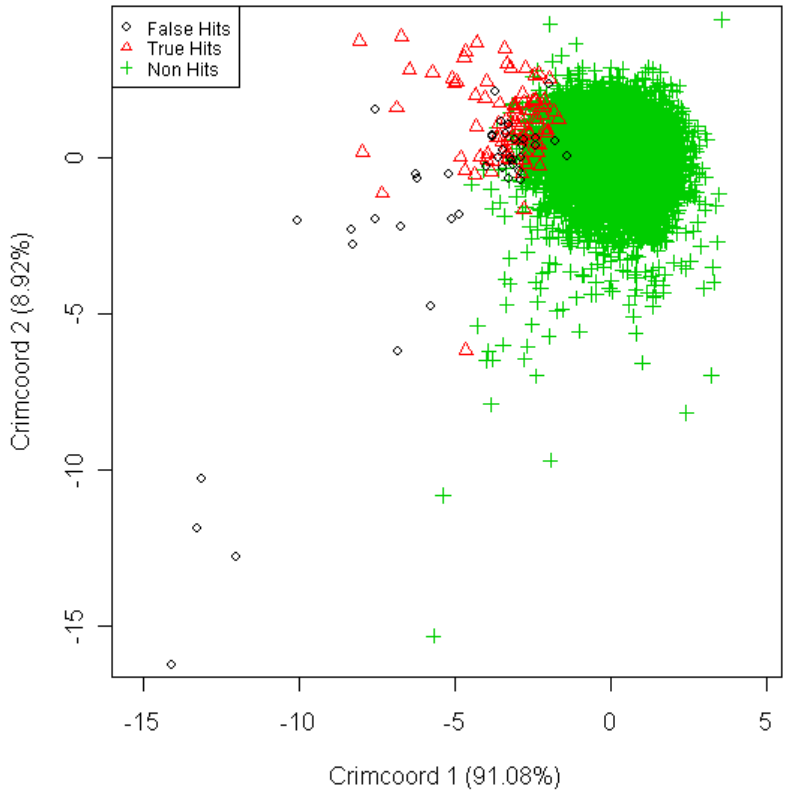
4.4 Application to Full Data

The aim of this Section is to apply the methodologies and extensions from 4.2 and 4.3 to the full set of supervised data that was described in Section 3.2 so that a comparison can be made with the existing one-parameter approach for classifying high content screening data. However, when attempting to apply the updating methodology computational memory problems were encountered. In order to solve these problems it was decided that the number of observations in each batch of data should be reduced whilst preserving the overall data structure.

The data was set up in the following way. The full training data consisting of 12,288 compounds was retained to train the models. The two test batches were then reduced in size from 33,941 and 33,408 compounds in batch A and B respectively to 1500 compounds in each. The compounds in each batch were selected by firstly taking the 81 compounds from batch A and the 39 compounds from batch B that were identified as true hits by the single parameter approach. The remaining 1419 and 1461 compounds for each batch were then selected randomly. By selecting the data for analysis in this way the objective changed to identifying if the updating methodology could correctly predict more or less true hits than the single parameter approach.

Applying the cross validatory method of model selection (described in Section 4.2.2) it was found that the model with equal volume, shape and orientation (model EEE in table and figure 4.1) had the smallest cross validation error. This model is the same as that selected for the analysis of the pre-screen data in Section 4.3 and is equivalent to classical linear discriminant analysis. The observations from the labelled training data are plotted on crimcoords in Figure 4.6.

Figure 4.6: Linear discriminant plot of training data



Examining the linear discriminant plot in Figure 4.6 it can be seen that the same features exist as with the linear discriminant plot in the pre-screen analysis (Figure 4.2). Firstly, the three groups are not clearly defined with there being a large overlap between all of the groups. In the case of the true hits and false hits the groups almost lie entirely on top of each other which suggests that large numbers of false positives and false negatives are expected. In addition, the problem of outlying data points is also a problem with all three groups having outlying

observations. The removal of these outliers using the two robust methods of estimating location and scale is shown in Figures 4.7 and 4.8.

Figure 4.7: Linear discriminant plot of training data with outliers removed by minimum covariance determinant method

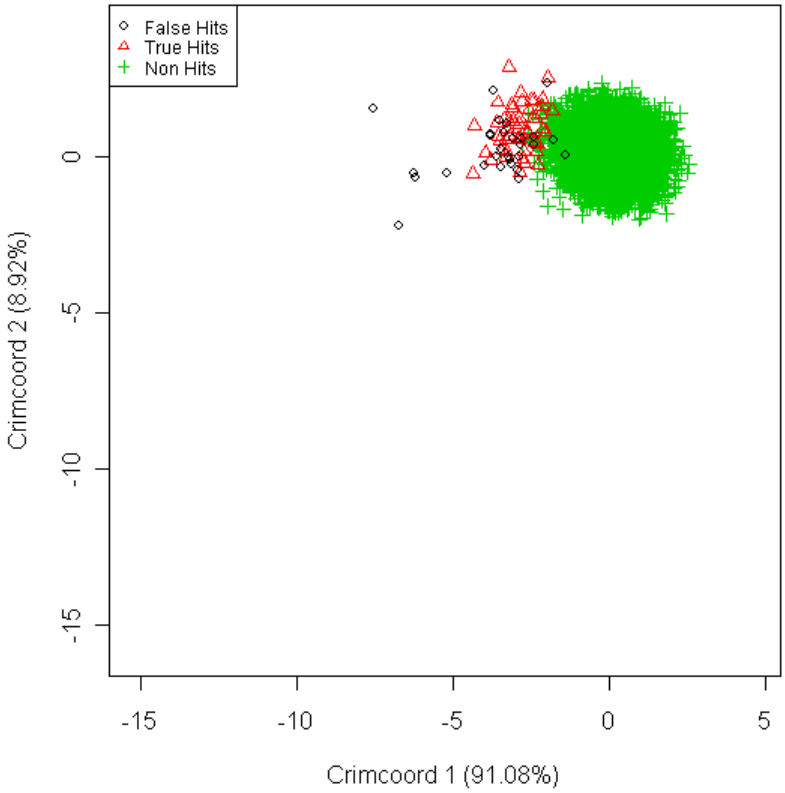
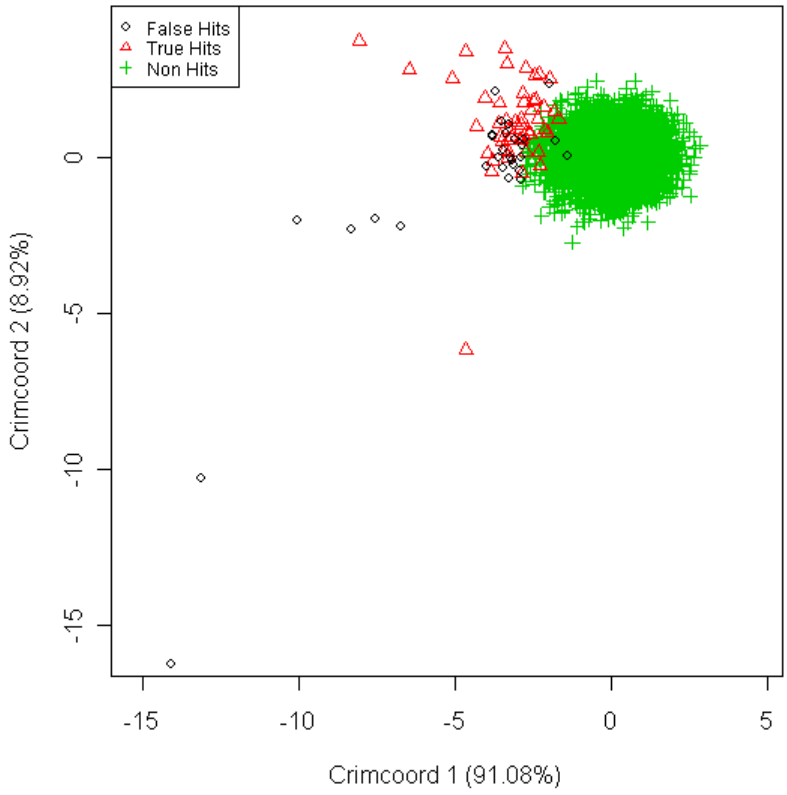


Figure 4.7 and 4.8 respectively show linear discriminant plots of the training data with outliers removed by the methods of minimum covariance determination and minimum volume ellipsoid estimation (see Section 4.3.1). A comparison of these plots with the plot of the full training data in Figure 4.6 shows that the groups look much more homogeneous after removing the outliers (especially in the case of minimum covariance determination). However, there is still a significant problem with the true hits and false hits overlapping (which was overcome in the previous analysis for the pre-screen data). Hence, the predictive performance of the classifier is not expected to improve when using the robust estimates.

Figure 4.8: Linear discriminant plot of training data with outliers removed by minimum volume ellipsoid method



4.4.1 Computational Problems

During the application of the methodology two computational problems were encountered. The first problem originated from the selection of starting values of the parameters in the CEM algorithm. During the previous applications of the algorithm the prior probabilities for each group were selected to be proportional to the number of observations in the group (calculated using the labelled training data) but when this approach was applied to the current data set it was found that the prior probability of an observation being a non-hit was much greater than that of being a true hit or false hit causing the algorithm to converge after one iteration with all of the unlabelled test data being classified as non-hits. As a solution to this problem the prior probabilities were changed so that they were equal for each group. This solution was considered adequate because the group probabilities are

recalculated at the start of each of the iterations and hence over a large number of iterations should have a negligible effect on the final classifications.

The second problem was concerned with the application of the minimum covariance determination method for calculating robust estimates of multivariate location and scale. When applying this methodology each iteration of the algorithm was taking approximately ten minutes on a Pentium 4 3GHz machine and therefore, to allow for one thousand iterations, the algorithm would take nearly seven days to complete. Hence, this methodology is not suitable for use in high content screening experiments and continued our investigation using the minimum volume ellipsoid method of estimation.

4.4.2 Comparison of Methodologies

One of the aims of the application section was to identify if the updating methodology predicts more or less true hits than the single parameter approach. Consequently, the results displayed in Tables 4.6 and 4.7 are the observed classifications (i.e. those made by the screening expert) of compounds that were predicted to be true hits by the single parameter approach, the non-robust updating model and the updating model with parameters estimated by the minimum volume ellipsoid method. In a change from the previous analyses the observed classifications of two of the groups have now been split into further sub-groups with the true hits being split into hits and good hits and the false hits being split into the nine groups shown in Table 3.2. This change in grouping reflects the change in assessment of classifier performance being used and allows a more in depth comparison to be carried out. This method of comparison shall be used in the remaining chapters of this thesis.

Table 4.6: Comparison of methodologies for batch A

Observed Classifications	Single Parameter Approach	Non-Robust Updating	MVE Updating
Hits	50	45	27
Good Hits	31	30	34
Non-Hits	0	2	1
Focus Error	1	0	0
High Background	4	2	4
Over Confluent	0	0	0

A comparison of the results of batch A in Table 4.6 shows that both the updating models fail to correctly predict as many hits as the single parameter approach. Further analysis revealed that thirty-one of the hits correctly predicted by the non-robust model were the same as those found by the single parameter method. This suggests that there is a minimum of sixty-four hits in this data set and highlights the number of false negatives associated with both methodologies. Further comparison shows that the robust model has correctly predicted three more good hits than the single parameter approach but this is not enough to justify its use over the single parameter or non-robust updating approaches given its performance in predicting hits. Note also that the ambiguity reject option of Section 4.3.2 was applied when using the robust updating model with forty-five observations being rejected (four of which were found to be good hits and one was found to be a hit).

A comparison of the results for batch B in Table 4.7 shows a contrast to those of batch A with both of the updating models correctly predicting more hits than the single parameter approach but less good hits. Further comparison of the false hits and non-hits shows that there is little difference in the numbers of false positives for each model. Note that in batch B ten observations were rejected using the ambiguity method and two of these observations were classified as hits by the screening expert. These results are discussed further in Section 4.5.

Table 4.7: Comparison of methodologies for batch B

Observed Classifications	**Single Parameter Approach**	**Non-Robust Updating**	**MVE Updating**
Hits	19	22	25
Good Hits	20	16	18
Non-Hits	0	3	3
Focus Error	1	1	1
High Background	1	0	0
Over Confluent	0	2	5

The comparison of the single parameter approach to the two methods of using unlabelled data to update classification rules has shown mixed results. For the first batch of compounds (batch A) the single parameter appears to perform the 'best' in terms of numbers of hits and good hits predicted correctly but for the second batch of compounds (batch B) the updating model with robust parameter estimation appears to outperform the single parameter approach. The data set used for these comparisons is much more akin to data that would be produced from a high content screening experiment than the data used in Section 4.3 and it is for this reason that this method of updating is not pursued further.

4.5 Summary and Discussion

The aims of this chapter were to firstly introduce the idea of using unlabelled data to update classification rules before applying the methodology to data from high content screening experiments. The analyses took two different forms with the first application being used on the pre-screen data as an initial assessment before applying the methodology to the full high content screening data. The results of both these analyses were mixed with the updating methods (both robust and non-robust) making an improvement over the single parameter approach, linear discriminant analysis and random forests for some criteria but failing to match the performance for other criteria. Overall there was not enough gain in predictive

performance from updating in this manner for it to be considered over the existing methodologies. In addition, the updating method is computationally much slower than when using the single parameter approach, linear discriminant analysis or random forests and therefore regardless of its performance may have not been considered suitable.

The ambiguity reject option from Section 4.3 was only applied to the data when the CEM algorithm would not converge. However, it is also possible to apply this method in a more general situation by defining criteria for when to reject. For example, it would be possible to reject all those observations whose largest posterior probability was not greater than 0.6. By using the ambiguity reject option in this way the classification results maybe improved because the uncertainty around decision boundaries would be reduced by using the screening expert's classifications. The disadvantage of this would be an increased number of images that were required to be visually classified. The idea of ambiguity rejection is considered further in the context of random forests in Chapter 8.

Throughout this chapter the methodologies that have been described and applied have all made the assumption that the training data is randomly sampled from the same distribution as the test data. In the case of the data used during the analysis in Section 4.3 this assumption is true because this is an artificial data set that was created by sampling from the pre-screen data. However, in the case of the data used in Section 4.4 this assumption is in fact false with the training data being selected for its known properties so that the experiment can be validated. It is data in this form that is representative of that from high content screening experiments.

There are many issues associated with training data not being representative of test data with the main problem being a reduction in the performance of the classifier being used. A particularly relevant problem to the work of this chapter is the estimation of parameters. McLachlan (1992) suggests that in the situation where classified data is not representative of an observed random sample from the

sample space of the feature vector appropriate steps must be taken, such as fitting truncated densities, or estimation of unknown parameters using only the data of known origin will be biased. In other words, by only using the unrepresentative training data to estimate the unknown parameters in Section 4.4 the estimates obtained were biased. However, this is not an issue investigated further.

The next chapter continues developing the idea of updating classifications rules with a new algorithm for classifying compounds in batches. This new algorithm is designed to take into account the training data not being representative of the test data and the possibility that the underlying populations change throughout the course of the experiment.

Chapter 5

Updating Algorithm

5.1 Introduction

Following the work in the previous chapter of exploring the use of unlabelled data to update classification rules, this chapter introduces a new classification method for batches of compounds where the rule is updated sequentially using information from the classification of previous batches. This continues the investigation of the classification of the high content screening data from Section 3.2.

The new updating methodology takes into account two problems that may cause a reduction in the performance of classical multivariate classifiers. Firstly, traditional multi-parametric methods of classification make the assumption that the data used to train the classifier are randomly sampled from the same distribution as the points to be classified in the future (Hand, 2006). In other words, any model that is based on an empirical fit to data collected under different circumstances to that currently being classified may not yield good predictions (Brentnall *et al.*, 2008). In the case of high content screening experiments the data in the training set are from compounds selected because of their known biological effects. Hence these data points may not be representative of the data to be classified in the future.

A second fundamental assumption of classical classification techniques is that the various distributions do not change over time (Hand, 2006) but in many situations this assumption fails to hold and may lead to a decline in performance of a classifier over time (for example, credit scoring models where there may be changes to distributions over time due to seasonal variation). With the application to high content screening data the class populations do not evolve over time but instead the changes are associated with compounds being analysed in batches. High content screens contain a finite number of compounds, so given the classifications of those compounds it would be possible to calculate the distribution of each class. However, as analysis is sequential these distributions are not known, therefore the distributions of the classes are based on the non-random sample of compounds in the training data. Each new batch of data brings with it compounds that may have different properties to those in the training data and those in previous batches so the class distributions in any model need to change to accommodate these new observations. In other words, the new data are from different distributions to those in the training data and changes to the distributions of the classes and hence the posterior distributions of class membership may be required sequentially so that classifier performance does not deteriorate (Kelly *et al.*, 1999).

The remainder of this chapter can be summarized as follows. Section 5.2 describes in detail the new updating methodology before Sections 5.3 and 5.4 respectively apply and compare the algorithm based on each of the four alternatives of a random forest, linear discriminant analysis, k-nearest neighbours and mixture discriminant analysis. Section 5.5 investigates the sensitivity of the batch orderings before the chapter concludes with a summary and discussion.

5.2 Updating Methodology

In this section the new method for updating classification rules is outlined, it is then used in the following section to classify data from a high content screening experiment. This is followed by a mathematical description of the algorithm in Section 5.2.1.

The methodology for the new updating algorithm is as follows. The training data are initially classified using the single parameter plus visual checking approach (described in Section 3.1) and a classifier is constructed using these data. This classification rule is then used to classify those compounds which were screened as part of the first batch (in our case batch A) into groups of true hits, non-hits and false hits.

The compounds identified as true hits by the classifier are examined visually by the screening expert to verify the predictions. At this stage all true hits that have been misclassified have their classification labels corrected. A new training data set is now created by combining the data from batch 1 (the original training data) with the visually checked compounds from batch A.

This new updated training data is used to construct a new classifier for the classification of Batch B. This part of the algorithm accommodates the possibility that the training data is not representative of the test data by correcting the assumptions on underlying distributions made from the training data.

For each new batch of data the training data is updated using the previous batches until the final classification rule that is constructed is the 'best' possible. At this stage it is recommended that each of the batches are classified again to see if any true hits were misclassified during the previous classification.

5.2.1 Algorithm

Let the pre-screen training data $L^{(0)}$ consist of data $\{(y_n,x_n),n=1,...,N\}$ where the y's are the class labels. Let the set of all data test data T consist of data $\{(y_m,x_m),m=1,...,M\}$ where the y's are the unknown class labels.

Step 1: (k = 0) Given the training set $L^{(0)}$, construct a classifier $\varphi_0(x^{(0)},L^{(0)})$, where given input x the class membership y is given by $\varphi(x,L)$.

Step 2: (k = 1) Classify batch k (denoted $t^{(k)}$) of the data using the classifier φ_{k-1} to give class labels $\varphi_{k-1}(t^{(k)},L^{(k-1)})$.

Step 3: Identify $x_i^{(k)} \in \{x_m^{(k)}\}$ such that $\varphi_{k-1}(x_i^{(k)},L^{(k-1)})$ = True Hit.

Step 4: Check the classification of the observations $x_i^{(k)}$ identified in step 3 and adjust any incorrect labels

Step 5: Construct a new training set $L^{(k)}$ consisting of data $\{(y_{n'}^{(k)},x_{n'}^{(k)}),n'=1,...,N'\}$ where $x_{n'}^{(k)}$ is the combined data $x_n^{(k-1)}$ and $x_i^{(k)}$.

Step 6: Construct a classifier φ_k using the data $L^{(k)}$.

Set k = k+1 and repeat steps 2-6 until k = B, the number of batches.

Step 7: Apply the classifier φ_B to batches 1,…,B to identify any true hits that have previously been misclassified.

5.3 Application to HCS Data Set

This section is concerned with applying the updating algorithm from Section 5.2 to the high content screening data set described in Section 3.2 using different classifiers. Four different classifiers were used for the application, random forests, linear discriminant analysis, k-nearest neighbours and mixture discriminant analysis. Random forests were chosen because this method produced the lowest misclassification rate during exploratory analysis (see Table 3.3 in Section 3.2.4) and there are comparisons in the literature (Svetnik *et al.*, 2003) which indicate it may generally have the highest classification accuracy. Linear discriminant analysis was applied because it is considered as a 'standard' method of discrimination and therefore is a good bench mark with which to compare other classifiers. K-nearest neighbour classification was used as it has received good reviews in terms of its performance in the statistical literature (for example Weiss and Kapouleas (1989) and Michie *et al*. (1994)). Finally, mixture discriminant analysis was applied for an extra comparison.

Due to the large number of compounds in each of the test batches of data (33,941 and 33,408 respectively) it was only possible to check the classification of those compounds that were predicted to be true hits (a full comparison would require all 67,349 compounds to be visually classified by the screening expert). This was considered to be sufficient because comparisons could be made between the number of true hits and false positives for each of the classifiers. In addition, by constructing a list of all true hits found during analysis the minimum number of false negatives for each method was identified.

It is also important when comparing the classification results to take into account the number of images that were required to be checked by the screening expert in order to achieve the final classification. This was considered to be important because of the time it takes for each image to be checked. For the updating algorithm this required counting the number of images to be checked at the end of each iteration in addition to those when applying the final model. For the non-

updating method the number of compounds selected as hits for both batches were counted. The methods were then compared by looking at the ratio of hits found to images checked. These comparisons are shown when evaluating all methods in Section 5.4.5.

5.3.1 Random Forests

This section applies the random forest classifier to the data set both as part of and independent of the updating algorithm. All of the analyses were carried out using the randomForest package in R (Liaw and Wiener, 2002). The random forest methodology is outlined in Section 2.3.3.

Table 5.1 compares the results of classifying the two batches of test data (A and B) using a random forest classifier as part of and independently of the updating algorithm. The first column of this table contains a list of all possible classifications and the second and third columns show the observed classifications (for batches A and B respectively) of those compounds that were predicted to be hits by the random forest with no updating. When applying the updating algorithm, each iteration (with the exception of the first) uses a different model for prediction than when there are no updates, therefore the fourth column shows the results of classifying batch B after the model has been updated using the addition information from batch A. The fifth and sixth columns contain the observed classifications (for batches A and B respectively) of those compounds that were predicted to be hits by the final updating model. In other words, these are the results of applying the random forest model that has been updated using all batches of data.

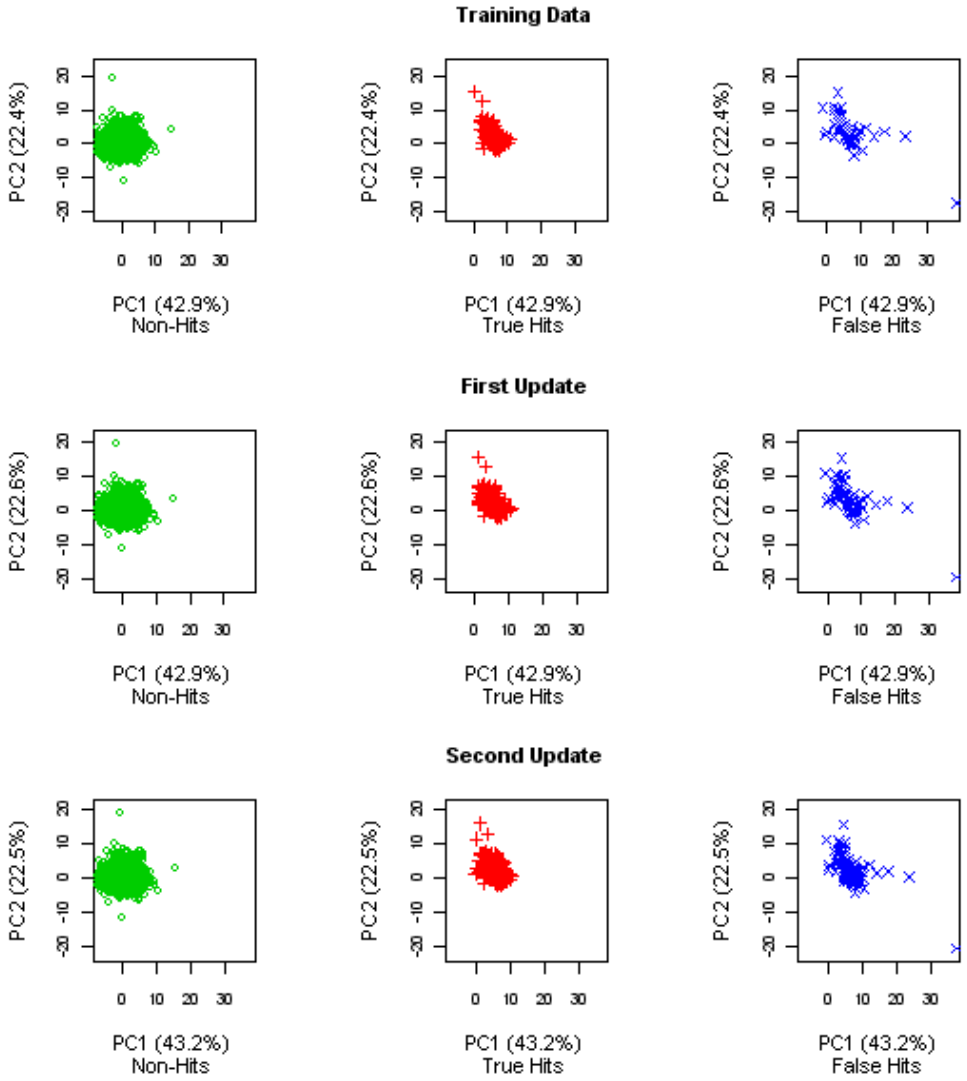
Table 5.1: Comparing updating with no updating using a random forest classifier

Observed Classifications	No Updating		Iteration	Final Model (Updating)	
	A	B	B	A	B
Hits	75	21	22	75	22
Good Hits	30	19	19	29	20
Non-Hits	15	35	31	0	1
Focus Error	5	2	2	0	0
Over Confluent	6	39	34	0	0
Toxic	7	18	1	0	0
High Background	0	6	1	0	0
No Image	0	0	1	0	0
Well Dry	0	1	0	0	0

Comparing the results of classifying without updating with those of the final model of the updating algorithm (i.e. comparing the second and third columns of Table 5.1 with the fifth and sixth columns) it can be seen that there is a noticeable reduction in the number of compounds that are classified incorrectly when using the updating algorithm. In particular, the final model of the updating algorithm has not misclassified any non-hits or false hits as true hits for batch A and has only misclassified one non-hit as a true hit for batch B; whereas classification without updating has made thirty-three mistakes when predicting true hits in batch A and one hundred and one mistakes when predicting true hits in batch B. However, in order to compare the two methods properly it is important to consider the misclassifications that were made during the iterative stages of the algorithm. Comparing the results of classifying batch B at the iterative stage of the updating with those using the no update method (columns four and three respectively in Table 5.1) it can be seen that the updating method reduced the number of non-hits, over confluent, toxic, well dry and high background compounds that were misclassified but did misclassify one compound with no image that was not selected by the other method.

Overall, the results suggest that the predictive capability of the random forest classifier is improved when using the updating algorithm. These results are compared further and also with the results of using other classifiers in Section 5.3.5.

Figure 5.1: Principal component plots for updating training data using random forests



In order to see how the distributions of the three groups (true hits, false hits and non-hits) changed as the training data was updated the three iterations of the data were plotted on principal components. The first row of principal component plots in Figure 5.1 show the three groups for the original training data. Each group is plotted on separate axes because all the groups overlap in the principal component space making it difficult to see any distributional changes. The second and third rows of plots show the three groups for the first and second update of the training data respectively. Each update of the training data is plotted on different principal components calculated using the relevant data but as only a relatively small

amount of new data is added at each update it is not expected that there will be much change in the axes.

Comparing the principal component plots in Figure 5.1 it is possible to see very subtle changes to the shapes of the groups for the true hits and false hits. These results suggest that principal component axes may not be the best method of visualising the data in order to see changes in the distributions of the groups. Further inspection of higher principal components (not shown here) did not reveal any improvement and neither did the principal component plots for updates using the remaining classifiers that are discussed in this chapter so with the exception of linear discriminant analysis (where principal components are compared to crimcoords) these plots are not shown.

5.3.2 Linear Discriminant Analysis

This section applies linear discriminant analysis to the data set both with and without the updating algorithm. In addition to discussing the results of classification, the training data is examined to see how it changed as it was updated. This was done by plotting the data on principal components and crimcoords.

Table 5.2: Comparing updating with no updating using linear discriminant analysis

Observed Classifications	No Updating		Iteration	Final Model (Updating)	
	A	B	B	A	B
Hits	55	18	18	39	18
Good Hits	31	19	19	30	15
Non-Hits	16	36	50	16	32
Focus Error	3	0	2	2	2
Over Confluent	5	9	46	1	19
Toxic	2	0	0	0	0
High Background	2	0	4	3	4
No Image	0	0	0	0	0
Well Dry	0	0	0	0	0

Table 5.2 compares the results of classifying the two batches of test data (A and B) using linear discriminant analysis as part of and independently of the updating algorithm. The first column of this table contains a list of all possible classifications. The observed classifications of those compounds that were predicted to be hits by linear discriminant analysis with no-updating for batches A and B are shown in the second and third columns respectively. When applying the updating algorithm each iteration (with the exception of the first) uses a different model for prediction than when there are no updates, therefore the fourth column shows the results of classifying batch B after the model has been updated using the addition information from batch A. The observed classifications (for batches A and B) of those compounds that were predicted to be hits by the final updating model are shown in the fifth and sixth columns respectively. In other words, these are the results of applying linear discriminant analysis that has been constructed using training data that has been updated using all batches of test data.

Comparing the results for batch A of classifying without updating with those of the final linear discriminant rule of the updating algorithm (i.e. comparing the second column of Table 5.2 with the fifth) it can be seen that the updating algorithm has reduced the number of compounds that are over confluent or have focus error from being classified as hits. However, this small reduction in false hits is at the expense of a reduction in the number of hits and good hits found, with the updating method finding sixteen less hits and one less good hit. Similar results are found when comparing the results of classifying batch B at the iterative stage of updating with those using the no update method and those using the final linear discriminant rule of the updating algorithm (columns four, three and six of Table 5.2 respectively). Here it can be seen that the same number of hits are found at each stage but less good hits are found by the final model than the no updating method or the iteration. In addition, there is a large increase in the number of non-hits and over confluent compounds classified as hits by the iteration when compared to the no update but these numbers are reduced when applying the final model.

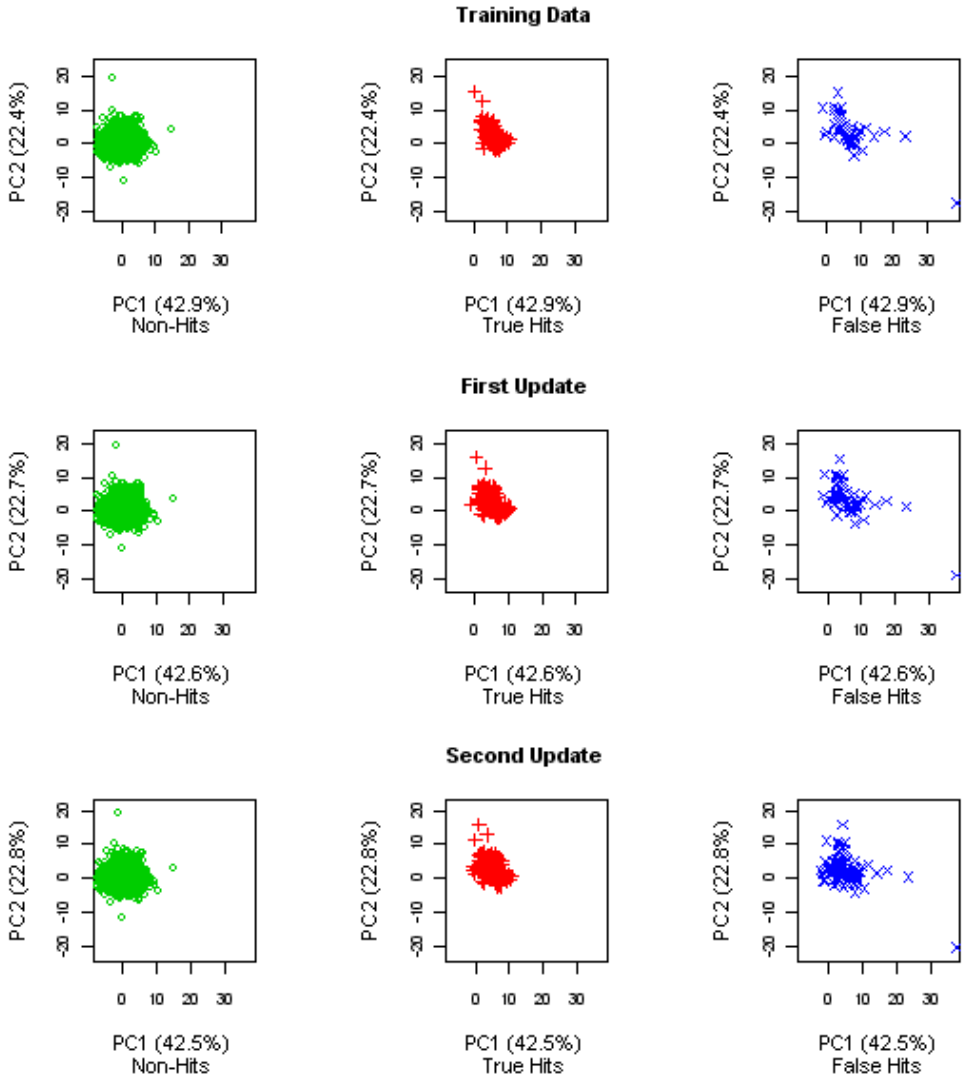
Overall, the results in Table 5.2 suggest that the final updated model is not an improvement on the classical linear discriminant model. This reduction in predictive performance maybe due to linear discriminant analysis being susceptible to the influence of the additional observations in the updated training data. In other words, linear discriminant analysis is less robust to outliers and other influencing observations than other classifiers (for example, the random forest applied in the previous section). The influence of observations is discussed in Campbell (1978) but is not pursued further here. These results are compared further and with the results of using other classifiers in Section 5.3.5.

The changes to the distributions of the three groups (true hits, false-hits and non-hits) as the training data was updated were investigated by plotting the data at each stage on principal components and on crimcoords. The first row of plots in Figures 5.2 and 5.3 show the original training data plotted on principal components and crimcoords respectively. Each group is presented on its own plot so that it is easier to identify any changes in the data clusters (when the data is shown on one plot some of the groups overlap). The second and third rows of both figures show the respective plots for the first and second updates of the training data. For both the plots of principal components and crimcoords each update of the training data is plotted on axes calculated using the relevant data. This means that there will be slight differences in axes but as only a relatively small amount of new data is added at each update it is not expected that there will be much change (and this is reflected in the plots).

The principal component plots in Figure 5.2 show that there are some subtle changes to the clusters as the updates take place. In particular, there is a much higher density of points around (0, 0) on the plot of false hits after the second update than there is for the original training data. However, as discussed in Section 5.3.1, these results suggest that principal component axes may not be the best method of visualising the data in order to see changes in the distributions of the groups. In contrast, the plots of crimcoords in Figure 5.3 give a much better indication of how the distributions of the true hit and false hit groups are changing

as the training data is updated. The true hit group has expanded so that its boundaries are at 0 and -12 on crimcoord 1 for the second update of the training data whereas the boundaries on crimcoord 1 for the original training data were at 1 and -8. In other words, the within group variability of the observations has increased and the group centroid has moved.

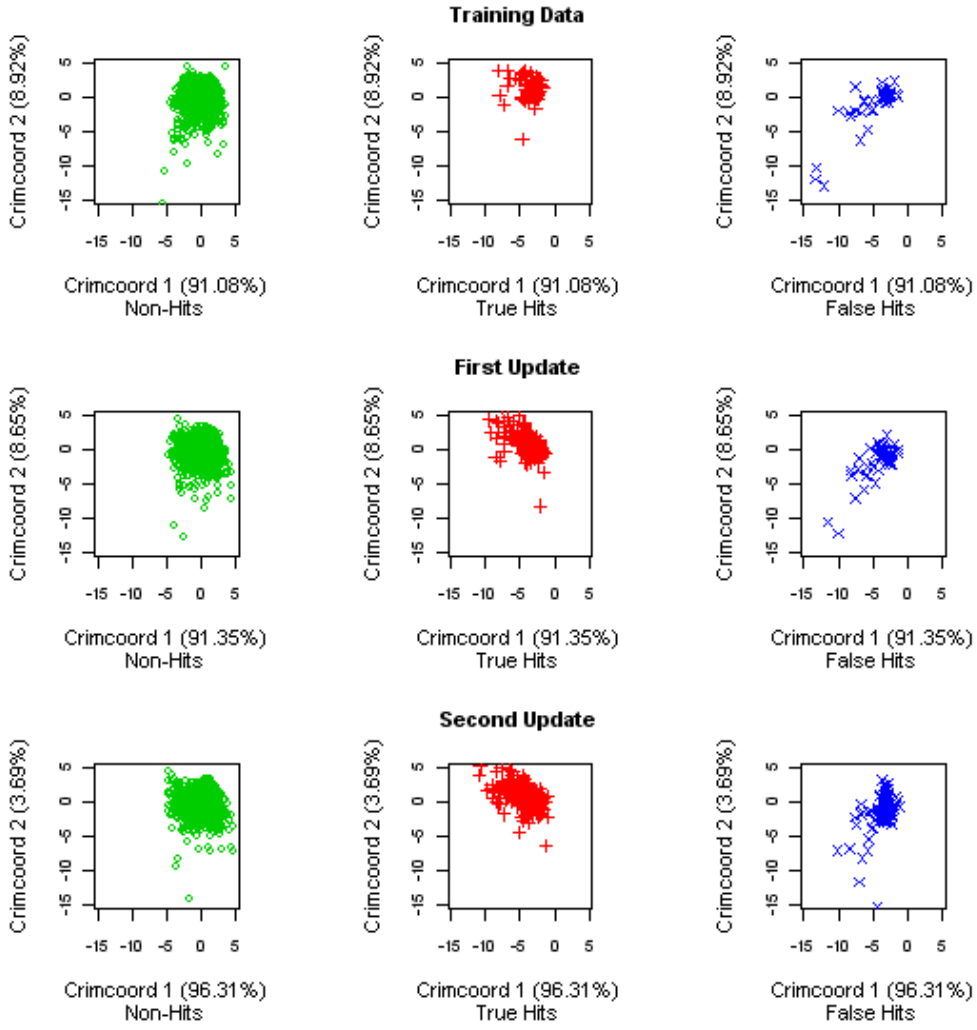
Figure 5.2: Principal component plots for updating training data using linear discriminant analysis



Further to the discussion earlier in this section about the results in Table 5.2 not suggesting that the updating algorithm is improving the predictive performance of

linear discriminant analysis, the plots in Figure 5.3 show that there is no clear distinction of the groups in linear discriminant space for any of the iterations of the training data. In other words, for the original training data and for the two updated training sets the three groups are overlapping each other and therefore the linear discriminants will make classification errors. This is not improved by the updating procedure because changes are not made at the overlapping boundaries of the groups to make them distinct.

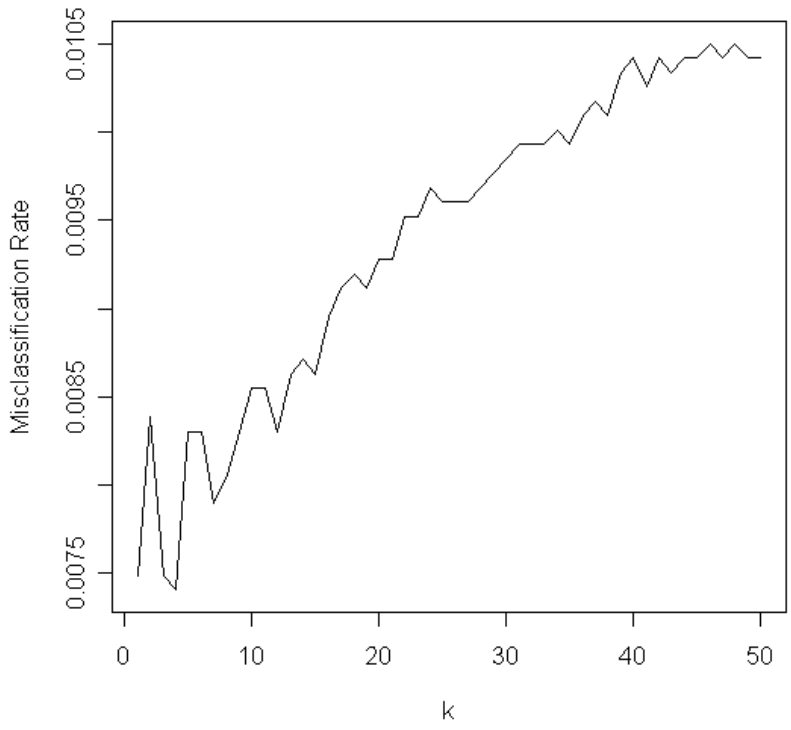
Figure 5.3: Linear discriminant plots for updating training data



5.3.3 K-Nearest Neighbours

This section involves the application of the k-nearest neighbour algorithm both as part of and independent of the updating algorithm. A comparison of the results of applying both methods is presented but first a description of the method for selecting the value of k (the number of neighbours) is given. Full details of the k-nearest neighbour algorithm can be found in Section 2.3.4.

Figure 5.4: Associated misclassification rates for numbers of nearest neighbours



When implementing the k-nearest neighbour algorithm the knn.cv function in R was used to apply a leave-one-out cross-validatory approach for selecting the value of the parameter k. For each observation in the training data, the k nearest (in Euclidean distance[†]) other training observations are found, the classification is then decided by majority vote with any ties broken at random. In the case of there being ties for the k^{th} nearest observation, all candidates are included in the vote

[†] Note that other distance metrics can be used (see Section 2.3.4).

(Molina *et al.*, 1994 and Venables and Ripley, 2002). Figure 5.4 is a plot of k (the number of nearest neighbours) against the leave-one-out cross-validated misclassification rate. In this case the expected optimum value for k would be small because the principal component plots in Figures 3.6, 5.1 and 5.2 show the three groups in the data to be occupying the same space. In other words, choosing a large value for k would increase the chance of selecting data points from the wrong group. This is reflected in Figure 5.4 where the misclassification rate is minimised when k equals four and the misclassification rate tends to increase as the value of k increases. The above procedure for selecting a value of k was repeated each time the training data was updated (not shown here).

Table 5.3: Comparing updating with no updating using a k-nearest neighbour classifier

Observed Classifications	No Updating		Iteration	Final Model (Updating)	
	A	B	B	A	B
Hits	32	13	12	36	17
Good Hits	24	11	18	27	16
Non-Hits	4	16	22	7	11
Focus Error	19	4	3	5	2
Over Confluent	5	1	26	1	5
Toxic	5	25	3	0	2
High Background	1	24	1	1	0
No Image	0	1	0	0	0
Well Dry	0	0	0	0	0

Table 5.3 compares the results of classifying the two batches of test data (A and B) using a k-nearest neighbour classifier as part of and independently of the updating algorithm. A list of all possible classifications is contained in the first column of this table and the observed classifications of those compounds that were predicted to be hits by k-nearest neighbour with no-updating for batches A and B are shown in the second and third columns respectively. When applying the updating algorithm each iteration (with the exception of the first) uses a different model for prediction than when there are no updates, therefore the fourth column shows the results of classifying batch B after the model has been updated

using the addition information from batch A. The fifth and sixth columns contain the observed classifications (for batches A and B respectively) of those compounds that were predicted to be hits by the final updating model. In other words, these are the results of applying the k-nearest neighbour classifier that has been updated using all batches of data.

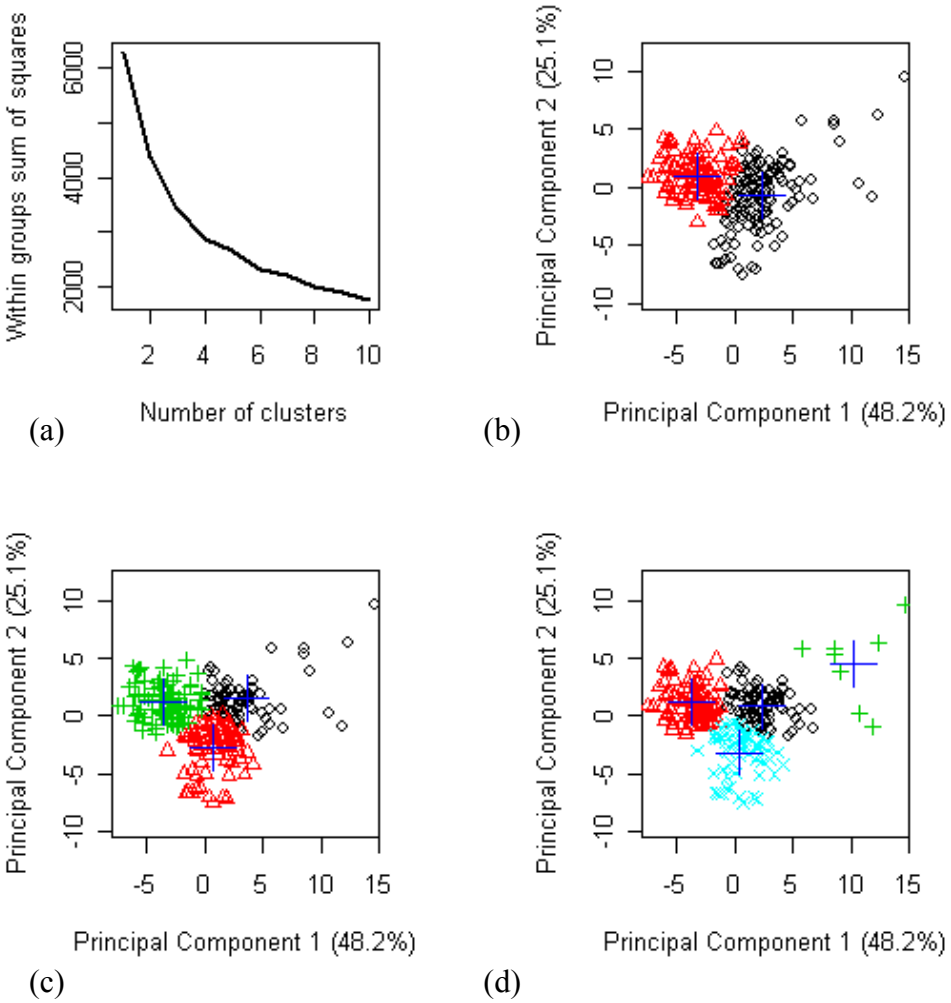
Comparing the results of classifying using k-nearest neighbours without updating with those of the final model of the updating algorithm (i.e. comparing the second and third columns of Table 5.3 with the fifth and sixth columns) it can be seen that for both batches the number of hits and good hits identified has been increased by using the updating methodology. The most noticeable difference between the two methods is seen when comparing the results for batch B with the final model classifying twenty-three fewer toxic compounds and twenty-four fewer compounds with high background as hits than the no updating method. Further reductions in the number of focus errors, over confluent compounds and toxic compounds are seen in Batch A but there is a slight increase in the number of non-hits.

Comparing the results of classifying batch B at the iterative stage of the updating algorithm with those using the no update method (columns four and three respectively in Table 5.3) it can be seen that there are an extra seven good hits selected by updating and a reduction in the number of toxic and high background compounds selected as hits (twenty-two and twenty-three less respectively). However, these improvements are at the cost of an extra six non-hits and twenty-five over confluent compounds being incorrectly classified.

5.3.4 Mixture Discriminant Analysis

In this section mixture discriminant analysis is applied to the data set both as part of and independent of the updating algorithm. A comparison of the results of applying both methods is presented but first a description of the method for selecting the number of clusters for each group is given. See Section 2.3.2 of Chapter 2 for details of the mixture discriminant analysis methodology.

Figure 5.5: Selecting cluster numbers for mixture discriminant analysis: (a) plot of number of clusters against within-group sum of squares; (b) principal component plot with two clusters; (c) principal component plot with three clusters; and (d) principal component plot with four clusters.



The mixture discriminant analysis was carried out using the mda package (Hastie *et al.*, 2006) within R. This function requires the number of clusters for each group to be chosen. In order to determine this, the informal method of examining the value of the within-group sum of squares associated with solutions for a range of different group sizes was used. As the number of groups increases the within-group sum of squares will decrease but a large change in the level of the plot may be indicative of the best solution (Everitt *et al*., 2001 and Everitt, 2005). This is exemplified by the top left-hand plot of Figure 5.5, where the number of clusters for the true hits class is being investigated. This plot of the number of clusters against within-group sum of squares shows that the main change in level occurs between 2 and 4 clusters. The decision is further aided by the remaining three plots of Figure 5.5 which show the first two principal components for the corresponding data with the top left-hand plot showing 2 clusters (as determined by a k-means clustering algorithm) for the data and the bottom left and right-hand plots showing 3 and 4 clusters respectively. The large blue crosses on each plot indicate where the mean is located for each cluster. Using this information it was decided that 4 clusters would be the most appropriate because this treats the small outlying group to the right of principal component one as a separate cluster. For alternative methods for choosing numbers of clusters see Everitt *et al*. (2001).

Table 5.4: Comparing updating with no updating using mixture discriminant analysis

Observed Classifications	**No Updating**		**Iteration**	**Final Model (Updating)**	
	A	**B**	**B**	**A**	**B**
Hits	85	15	22	64	20
Good Hits	15	16	19	31	18
Non-Hits	36	27	36	20	34
Focus Error	27	5	9	21	9
Over Confluent	27	25	30	20	28
Toxic	39	45	1	1	0
High Background	5	4	3	3	0
No Image	0	0	0	3	6
Well Dry	0	0	0	0	1

Table 5.4 compares the results of classifying the two batches of test data (A and B) using mixture discriminant analysis as part of and independently of the updating algorithm. The first column of this table contains a list of all possible classifications and the observed classifications of those compounds that were predicted to be hits by mixture discriminant analysis with no-updating for batches A and B are shown in the second and third columns respectively. When applying the updating algorithm each iteration (with the exception of the first) uses a different model for prediction than when there are no updates, therefore the fourth column shows the results of classifying batch B after the model has been updated using the additional information from batch A. The observed classifications (for batches A and B) of those compounds that were predicted to be hits by the final updating model are shown in the fifth and sixth columns respectively. In other words, these are the results of applying mixture discriminant analysis that has been constructed using training data that has been updated using all batches of test data.

Comparing the results of classifying batch A without updating with those of using the final mixture discriminant model of the updating algorithm (i.e. comparing the second column of Table 5.4 with the fifth) it can be seen that by updating the rule there is a reduction in the number of false hits and non-hits classified as hits. 57% of compounds that were predicted to be hits by the none updating model were found to be misclassifications whereas 42% of compounds predicted to be hits by the final updating model were misclassifications. In addition, the final updating model predicted an extra sixteen good hits but this was at the expense of classifying twenty-one fewer hits than the non-updating model.

A comparison of the results of classifying batch B at the iterative stage of updating with those using the non-updating model and those using the final mixture discriminant model (fourth, third and sixth columns of Table 5.4 respectively) shows that both at the iterative stage and when using the final model there is an increase in the number of hits and good hits found and a large decrease in the number of toxic compounds classified as hits. However, there are small

increases in the numbers of non-hits, over confluent compounds and compounds with focus error that were classified as hits.

5.4 Overall Comparison of Classification Methods

The next stage of the analysis was to collate all of the results from sections 5.3.1 to 5.3.4 so that a comparison of the four different multivariate classifiers (both with and without using the updating algorithm) could be made with the current single parameter approach. This section presents the subsequent findings.

Table 5.5 compares the results of classifying the two batches of test data (A and B) using the single parameter approach (see Section 3.2.3 of Chapter 3 for details) and the four classifiers (random forests, linear discriminant analysis, k-nearest neighbours and mixture discriminant analysis) from Sections 5.4.1 to 5.4.4. With the exception of the single parameter approach the classifiers are applied using the classical non-updating methods and the updating algorithm. The results of applying the updating algorithm differ from those in the previous four sections because the results of the iterative stages of classification have been combined with classifications of the final models (this corresponds to step 7 of the algorithm). The first column of the table contains a list of all possible observed classifications. The remaining columns show the observed classifications of those compounds that were predicted to be hits by the nine different methods. Note that the results of updating displayed in Table 5.5 are the final results of applying the algorithm and do not represent the number of images that were required to be checked to achieve the final classifications. These results are compared later in Table 5.6.

Table 5.5: Comparing the single parameter classifier with multi-parameter classifiers

		Single Parameter Approach		Random Forest		Random Forest (Updated)		LDA		LDA (Updated)		KNN		KNN (Updated)		MDA		MDA (Updated)	
		A	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Observed Classification	Hits	50	19	75	21	75	22	55	18	58	19	32	13	43	17	85	15	101	23
	Good Hits	31	20	30	19	30	20	31	19	34	19	24	11	27	18	15	16	36	20
	Non-Hits	0	0	15	35	0	1	16	36	7	2	4	16	5	2	36	27	11	7
	Focus Error	21	10	5	2	0	0	3	0	1	1	19	4	3	0	27	5	6	3
	Over Confluent	2	3	6	39	0	0	5	9	0	1	5	1	0	0	27	25	10	7
	Toxic	3	7	7	18	0	0	2	0	1	0	5	25	0	1	39	45	1	0
	High Background	12	7	0	6	0	0	2	0	2	0	1	24	1	0	5	4	2	0
	No Image	3	7	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3	6
	Well Dry	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	Low Draq	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Comparing the results in Table 5.5 the most noticeable difference between the classifiers is the number of extra hits and good hits that have been found when using mixture discriminant analysis. The updated version of this classifier finds one hundred and one hits in batch A which is sixteen more than the non-updated version and twenty-six more than the closest alternative (random forests). In general, it can be seen that with the exception of the two versions of k-nearest neighbours all the multi-parameter classifiers find more hits for batch A than the single parameter approach and the two versions of random forests and updated mixture discriminant analysis find more in batch B. However, in the case of finding good hits, only the updated versions of linear discriminant analysis and mixture discriminant analysis perform better than the single parameter approach for batch A and, updated random forests and updated mixture discriminant analysis find the same number for batch B.

Focusing on the comparison of the number of non-hits and false hits predicted to be hits by the different classifiers shows that mixture discriminant analysis makes the most mistakes with two hundred and forty misclassifications for the two test batches combined. The best performance in terms of accuracy was the updated random forest which only made one mistake in final classification. However, it is important to remember that the updating results in Table 5.5 are obtained from applying the full updating algorithm and do not show the mistakes made during each of the iterations (these can be seen in Tables 5.1 to 5.4). A further comparison of misclassifications is made later in Table 5.6 when the ratio of hits found to images checked by the expert is examined.

A further important criterion in comparing the classifiers is the number of false negatives misclassified by each model; in other words, the number of hits that were classified incorrectly as non-hits or false hits. As described at the beginning of this application section, the large number of compounds in each test batch means that it was not possible to check the true classifications of the false hit and non-hit groups but by compiling a list of all compounds found to be hits by the different classifiers an idea of how they are performing can be gained.

Completing this analysis shows that in batch A a total of one hundred and twenty hits and fifty-one good hits were found and in batch B, thirty-five hits and twenty-one good hits were found. These results show that for updated mixture discriminant analysis (which identified the most hits) there were a minimum of nineteen false negatives (i.e. nineteen hits classified as non-hits or false hits that were in fact hits).

Table 5.6 compares the number of hits and good hits found by each classifier to the number of images that were required to be checked in order to achieve the classification. The first column is a list of all classifiers used and, the second, third and fourth columns show the number of hits found, good hits found and images checked respectively. The final column shows the percentage of images checked that were found to be hits for each method. The aim of this analysis is to look at the time and effort required by the expert to check images with respect to the number of good compounds found.

Table 5.6: Comparing hits found to number of images checked for different classifiers

Classifier	**Hits Found**	**Good Hits Found**	**Images Checked**	**% Images Checked That Were Found to be Hits**
Single Parameter	69	51	202	59 %
Random Forest	96	49	272	53 %
Random Forest (Updated)	97	50	250	59 %
LDA	73	50	196	63 %
LDA (Updated)	77	35	256	44 %
KNN	45	35	185	43 %
KNN (Updated)	60	35	196	49 %
MDA	100	31	371	35 %
MDA (Updated)	124	56	450	40 %

The updated mixture discriminant analysis identifies the most hits and good hits of all the classifiers but as the results in Table 5.6 show, this method also requires the largest number of images to be checked in order to achieve this classification

(only 40% of images checked turned out to be hits). Conversely, linear discriminant analysis has the largest percentage of images checked that turn out to be hits (63%) but it identifies fifty-one less hits and six less good hits than mixture discriminant analysis.

It is clear that the optimum classifier would find a large number of hits and these hits would correspond to a high percentage of the images checked. The evidence in Table 5.6 suggests that the updated random forest is the optimum classifier. Using this method the joint second (with the single parameter approach) highest percentage of images checked that turn out to be hits (59%) is found but the method also identifies twenty-eight more hits than the single parameter. However, the single parameter does identify one extra good hit than the updated random forest.

5.5 Sensitivity of Batch Orderings

The previous sections of this chapter have concentrated on describing the new methodology and applying it based on a number of different classifiers. This section will concentrate on a further aspect of evaluating the methodology by investigating how sensitive the algorithm is to the ordering of the batches. In other words, does the algorithm produce the same classification results regardless of the order of the batches? During this investigation, the updating algorithm was applied with the random forest classifier.

If the exact form of the final models as determined by the updating algorithm could be written down then the results produced by different batch orderings could be compared. However, as a random forest classifier is being used, this is not possible. Therefore, to empirically investigate the sensitivity of the batch orderings a comparison was made between the results of classifying the compounds as batches were permuted. In particular, each of the two batches of test data were randomly divided into two sub-batches (A1, A2, B1, B2); a random

order was then assigned to the sub-batches and the updating methodology was used to predict the class of the compounds. This process was repeated 8 times so that the predictions made by the model for each sub-batch order could be compared.

Table 5.7 shows the results of investigating the sensitivity of the classification results when permuting the batch orders. The first column lists the eight different permutations of batches used for classifications and the remaining seven columns show the true classifications of all those compounds identified as hits. These results suggest that there is some variation in the predicted classifications for the different batch orderings. The most noticeable difference appears to be between those orderings which start with batch A and those which start with batch B. A detailed comparison of this difference shows that when a sub-batch from batch A is the first to be classified there are more hits identified than when a sub-batch from batch B is the first to be classified. However, when comparing the numbers of compounds that were found to have been misclassified (i.e. the non-hits and the false hits) it can be seen that there is little difference between the results for all eight permutations.

Table 5.7: Observed classifications of hit selected compounds using updated random forests with different batch orders

Batch Order	Hit	Good Hit	Non-Hit	Focus Error	High Background	Over Confluent	Toxic
A1, B1, A2, B2	95	52	1	2	0	3	1
A1, B2, A2, B1	96	50	1	1	0	0	0
A2, B1, B2, A1	94	49	3	1	0	2	0
A2, B2, A1, B1	91	50	2	0	0	0	0
B1, A2, B2, A1	83	48	2	0	0	0	0
B1, B2, A1, A2	84	49	2	1	3	0	0
B2, A1, A2, B1	87	49	3	0	0	0	0
B2, A1, B1, A2	83	49	2	0	1	0	0

In conclusion, the results of Table 5.7 show that the results of classification vary as the batch ordering is permuted but it is suggested that in this case the updating

algorithm approximately converges to the same classification. However, it may be of interest to investigate this further by considering larger numbers of batches (this has not been done here because the process of classifying the data in different batch orders is very time consuming for the imaging expert). Further discussion of this is given in Section 5.6.

5.6 Summary

The aim of this chapter was two-fold. The first was to introduce the methodology for a new classification method for batches of compounds where the rule is updated sequentially using information from the classification of previous batches. The aim of this new algorithm was to take into account a reduction in classifier performance due to the training data not being representative of the test data and that the distributions of the groups in the data change as new batches of compounds are introduced. The second aim of the chapter was to apply this new methodology to our example data set using a number of different classifiers so that the results could be compared with standard methods of classification and the current single parameter methodology from high content screening data analysis.

The updating algorithm has been shown to improve the predictive capability of the four classifiers it has been tested on. In particular, all of the updated classifiers have identified more hits than their equivalent non-updated classifiers and the single parameter approach. In addition, by updating random forests, k-nearest neighbours and mixture discriminant analysis it has been shown that they are more efficient in terms of the number of images that are required to be checked by the screening expert per hit found than when not updating.

Further analysis concentrated on showing how the distributions of the groups in the training data changed as new batches of data were introduced and the training data was updated. From the results of this analysis it was concluded that principal component analysis is not useful in visualising data for this purpose but by

plotting the data on crimcoords calculated during linear discriminant analysis it is possible to show how the shape of the true hit and false hit clusters change as the data is updated; hence, indicating an increase in within group variability and a change in the position of the centroid of the groups.

The results of investigating the sensitivity of the classification results when changing the batch orderings showed that there is some variation when permuting the batches but it is suggested that in this case the updating algorithm approximately converges to the same classification. However, without further analysis it is not possible to comment generally on how sensitive the algorithm is to changes of batch orderings. It is possible to optimise the orderings of the batches or the compounds within the batches before they are classified by the updating algorithm using methods such as Willett (2006). However, any increase in accuracy from doing this would have to be balanced against the time taken to implement it.

In addition to looking at the sensitivity of classifications in future high content screening data sets, it may also be of interest to investigate this area further by considering larger numbers of batches (the batches used in our analysis only make up a small proportion of a full high content screening experiment). Further analysis could also focus on finding an optimal batch size (i.e. finding the most appropriate number of compounds to be in each batch). This could vary from only classifying a single compound at a time (this would not be practical in the case of high content screening experiments because of the very large number of compounds to be classified but may be appropriate in other contexts) to classifying batches of many thousands of compounds.

The objectives of the compound hit selection data set out in Section 3.2.5 identified that any multi-parametric approach should aim to identify more hits than the single parameter approach while reducing the number of false positives and false negatives. The results of the analyses in this chapter have shown that when applying the updating algorithm with the random forest, k nearest neighbour

and mixture discriminant classifiers these objectives are fulfilled. However, the study of the new classification algorithm has so far focused on the application of a number of classifiers to one high content screening experiment data set. In the next chapter a new high content screening case study is introduced so that the application of the updating algorithm can be investigated further. This work will focus on using the random forest classifier as this was found to perform 'best' during comparisons made in Section 5.4.

Chapter 6

Second Case Study

6.1 Introduction

This chapter continues the study of the new classification updating algorithm that was introduced in Chapter 5. The methodology is applied to a new high content screening case study with a different biological assay to that described in Chapter 3 as an illustration of its application in a different context. All multi-parametric classifications in this chapter are made using a random forest as this was found to perform 'best' in the previous chapter and comparisons are made with the single parameter approach.

The data to be analysed in this chapter comes in two forms. The first uses the same imaging algorithms as were used to produce the variables for the data described in Chapter 3. The second uses some new imaging algorithms which are supposed to be more accurate in their measurements of the biological features. Therefore, a further aim of this chapter is to make comparisons of the results produced using these different methods of producing variables.

The outline of the remainder of this chapter is as follows. Section 6.2 gives details of the data and a brief discussion of the features that are essential to the analysis. Section 6.3 applies the random forest and updated random forest

methods of classification to the two sets of data. Comparisons are then made between using the old and new sets of variables and the multi-parametric classifiers are compared with the single parameter approach. Finally, Section 6.4 summarises and discusses the results of this chapter.

6.2 Data Description

As with the first case study, the data available from the screening experiment were collected in three batches. The first of 7,680 compounds form the training data. These compounds were selected because of their known properties and were used in a pre-screen to validate the experimental procedures. The remaining two batches each of 15,360 compounds form the test data.

All three batches of data (training and two test sets) were originally classified using the single parameter approach that is described in Section 3.2.3. However, this case study differs from that of the previous one in that an observation was classified as a hit when it was two and a half standard deviations away from the median rather than three. This change in the selection procedure is not uncommon and in this case it was due to the low numbers of hits selected if the threshold were three standard deviations. This is reflected in the classifications shown in Table 6.1.

Table 6.1: Single parameter classification

Observed Classifications	Single Parameter		
	Training	Batch A	Batch B
Hits	0	3	2
Good Hits	12	37	6
False Hits	77	112	113
Non-Hits	1	1	1

Table 6.1 shows the results of classifying the training data and batches A and B of the test data using the single parameter approach. Note that although there are

two sets of variables for this data set, the variable used for the single parameter classifications remains the same throughout. Examining the results of classifying the training data a number of features can be seen. Firstly, no hits and only a small number of good hits have been identified, and secondly, the classifier has found a comparatively large number of false hits. This suggests that the single parameter approach cannot accurately distinguish between the hits and false hits and that there are either a small number of hits and good hits to be found or there are false negatives (i.e. hits classified as non-hits). Moreover, any multi-parameter classifier that is based on this training data will have no information from which to classify hits in the test batches and hence provides evidence that the updating algorithm may benefit classification accuracy.

6.3 Application of Updating Algorithm

This section is concerned with applying the updating algorithm for Section 5.2 to the second high content screening data set that was described in the previous section. All classifications in this section have been made using the random forest classifier (both with and without updating) as this was found to be the most successful classifier of those tested in the previous chapter. Section 6.3.1 concentrates on the classification of the data generated using the old imaging algorithms, Section 6.3.2 focuses on classifying the data generated using the new imaging algorithms and Section 6.3.3 compares the results of the two multi-parametric models with that of the single parameter approach.

6.3.1 Old Variables

Table 6.2 compares the results of classifying the two batches of test data (A and B) using a random forest classifier as part of and independently of the updating algorithm. The results shown in this table are from data generated by using the old imaging algorithms. The first column of this table contains a list of all

possible classifications and the second and third columns show the observed classifications (for batches A and B respectively) of those compounds that were predicted to be hits by the random forest with no updating. When applying the updating algorithm, each iteration (with the exception of the first) uses a different model for prediction than when there are no updates, therefore the fourth column shows the results of classifying batch B after the model has been updated using the addition information from batch A. The fifth and sixth columns contain the observed classifications (for batches A and B respectively) of those compounds that were predicted to be hits by the final updating model. In other words, these are the results of applying the random forest model that has been updated using all batches of data.

Table 6.2: Comparing updating with no updating using a random forest classifier

Observed Classifications	**No Updating**		**Iteration**	**Final Model (Updating)**	
	A	**B**	**B**	**A**	**B**
Hits	2	0	0	2	0
Good Hits	27	3	3	30	4
False Hits	2	1	0	1	0
Non-Hits	0	0	0	0	0

Comparing the results of classifying without updating with those of the final model of the updating algorithm (i.e. comparing the second and third columns of Table 6.2 with the fifth and sixth columns) it can be seen that for both batches there is a slight increase in the number of good hits found with the updating methodology identifying three extra good hits in batch A and one extra in batch B. In addition, the updating methodology has reduced the number of false hits from that of the non-updated random forest with two fewer being found over the two batches. Further comparison with the classifications at the iterative stage of the updating algorithm (column four in Table 6.2) shows that the reduction in false hits for batch B occurs after updating the training data with the classifications from batch A. The reduction in false hits for batch A and the increases in good

hits found in batches A and B occur when applying the final model to the two test batches.

6.3.2 New Variables

Table 6.3 compares the results of classifying the two batches of test data (A and B) using a random forest classifier as part of and independently of the updating algorithm. The results shown in this table are from data generated by using the new imaging algorithms and variables. A list of all possible classifications is contained in the first column of the table and the observed classifications of those compounds that were predicted to be hits by the random forest with no-updating for batches A and B are shown in the second and third columns respectively. When applying the updating algorithm each iteration (with the exception of the first) uses a different model for prediction than when there are no updates, therefore the fourth column shows the results of classifying batch B after the model has been updated using the additional information from batch A. The fifth and sixth columns contain the observed classifications (for batches A and B respectively) of those compounds that were predicted to be hits by the final updated model.

Table 6.3: Comparing updating with no updating using a random forest classifier

Observed Classifications	No Updating		Iteration	Final Model (Updating)	
	A	B	B	A	B
Hits	2	0	0	2	0
Good Hits	26	4	4	28	4
False Hits	2	1	0	0	0
Non-Hits	0	0	0	0	0

Comparing the results of classifying without updating with those of the final model of the updating algorithm (i.e. comparing the second and third columns of Table 6.2 with the fifth and sixth columns) it can be seen that for both batches there is a slight increase in the number of good hits found with the updating

methodology identifying two extra good hits in batch A. In addition, the updating methodology has reduced the number of false hits from that of the non-updated random forest with the updating model classifying no false hits as true hits.

6.3.3 Comparison of Classifiers

The next stage of analysis was to collate the results of the single and multi-parameter approaches from sections 6.2, 6.3.1 and 6.3.2 so that a comparison could be made. This section presents the subsequent findings.

Table 6.4 compares the results of classifying the two batches of test data (A and B) using the single parameter approach and the two random forest classifiers (one using the old variables and the other using the new) from sections 6.3.1 and 6.3.2. Both of the random forest classifiers are applied using the classical non-updating method and the new updating algorithm. The first column of the table contains a list of all the possible observed classifications. The remaining columns show the observed classifications of those compounds predicted to be hits by the five different methods. Note that the results of updating displayed in Table 6.4 are the final results of applying the algorithm and do not represent the number of images that were required to be checked to achieve the final classifications. These results are compared later in Table 6.5.

Focusing on the results of classifying using the random forest in Table 6.4 it can be seen that there is little difference between using the old and new variables. The results of classifying using the random forest with no update (columns three and five of Table 6.4) show that when using the old variables an extra good hit is found in batch A but when using the old variables an extra good hit is found in batch B. However, the results of applying the updating algorithm (columns four and six of Table 6.4) show that two extra good hits are found in batch A when using the old variables while the results for batch B were found to be the same for

both sets of variables. These results suggest that the updating algorithm performs better in terms of classification when using the old set of variables and therefore the remainder of the analysis in this chapter will concentrate on using these variables.

Table 6.4: Comparing the single parameter approach with multi-parameter classifiers

Observed Classifications	Single Parameter Approach		Old Variables				New Variables			
			Random Forest		Random Forest (Updated)		Random Forest		Random Forest (Updated)	
	A	B	A	B	A	B	A	B	A	B
Hits	3	2	2	0	2	0	2	0	2	0
Good Hits	37	6	27	3	30	4	26	4	28	4
False Hits	112	113	2	1	1	0	2	1	0	0
Non-Hits	1	1	0	0	0	0	0	0	0	0

Further comparison of the results in Table 6.4 shows that the most noticeable difference between the single parameter approach and the random forest is the number of false hits that have been incorrectly classified as hits. The updated version of the random forest (using the old variables) only misclassifies one false hit as a hit over the two batches of test data whereas the single parameter approach misclassifies two hundred and twenty five compounds. However, the single parameter approach does identify more hits and good hits than the updated random forest for both batches. These results show that the updated random forest has achieved the objective of reducing the number of false positives but has more false negatives than the single parameter approach. As stated in Section 3.2.5, only a limited number of compounds can be taken forward to the next stage of screening and therefore the most important objective is to reduce the amount of manual image inspection. The results of investigating this are shown in Table 6.5.

Table 6.5 compares the number of hits and good hits found by each classifier to the number of images that were required to be checked in order to achieve the

classification. The first column lists all the classifiers used and, the second, third and fourth columns show the number of hits found, true hits found and images checked respectively. The final column shows the percentage of images checked that were found to be hits. The aim of this analysis is to look at the number of images that are required to be checked by the imaging expert for each of the hits and good hits found.

Table 6.5: Comparing hits found to number of images checked for different classifiers

Classifier	**Hits Found**	**Good Hits Found**	**Images Checked**	**% Images Checked That Were Found to be Hits**
Single Parameter	5	43	275	17.5%
Random Forest	2	30	35	85.7%
Random Forest (Updated)	2	34	38	94.7%

Comparisons of the results in Table 6.4 have shown that the single parameter approach identifies more hits and good hits than both the updated and non-updated random forest. However, as the results in Table 6.5 show, the single parameter approach also requires the largest number of images to be checked in order to achieve this classification (only 17.5% of images checked turned out to be hits). If this is compared to the non-updated random forest it can be seen that there is a great reduction in the number of images that are required to be checked per hit found with 85.7% of the images checked being hits. This figure is then improved upon further by using the updated random forest with 94.7% of the images checked being found to be hits.

From the objectives set out in Chapter 3 and the previous analysis described in Chapter 5, the optimum classifier for the high content screening experiments described in this thesis would find a large number of hits (or good hits) and these hits would correspond to a high percentage of the images that are checked by the screening experiment. The results shown in Tables 6.4 and 6.5 suggest that the

updated random forest fits these criteria better than the single parameter approach and therefore the updated random forest is believed to be the optimal classifier of those tested.

6.4 Summary

The analyses in this chapter have applied the new updating algorithm that was introduced in Chapter 5 to a second high content screening case study. The primary aims of this were to demonstrate that the algorithm can be used on more than just one data set and to further investigate the algorithm with respect to the objectives set out in Section 3.2.5.

The results presented in this chapter have again shown that the updating algorithm improves the predictive capability of the random forest when sequentially classifying batches of compounds. In particular, when using the updating methodology there was an increase in the number of good hits found and a reduction in the number of false hits than when classifying using the non-updated random forest. However, comparison has also shown that the single parameter approach identifies more hits and good hits than both the updated forests but the single parameter approach is much less efficient in terms of the number of images that are required to be checked by the screening expert per hit found. Hence, in this case the updating algorithm has not fulfilled all the objectives that were set out in Section 3.2.5 but has still been shown to be a much more efficient method of classifying high content screening data when compared to the single parameter approach.

The secondary objective of this chapter was at the request of the screening expert and involved comparing classifications when using two different sets of variables. The first set of variables were produced using the same imaging algorithms as were described in Chapter 3 and the second set of variables were produced using a new set of imaging algorithms. The results of these comparisons showed that

when using the updating algorithm there was little difference between the classifications but there were two extra good hits found in batch A when using the old set of variables. It is however important to note that the comparison of the different sets of variables has only been conducted on one data set and further analysis would be required to make general conclusions about which variables are most suitable for classification.

The multi-parameter classifications that have been conducted in this chapter have used the random forest as this was found to be the 'best' method of classification in Chapter 5. Further analysis of the data in this chapter may focus on testing further classifiers to compare with the random forest. It is suggested that different classifiers may perform optimally for data generated using different biological assays or from different experiments.

All of the multi-parameter classifiers that have been applied to the data in this chapter were trained using all of the available variables. In other words, no variable selection has been applied to the data before making classifications and comparisons between classifiers and different methods of generating data. Any further comparisons between classifiers the old and new variables may wish to take this into account as the variables selected may have an effect on the final classifications. This is discussed in further detail and in the wider context of all supervised classifiers in this thesis in Chapter 8.

Chapter 7

Clustering Dose Response Data

7.1 Introduction

This chapter is concerned with using data from high content screening experiments to cluster compounds based on similarities of their dose response on liver cells. As previously described in Section 3.3, this type of analysis is an important step in the evaluation of potential drugs because drug induced liver injury is the most common cause for non-approval, withdrawal, limitation in use, and clinical monitoring by the Food and Drug Administration (Ainscow, 2007a).

The main focus of the work reviews and applies the methodology of Perlman *et al*. (2004a). The results of these analyses show that in its current form this methodology is not suitable for clustering the compounds in the data set being considered and therefore alternative approaches are suggested and applied.

The remainder of this chapter is structured in the following way. A review of the existing methodologies in the literature is given in Section 7.2. The approach of Perlman *et al*. (2004a) is applied to the data set of interest in Section 7.3 before Section 7.4 describes and applies possible changes to the methodology. Finally, Section 7.5 summarizes the chapter.

7.2 Existing Methodologies

There are two specific examples of clustering the dose response effects of compounds on cells in high throughput screening experiments in the literature; Perlman *et al*. (2004a, 2004b) and O'Brien *et al*. (2006). Both of these are described and discussed below.

Perlman *et al*. (2004a) describe an approach of multidimensional drug profiling using the Kolmogorov-Smirnov statistic and titration-invariant similarity scores. (Note that titration is the process of gradually adjusting the dose until a desired effect is achieved. The experiment being considered here defines the doses a priori and therefore the titration terminology should be replaced with dose). In order to apply this methodology a population histogram and cumulative distribution function are generated for each given compound, titration (dose) and descriptor (variable). Given c = 1,…,C compounds, t = 1,…,T titrations and d = 1,…,D descriptors then the effect of a compound c at titration t is assessed by calculating the Kolmogorov-Smirnov statistic $KS_{c,d,t} = KS_{cdt}(p_{c,d,t},q)$ for each of the D descriptors. The KS statistics provides a measure of the maximum vertical distance between the cumulative distribution function of the population response $p_{c,d,t}$ and the cumulative distribution function of the control q. From this z-scores are computed, $z_{c,d,t} = KS_{c,d,t} / std(q_d(n))$, where n is the population size of the cells used to determine $p_{c,d,t}$.

The next stage involves comparing different compounds independently of their titration (dose). Perlman *et al*. (2004a, 2004b) suggest using a titration-invariant similarity score (TISS). The methodology is described as follows. For each compound c, a vector of z-scores across D descriptors and T titrations,

$$X_c = \left(z_{c,1,1},...,z_{c,D,1},...,z_{c,1,T},z_{c,D,T}\right),$$

is formed. In order to allow comparisons of compounds with different titration starting points, a titration sub-series is defined as

$$X_c(s) = \left(z_{c,1,1}, ..., z_{c,D,1}, ..., z_{c,1,T-s}, ..., z_{c,D,T-s} \right)$$

and

$$X_c(-s) = \left(z_{c,1,1+s}, ..., z_{c,D,1+s}, ..., z_{c,1,T}, ..., z_{c,D,T} \right).$$

This truncation of the starting or ending titrations allows the starting point of the series to be "shifted". The s-correlation for all vectors X_i and X_j is then defined as

$$x_{ij}(s) = \left\langle X_i, X_j \right\rangle(s) = \left\langle X_i(s), X_j(-s) \right\rangle \Big/ \left(\left\| X_i(s) \right\| \left\| X_j(s) \right\| \right).$$

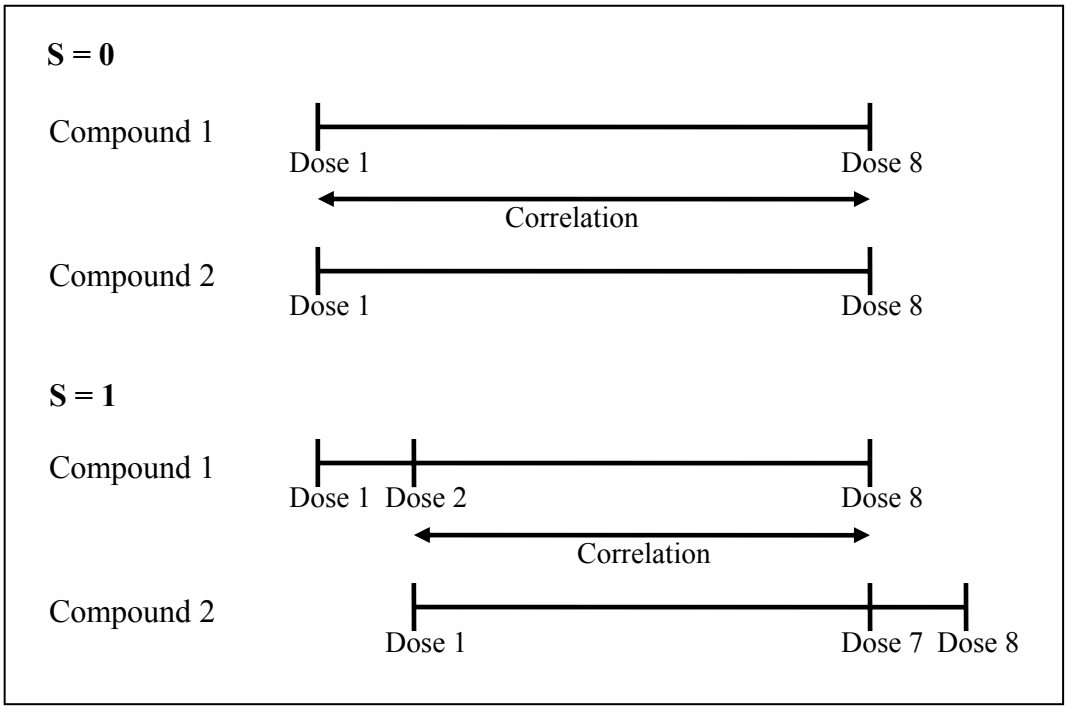
The "shifting" of compounds is illustrated in Figure 7.1. From this figure it can be seen that when s has a value of zero the two compounds are compared with the same range of dose but when s has a value of one this allows each dose of compound 2 to be compared to the higher dose of compound 1.

The final stage of Perlman *et al*. (2004a) involves finding a value of s from a range $-S \le s \le S$ that maximizes the correlation. However, since the s-correlations of compound vectors are not directly comparable for different values of s, a non-parametric ranking is used to normalize these values. Pearlman *et al*. (2004b) suggest that as each matrix followed an approximate Gaussian distribution an s-similarity score could be defined by:

$$\varphi_{ij} = \min\left(\varphi_{ij}(s) \right) = \min\left\{ \#\ \textit{entries in}\ X(s) \ge \left(X_{ij}(s) - 1 \right) \Big/ C^2 \right\}, \qquad (7.1)$$

where C is the number of entries in each matrix. A value of zero corresponds to the most correlated pairs of compounds and a value of one corresponds to the least correlated pairs of compounds.

Figure 7.1: Calculating correlation when "shifting" doses of compounds



The methodology outlined by Perlman *et al*. (2004a) is shown to work when applied to the data set in their paper. However, there are some aspects which raise concerns. The main area of concern relates to the non-parametric ranking of the s-correlations. Using the definitions provided it is not possible to get a full range of similarity scores. Using equation 7.1 the following results can be obtained:

$$X_{ij}(s) = -1 \Rightarrow X_{ij}(s) - 1 = -2 \Rightarrow \varphi_{ij} = 1,$$

$$X_{ij}(s) = 0 \Rightarrow X_{ij}(s) - 1 = -1 \Rightarrow \varphi_{ij} = 1,$$

$$X_{ij}(s) = 1 \Rightarrow X_{ij}(s) - 1 = 0 \Rightarrow \varphi_{ij} \approx \tfrac{1}{2}.$$

Hence it is not possible to get a score of zero if the Gaussian assumption is correct. As an alternative the s-similarity score could be defined as:

$$\varphi_{ij} = \left\{ \#\ \textit{entries in}\ X(s) \geq X_{ij}(s) \middle/ C^2 - C \right\}$$

The first change that has been made is to the denominator of the expression. It is not necessary to consider the diagonal values of the correlation matrix because in the case where the doses are not being shifted $X_{ij}(0) = 1$. This will skew the distribution for calculating a similarity score and in the case where the doses are shifted it does not make sense to compare the same compound with different dose scales. The second change that has been made is to the expression $X_{ij}(s) - 1$ which has been replaced with $X_{ij}(s)$. This makes it possible for φ_{ij} to range between zero and one. Using this definition the following results are obtained:

$$X_{ij}(s) = -1 \ \Rightarrow\ \varphi_{ij} = 1,$$
$$X_{ij}(s) = 0 \ \Rightarrow\ \varphi_{ij} \approx \tfrac{1}{2},$$
$$X_{ij}(s) = 1 \ \Rightarrow\ \varphi_{ij} \approx 0.$$

Where a value of zero corresponds to the most correlated pairs of compounds and a value of one corresponds to the least correlated pairs of compounds.

O'Brien et al. (2006) present a rather ad hoc method for discriminating between positive (toxic) compounds and negative compounds. The method is based around quantifying a dose response relationship for each of five parameters using the IC_{50}, the concentration causing 50% inhibition. Discrimination of positive test results is based on several criteria. Firstly, a minimum of two parameters has to show an effect greater than the variance of their measures across the wells within the plate. Secondly, there needs to be a clear concentration-response relationship for the parameters in question. Finally, the effect is categorized as either equivocal, positive of strongly positive. When a change in parameter is between 1 and 2 times the coefficient of variation of the controls, the test is designated equivocal or weakly positive. If a parameter has a change of 2 or 4 times the coefficient of variation the test is designated positive of strongly

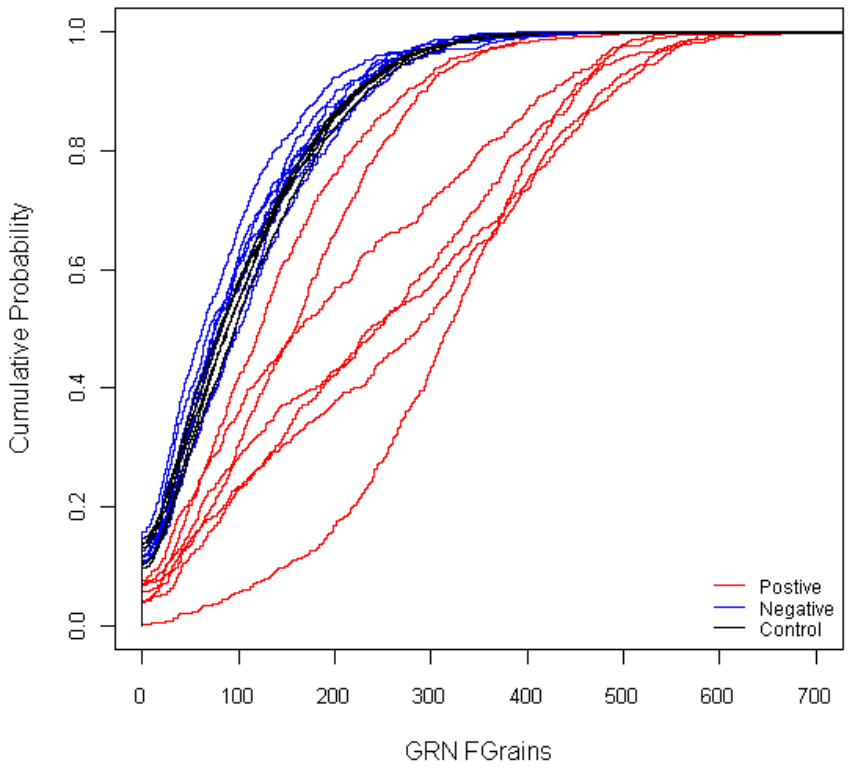
positive. A negative test result is designated to be a result where the above criteria are absent.

The methodology descried by O'Brien et al. (2006) clearly works for the data set that they are considering. However, this approach is very data set specific and does not appear to be suitable as an automated method of classification. For this reason this method has not been pursued any further. The next section will concentrate on the application of the Perlman *et al*. (2004a) procedure.

7.3 Application of Perlman *et al*. Methodology

This section concentrates on applying and evaluating the technique of Perlman *et al*. (2004a, 2004b). The main method of evaluating the results of the analysis was by examining the clusters of compounds that were produced when visualising the calculated similarities through principal coordinate plots. Discussion is also included on the different ranges of dose shifts that are allowed through the methodology and the effect these have on the clustering.

Figure 7.2: Cumulative distribution function curves



The first stage of applying the Perlman *et al*. (2004a) procedure involved plotting cumulative density curves. This was done for a number of reasons. Firstly, the controls from a number of different plates were compared in order to assess any between plate variability. The plot of these curves in Figure 7.2 shows that there is a small amount of variability between the plates (as is to be expected) but not enough to suggest that there could be any plate effect on the compound measurements. Secondly, curves were plotted for a number of compounds (both positive and negative) at different doses to see how these compounds differed from the controls. Examining Figure 7.2 it can be seen that the cumulative density curves corresponding to the negative compounds are clustered around the control compound with small Kolmogorov-Smirnov distances (both positive and negative). The positive compounds are located away from the cluster of negative and control compounds with larger negative Kolmogorov-Smirnov distances.

Figure 7.3 shows a plot of the first two principal coordinate axes of compounds with no shift in dose allowed. The corresponding plots in Figures 7.4, 7.5 and 7.6 show compounds with a maximum dose shift of one, two and three respectively. Each compound on the plots is coloured according to the level of phospholipidosis expected and the reference compounds are labelled using letters.

Examining the principal coordinate plots in Figures 7.3, 7.4, 7.5 and 7.6 it can be seen that there are no coherent clusters. In particular, with the exception of a couple of compounds, those compounds which are positive are positioned in the same cluster as those which are negative (although plots of principal coordinates one and two are displayed here plots of higher dimensions were also examined). In addition, some of the reference compounds that have been replicated in the screen are not positioned in the same principal coordinate subspace (for example, CCCP and Fluoxetine in Figure 7.3) suggesting that either the measurements taken from the experiment are not consistent for identical compounds that have been replicated or the analysis being used is not appropriate for capturing the biological activity in the data set.

Figure 7.3: Principal coordinate plot of compounds with no dose shift allowed

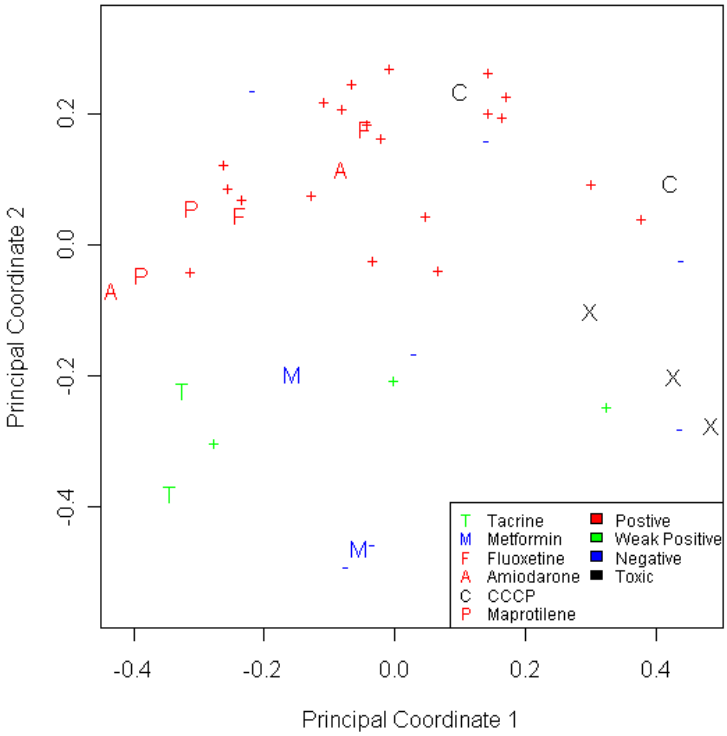


Figure 7.4: Principal coordinate plot of compounds with maximum of one dose shift allowed

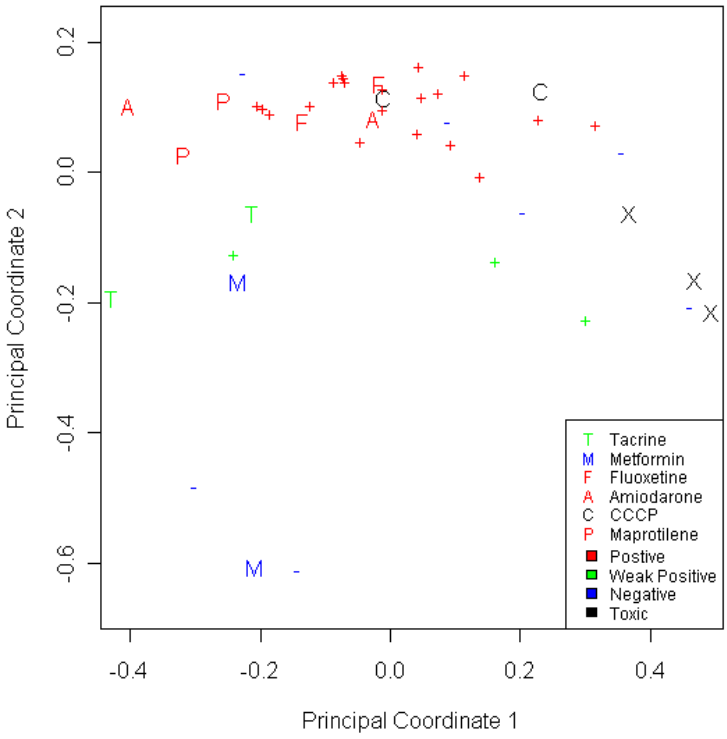


Figure 7.5: Principal coordinate plot of compounds with maximum of two dose shifts allowed

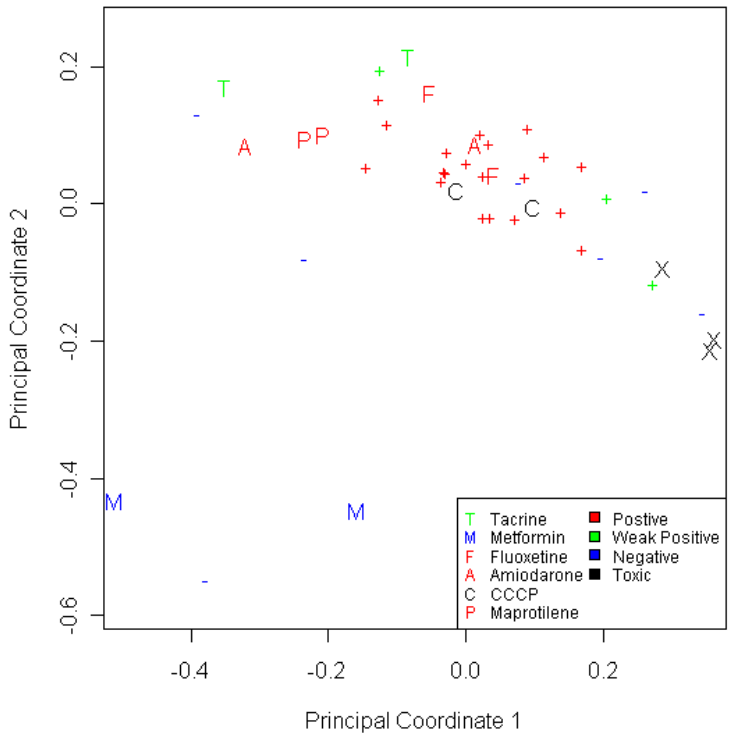
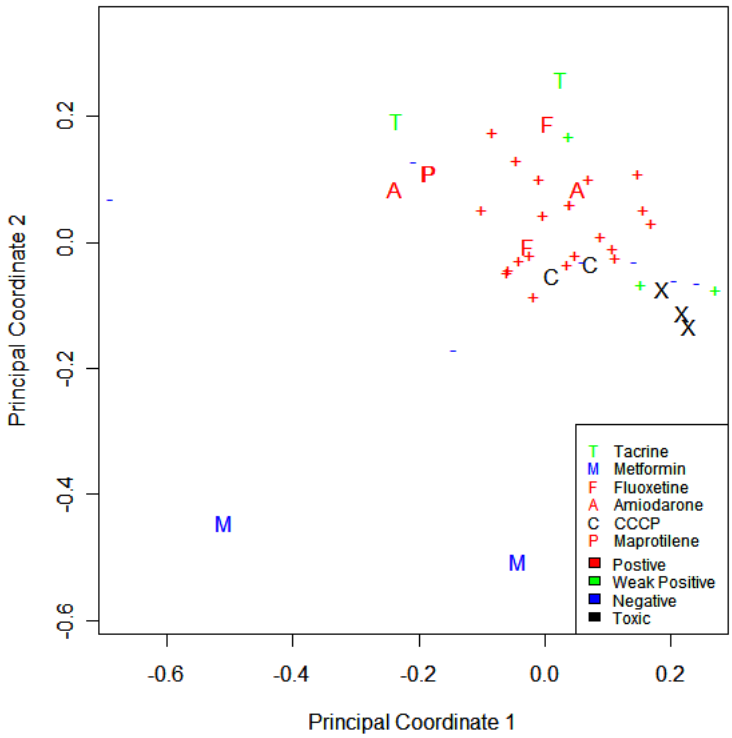


Figure 7.6: Principal coordinate plot of compounds with maximum of three dose shifts allowed



Further investigation of the reference compounds in Figures 7.3, 7.4, 7.5 and 7.6 was carried out by looking at which shift in dose (-3, -2, -1, 0, 1, 2 or 3) minimized the dissimilarity between replicates of identical compounds. The screening expert suggested that variability in the experiment may account for a maximum of one shift in dose between the identical compounds. In other words, when comparing two different replicates of an identical compound the shift that minimizes the dissimilarity should have a value of -1, 0 or 1. Examining the compounds it was found that the minimising dose shifts for all compounds could be explained by the variability in the experiment with the exception of Metformin compounds which were found to have a minimizing dose shift of 2 or -2 (depending on which order the compounds are labelled). This suggests that the features of the reference compounds seen in the principal coordinated plots may be a feature of the statistical methodology rather than the experimental procedure.

In addition to the previously described features of Figures 7.3, 7.4, 7.5 and 7.6, it can also be seen that the majority of compounds are clustered around an arc with a small number of negative compounds (including Metformin) at the centre. This feature appears to be a consequence of using correlation as a measure of similarity and the position of the compounds relative to the controls. It may also explain why negative compounds are not separated from positive ones. As described in Section 7.2, the first stage of analysis is to calculate the KS distance between the control and each compound (for each descriptor at each dose). In the data set that is being used the control is negative (specifically it is Metformin at a fixed dose) and therefore is located at one extreme of the toxicological scale. Hence, the majority of KS distances will have the same sign (i.e. positive or negative) with the majority negative compounds being a short distance from the control and the positive compounds being further away. Therefore using correlation (measured as the cosine of the angle θ between two vectors) as a measure of similarity will not separate those compounds which are negative from those that are positive. This is demonstrated graphical using a two-dimensional example in Figure 7.7.

Figure 7.7: Clustering using correlation as a measure of similarity:
(a) KS distances represented as vectors for a control at an extreme; (b) KS distances represented as vectors for a centred control; (c) principal coordinate plot corresponding to a control at an extreme; and (d) principal coordinate plot corresponding to a centred control.

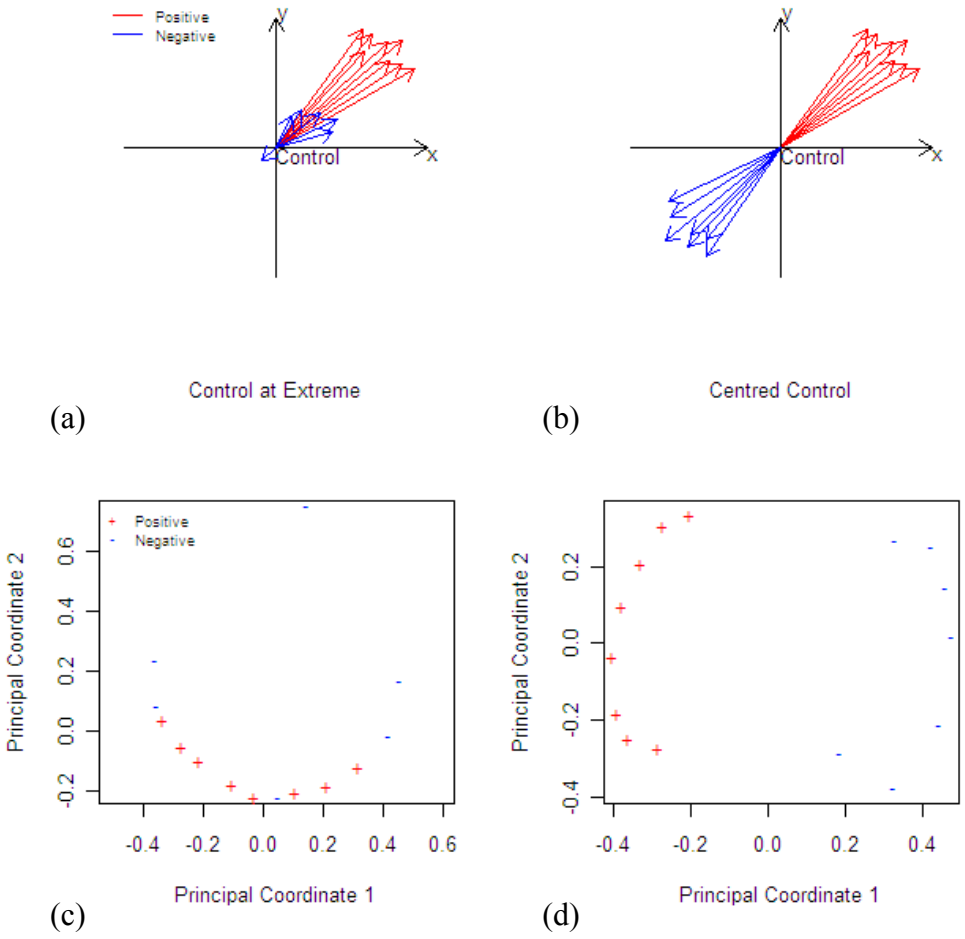


Figure 7.7 shows two artificial examples of the use of KS distances and ranking correlation to calculate dissimilarity. The top left-hand plot (a) shows fourteen compounds represented as vectors. Each axis represents the KS distance between the compound and a control compound that is located close to an extreme. With the exception of one compound each of the KS distances is positive with the values for the positive compounds being greater than those of the negative. The final negative compound has negative KS distances for each variable. Calculating dissimilarity using the ranked correlation method of Perlman *et al*. (2004a, 2004b) the compounds are displayed on principal coordinates in the bottom left-hand plot

(c). Examining this plot and comparing the compounds to the vector diagram it can be seen that the those compounds that have positive KS distances are positioned in an arc with the second principal coordinate having values of less that 0.3 and the single compound with negative KS values is separated with the second principal coordinate having a value of approximately 0.7. These features correspond to those seen in Figures 7.3, 7.4, 7.5 and 7.6, and therefore in this case correlation is not an appropriate measure of similarity.

One solution to the problem of using correlation as a measure of similarity may be to use a control that is positioned between the two groups of interest and not at an extreme. This case is presented in the second example in Figure 7.7. The top right-hand plot (b) shows the compounds represented as vectors but this time all the positive compounds have positive KS distances for both variables and all of the negative compounds have negative KS distances for both variables. Plotting these compounds on principal coordinates (plot (d) in Figure 7.7) using the same methodology of calculating dissimilarity as before it can be seen that the two groups are separated with the positive compounds having negative values for the first principal coordinate and the negative compounds have positive values.

7.4 Euclidean Distance as a Measure of Similarity

In Section 7.3 the approach of Perlman *et al*. (2004a, 2004b) was applied to the dose response data from Chapter 3. It was shown that in its current form this method cannot distinguish between the different groups in the data. In particular, the use of correlation as a measure of similarity was found to be inappropriate for the data in question. This section concentrates on using Euclidean distance as a measure of similarity to see what effect is made on the clustering of the compounds.

Considering again the top left-hand plot of Figure 7.7 suggests that when the control compound is situated at an extreme, Euclidean distance maybe an appropriate measure for distinguishing between the different groups of compounds. However, simply replacing the correlation measure with Euclidean distance in the Perlman *et al*. (2004a, 2004b) technique would not work as the non-parametric ranking method for comparing different ranges of dose (see Section 7.2) would not be appropriate. Therefore an initial analysis was carried out ignoring any shift in dose so that judgement could be made on how well this change performed.

Figure 7.8: Principal coordinate plot using Euclidean distance as a measure of similarity

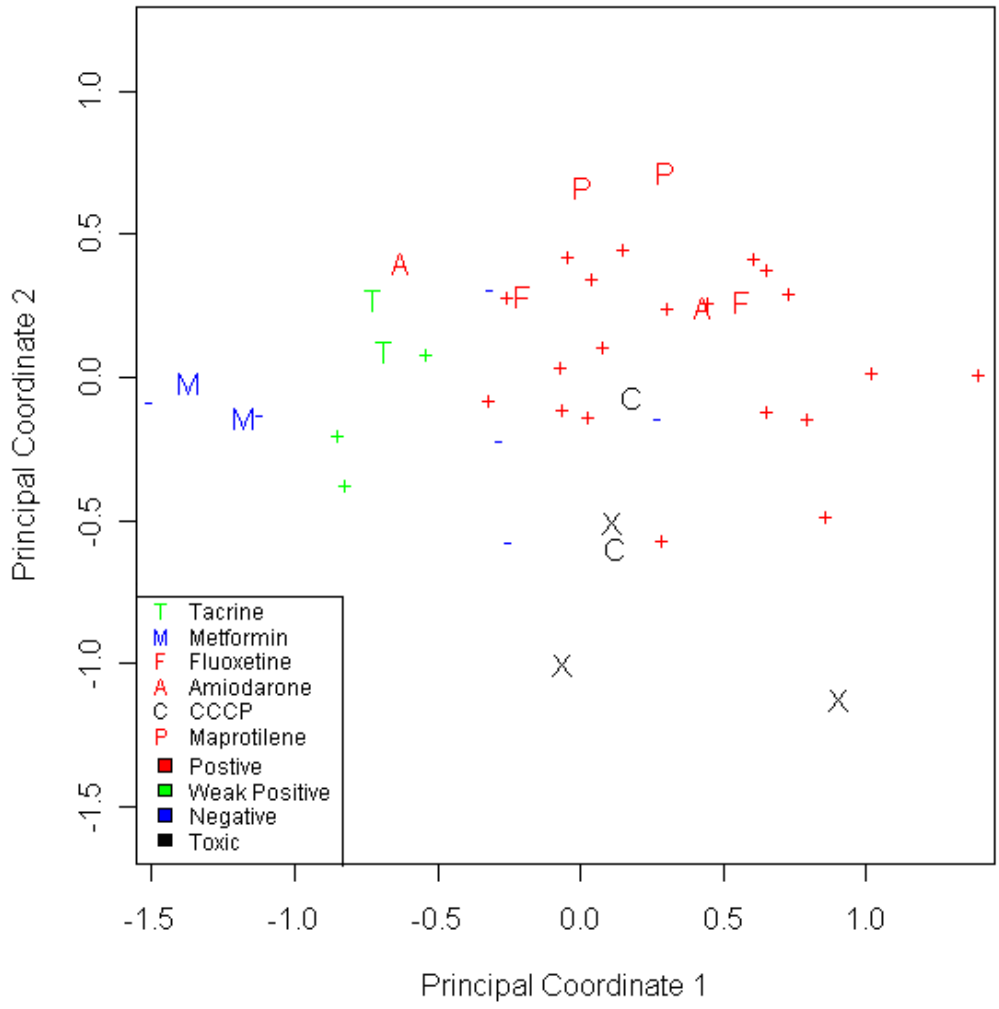


Figure 7.8 shows the principal coordinate plot produced by replacing correlation with Euclidean distance as the measure of similarity. Examining this plot it can be seen that there is some evidence that the clusters correspond to the negative, weak positive and positive compounds. However, there are four negative compounds with are located in or around the cluster of positive compounds. Looking at the reference compounds it can be seen that the replicates of Metformin, Tacrine and Maprotilene are positioned close to each other but the replicates of CCCP, Amiodarone and Fluoxetine are not positioned as close to each other as expected (however, this may be due to variation in the experiment and therefore may change when taking shifts of doses into account). Comparing Figure 7.8 with Figures 7.3, 7.4, 7.5 and 7.6 it can be seen that the compounds are no longer positioned in an arc around the Metformin compounds. This suggests that Euclidean distance is a more appropriate measure of similarity for this data than correlation.

In order to investigate this further an alternative approach to the original non-parametric ranking is required (see Section 7.2 for details). This method must minimize over Euclidean distances that have been calculated in different numbers of dimensions (for example, it would be incorrect to compare a distance that was calculated in two dimensions to one that was calculated in three). It may be possible to compare distances that have been calculated in different numbers of dimensions by weighting the measurements. In other words, each distance matrix could be multiplied by a different constant depending on how many dimensions were used to calculate the distance. It would then be possible to minimize over the different matrices. However, before implementing such a method the weightings would need to be calculated. It may be possible to calculate appropriate weightings by simulating distance matrices. This analysis has not been pursued.

7.5 Summary

This chapter has described and evaluated a number of different methods for clustering dose response data from high content screening experiments. Through the application of the Perlman *et al*. (2004a) procedure it has been shown that clustering using correlation as a measure of similarity is not appropriate when the experimental control is located at an extreme (as in the case with the data set in this thesis). Further analysis replacing correlation with Euclidean distance has indicated that there may be some potential in this method of clustering; however, no clustering has been performed using Euclidean distance when the doses of compounds have been 'shifted'.

In order to continue the clustering of dose response data using this methodology, a procedure for comparing compounds that have been 'shifted' would be required. One possible approach to comparing compounds in this way may be to use simulation to calculate weightings for distances calculated in different numbers of dimensions. By weighting the measurements it would be possible to minimize distances between compounds with different dose ranges.

When investigating the clustering of dose response data in this chapter a number of further analyses were conducted which did not prove to be useful for the data set being considered but which may be useful when analysing future data from high content screening experiments or any other problems which involve the clustering of dose responses. These are discussed in Chapter 8.

Chapter 8

Further Work

8.1 Introduction

This chapter describes some suggestions for alternative analyses and further work for some of the methods applied in this thesis. All the suggestions made about the analysis of the compound selection data (Chapters 4, 5 and 6) concern areas of further analysis while some of the suggestions made about the dose response data (Chapter 7) are applied but shown not to be appropriate in this particular case. However, the methods described may be appropriate for the analysis of data arising from different contexts. The first three sections concentrate on the analysis of the compound selection data. Section 8.2 suggests altering the size of batches for the updating algorithm introduced in Chapter 5 before Section 8.3 discusses variable selection for classifiers. A method of ambiguity rejection using random forest is suggested in Section 8.4. The next two sections concentrate on the problem of clustering dose response data with Section 8.5 applying the method of unsupervised random forests before Section 8.6 introduces profiling using single cells. Finally, Section 8.8 contains a summary and discussion.

8.2 Batch Size

The two data sets used to test the new updating algorithm in Chapters 5 and 6 were divided into batches based on runs of the experiment. In other words, each batch of compounds corresponds to a different run. This approach seems sensible as the analysis of one batch can be carried out whilst the next batch of compounds is being processed through the experiment. However, the number of compounds in each batch may have an effect on the overall classification results of the algorithm. For example, by reducing the number of compounds in each batch and therefore increasing the number of batches, the algorithm will update more times and this may produce increased classification accuracy. This should be investigated further to see if there is an optimal batch size in terms of the number of hits selected, the cost (time) of analysing the batches and the number of images that are required to be checked by an imaging expert.

8.3 Variable Selection

All of the multivariate analyses conducted in this thesis have used the full set of variables available. In terms of the number of variables measured this has not been a problem (sixteen variables is not particularly many by comparison with classification problems in genetics or proteomics with hundreds or even thousands of variables). However, by reducing the number of variables models may be simplified and this may have an impact on classification performance.

Further analyses of the updating algorithm should investigate the effect of variable selection on the results produced. In terms of the random forest classifier there are methods available in the R package to rank the variables in order of performance. However, choosing the best k variables based on their rank may not result in the selection of the best subset of variables (Hand, 1997).

Alternative methods of variable selection look for a 'best' subset of variables. An example of this is stepwise selection. Stepwise methods begin by comparing all the variables individually. The 'best' is chosen by using some measure of impurity or separability. All remaining variables are then examined to identify which yields greatest between class separability when combined with the first. At each subsequent step the variable that produces the best results when combined with the previously chosen variables is added to the subset. However, the subset of variables that are chosen may still not be the best because many potential subsets will not have been examined by this procedure (Hand, 1997). A review of further methods for variable selection can be found in Hand (1997) and Guyon and Elisseff (2003).

8.4 Random Forest Ambiguity Rejection

In Chapter 4 the ambiguity reject option was applied to the method of using unlabelled data to update classification rules when the CEM algorithm would not converge. In this section the use of the ambiguity reject option is investigated further but this time in the context of the random forest classifier.

The previous applications of the random forest in this thesis classified an observation to a group based on the majority vote from all trees in the forest. Using this method of majority vote means that the classifications of some observations have more uncertainty associated with them than others. For a simple illustration of this consider a two class problem as an example. The first observation is classified to group 1 with 90% of the votes and the second observation is also classified to group 1 but with only 51% of the votes. This suggests that the random forest is "less certain" about the classification of the second observation (with 49% of the votes against the given classification) than the first observation (with 10% of the votes against the given classification). Another way to view this is that the second observation is located near to a decision boundary. In the case of high content screening data there are three

groups in which to classify the observations but the idea of uncertainty associated with classifications is just the same as the two group example.

Table 8.1: Comparing classification results using different random forest thresholds

Percentage of Votes Required	Hits	False Negatives	False Positives	Images Checked
50	103	6	31	134
55	98	11	29	127
60	87	22	23	110
65	79	30	20	99
70	70	39	11	81
75	61	48	6	66
80	54	55	1	55
85	40	69	1	41
90	29	80	0	29
95	9	100	0	9

Initial investigation into random forest ambiguity rejection calculated the number of hits, false negatives, false positives and the number of images checked for differing percentages of votes required for classification. The results of these calculations for batch A of the data described in Section 3.2 are displayed in Table 8.1 (note that the column labelled hits is the combined total of hits and good hits). Examining the results it is clear to see that the number of hits found reduces and the minimum number of false negatives increases as the percentage of votes required increases. In addition, the number of false positives and images checked reduces as the percentage of votes required increases.

To investigate random forest ambiguity rejection further a method of comparing the results of classifying using different thresholds of votes would be required (i.e. a method to compare classifications when 90% of votes are required against classifications when 60% of votes are required). One method of comparing classifiers in this way is to use a ROC (Receiver Operating Characteristic) curve (see Hand (1997) for details). However, this requires the true classifications to be known for each of the groups so that sensitivity and specificity can be calculated.

As originally described in Section 2.4 the true classifications of all the compounds in the high content screening data sets analysed in this thesis are not known (they would all need classifying by eye by a high content screening expert). Hence, this method is not appropriate in its current form.

8.5 Classification using Unsupervised Random Forests

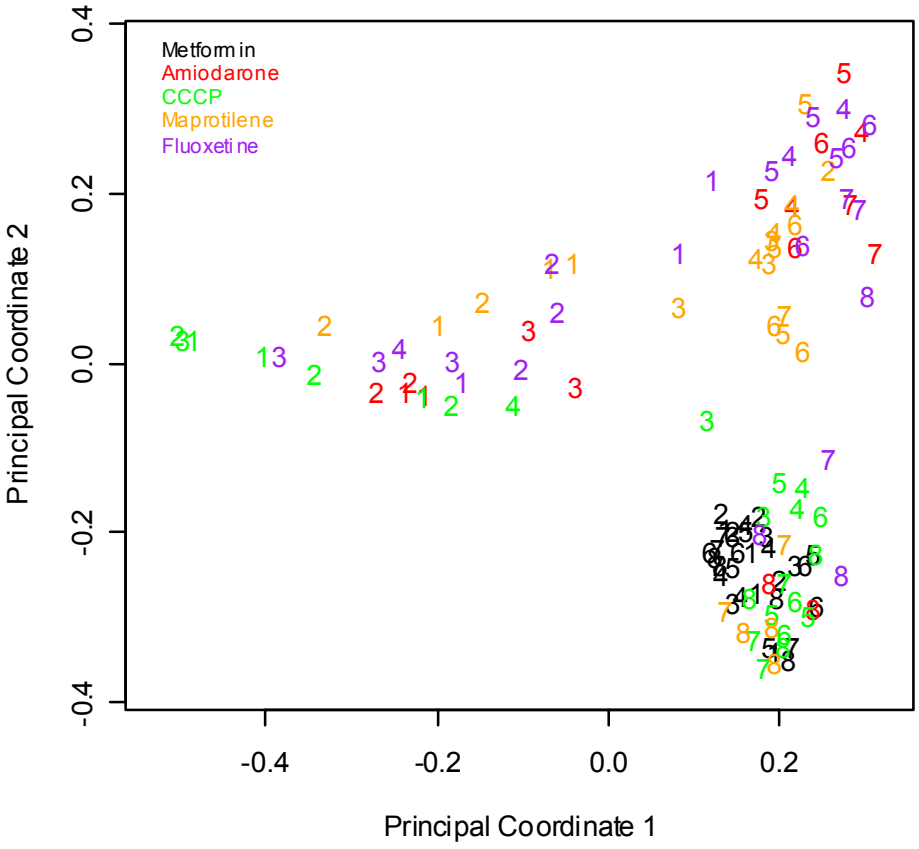
In Chapters 5 and 6 random forests were used as a method of supervised classification. However, it is also possible to use random forests for unsupervised classification. This is done by creating an artificial class label for the unsupervised data that distinguishes the 'observed data from suitably generated 'synthetic' data. The approach of random forests to unsupervised learning is to consider the original data as class 1 and then create a synthetic second class of the same size that will be labelled as class 2. The class 2 data can be sampled in two different ways. The first is to sample by independent bootstrap each variable separately and the second is to randomly sample from the hyper-rectangle that contains the observed data. The supervised random forest method can then be used on this artificial two-class problem (Shi and Horvath, 2006).

The idea behind the unsupervised random forest is that real data points that are similar to one another will frequently be found in the same terminal node of a tree. This frequency can be measured in a proximity matrix. This proximity matrix can then be taken as a similarity measure, and clustering or multi-dimensional scaling using this similarity can be used to divide the original data points into groups for visual inspection (Liaw and Wiener, 2002).

As an initial investigation of using unsupervised random forests to cluster compounds the measurements from five of the reference compounds were used. In this case each dose of each compound was treated as being a separate observation. Figure 8.1 shows the results of displaying the compounds on the first two principal coordinates. Each point on the plot is coloured to identify which

compound it is and numbered to show which dose it corresponds to (doses are numbered from one to eight with one corresponding to the weakest dose and eight corresponding to the strongest). Inspecting this plot it can be seen that the clusters are not representative of the groups that are expected. For example, the cluster with principal coordinates (0.2, -0.2) contains the compound Metform (which is negative) and some of the higher doses of CCCP (which is toxic) and Maprotilene (which is positive). This suggests that this is not a suitable method for clustering the data being considered.

Figure 8.1: Principal coordinate plot of unsupervised random forest clustering



8.6 Profiling of Drug Responses using Single Cells

The methods of clustering dose response data in this thesis have used approaches that summarize the information from individual cells (for example, Perlman *et al*. 2004a) draw cumulative distribution function curves and calculate Kolmogorv-Smirnov distances). An alternative approach is to use the measurements from individual cells for analysing the effects of compounds. Loo *et al*. (2007) and Young *et al*. (2008) have both described different methods of using individual cell measurements in this way.

Loo *et al.* (2007) describe an approach that uses support vector machines to classify untreated and treated cancer cells based on phenotype measurements. For each dose of each compound a support vector machine algorithm is used to determine a hyperplane that separates the treated and control distributions. The classification accuracy score of the hyperplane indicates the degree to which the populations are separated and the normal vector, used as a multivariate profile, indicates the line of greatest separation. Redundant and non-informative variables are then removed before the titration series for each compound is partitioned into ranges with minimum dissimilarity. For each partition a representative dosage range (d-profile) is obtained by averaging the constituent profiles. These d-profiles are then compared with the profiles of reference compounds to see which is most similar.

Young *et al*. (2008) describe a method of using factor analysis to profile compound activities. Given a matrix containing n variables and m cells, factor analysis is used to reduce the dimensionality of the data to a k-dimensional space described by a set of k factors. These k factors reflect the underlying attributed measured in the data. Based on the factor model, regression is used to estimate scores for each factor on a cell-by-cell basis for each compound. Each compound is then summarized as the mean score on each factor (i.e. the average of each factor for all cells in the well). Hierarchical clustering of the factor scores is then used to profile the biological activity of the compounds.

The approaches of profiling using single cells by Loo *et al.* (2007) and Young *et al*. (2008) are possible methods to be considered for future analysis. However, using the information from single cells for analysis may cause problems. For example, visualising any data would be much more difficult. Instead of one point being plotted for each dose of each compound there would be between two hundred and fifty and three hundred. This would be very confusing and any useful information about the structure of the data may be lost. In addition, the handling of large quantities of data may become an issue because relatively small screens will still have many millions of individual measurements.

8.7 Summary and Discussion

This chapter provides a review of some suggestions of alternative analyses and further work to the research in this thesis. Although the alternatives analyses did not improve upon those carried out in previous chapters, these methods may be more suitable for different data sets and other contexts.

The main suggestions regarding the compound selection data have been concerned with further work. The analysis that has been suggested for investigating batch size and variable selection is important for optimizing the updating algorithm in terms of hits selected as well as the number of images that are required to be visually checked by the imaging expert. Further to this, the random forest ambiguity rejection may also be used for improving these criteria but it would also allow some flexibility for classification of different data sets with different objectives. However, a method is still required for comparing classifications made with different thresholds of the reject option.

Sections 8.5 and 8.6 described some alternative methods for analysing the dose response data. Unsupervised random forests and the fitting of hyperbolic distributions were found not to be appropriate for the analysis of the high content

screening data set in question. Nevertheless, there are examples in the literature where these methods have been applied successfully in different contexts. The final section discussed the use of profiling drug responses using single cells. The two approaches of Loo *et al*. (2007) and Young *et al*. (2008) were described as possible alternative methods of analysing the dose response data. However, as with many of the existing methods of clustering found in the literature it is not clear how the approach of Young *et al*. (2008) would be adapted to allow compounds that have different dose ranges to be compared.

The investigation of batch size would be the most useful next step in analysis of the compound hit selection problem. By finding an optimal batch size the predictive performance of the updating algorithm would be improved. However, this work would be very time consuming in terms of conducting the statistical analysis and having the compound images checked by a screening expert. The most promising approach of those outlined in this chapter for clustering dose response data is that of Loo *et al*. (2007). Unlike the other methods described, this one takes into account the different dose ranges of the compounds. However, there may be difficulties in implementing this technique because the single cell data would have to be extracted from the high content screening software and there may be issues with handling the data as there would be many millions of individual measurements.

Chapter 9

Summary and Conclusions

This chapter provides a brief summary of the main chapters of this thesis and the conclusions to be drawn about the methodologies used and their application to both high content screening experiments and in other contexts.

An overview of multivariate classifiers was given in Chapter 2. A taxonomy of classifiers highlighted the differences between supervised and unsupervised classification before subgroups of methods belonging to each were discussed. The taxonomy allows approaches to classification to be compared in groups rather then comparing individual classifiers. Furthermore it can be used to identify if a particular group of techniques performs better than another. For example, if tree classifiers perform better than Fisher's linear discriminant analysis this may suggest that structural approximation of classification boundaries are more suited to the problem than functional approximation. The remainder of the chapter gave specific details of the methodologies used in subsequent chapters.

Chapter 3 provided an introduction to the case studies that motivated the work in this thesis. Exploratory analyses of the data from each case study were used to identify a number of key features and problems. Visualisation of the compound hit selection data demonstrated that it would be difficult to distinguish between the three groups of interest. This was further reflected in the number of true hits that were misclassified as false hits when examining the results of classifying this data using a number of existing multivariate classifiers. The exploratory analyses

that were conducted for the dose response clustering problem all concentrated on methods of visualising the compounds and their responses. However, the methods applied did not take into account the problem of how to compare responses over different dose ranges. The descriptions of key problems and exploratory analyses from this chapter may be used to identify other areas of research that may benefit from using the statistical techniques described in later chapters.

Chapter 4 was concerned with the updating of classification rules using unlabelled data. This analysis was performed using the method described by Dean *et al*. (2006) and extended by incorporating robust estimates of multivariate location and scale. Overall it was concluded that there was not enough gain in predictive performance from updating in this manner for it to be considered over existing methodologies. However, the investigation of this method did bring to our attention the idea of updating classification rules and it also high lighted a key problem. The methodologies that were being applied all make the assumption that the training data is randomly sampled from the same distribution as the test data. With the data used for supervised classification in the majority of this thesis this assumption is false as the training data are selected because of their known properties. With this in mind the new updating algorithm in Chapter 5 was developed.

Chapter 5 introduced a new classification method for batches of compounds where the rule is updated sequentially using information from the classification of previous batches. The aim of this new algorithm was to take into account a reduction in classifier performance due to the training data not being representative of the test data and that the distributions of the groups in the data change as new batches of compounds are introduced. The application of the updating algorithm showed that it improved the predictive capability of the four classifiers it was tested on. In particular, all of the updated classifiers have identified more hits than their equivalent non-updated classifiers and the single parameter approach. In addition, by updating random forests, k-nearest

neighbours and mixture discriminant analysis it has been shown that they are more efficient in terms of the number of images that are required to be checked by the screening expert per hit found than when not updating. In general, the updating algorithm can be used in any situation where data is analysed in batches. It is expected to make an improvement on the results of any classical classifier when the underlying distributions change as new batches of data are analysed. Furthermore, it can be used with any classifier so it is flexible enough to adapt to different problems.

Chapter 6 introduced a new case study so that the new updating algorithm that was introduced in Chapter 5 could be studied further. All multi-parameter analyses that were conducted in this chapter used the random forest classifier. The results indicated that by using the updating algorithm the predictive capability of the random forest was improved. In addition, the updated random forest was shown to be more efficient (in terms of the number of images that needed checking) than the single parameter approach. Ninety-five percent of the compounds that were selected by the updating algorithm were found to be hits whereas only eighteen percent of compounds selected by the single parameter were found to be hits. However, the single parameter approach found less false negatives than the updated random forest suggesting that the results can be improved upon further.

A number of different methods from clustering high content screening dose response data were described and evaluated in Chapter 7. Through the application of the Perlman *et al*. (2004a) methodology it was shown that clustering using correlation as a measure of uncertainty is not appropriate when the experimental control is located at an extreme. Further analysis replacing correlation with Euclidean distance in the Perlman *et al*. (2004a) methodology has shown that there may be some potential in this method of clustering; however, no clustering has been performed using Euclidean distance when the doses of compounds have been 'shifted'. There is much further work to be done in this area before a satisfactory method for clustering compounds with different dose ranges is found.

Chapter 8 described some suggestions for alternative analyses and further work for some of the methods applied in this thesis. All the suggestions made about the analysis of the compound selection data concern areas of further analysis while some of the suggestions made about the dose response data are applied but shown not to be appropriate in this particular case. The area of most interest for further work would be the investigation of batch size for the updating algorithm. By finding an optimal batch size the performance of the algorithm would be improved further but this work would be time consuming both in terms of statistical analysis and the checking of images by the screening expert.

The analyses that have been conducted in this thesis have focused on specific problems from high content screening experiments. However, the methods of selecting compounds that have been described are robust enough to be applied to other screens that use different biological assays and to problems from completely different contexts. The new updating algorithm that was described and applied to two data sets in Chapters 5 and 6 has been shown to improve the predictive capability of the classifiers it has been applied with. It has also been shown to reduce the number of images that are required to be manually checked by a screening expert. The analyses in Chapter 5 identified the random forest as being the 'best' classifier to analyses the data in question but as the algorithm can be used with any classifier it can adapt to situations where the random forest does not perform well or is unsuitable.

The problem of clustering dose response data is a difficult one especially when responses are measured over different dose ranges. The work in this thesis has investigated a number of possible approaches to solving the problem and has identified some details in existing methodologies that make them inappropriate for the data in question. It is clear that there is much more work required before a satisfactory procedure is established.

References

[1] Ainscow, E.K. (2007a) Definiens Cellenger Analysis for Hepatocyte Phospholipidosos (PLD) Screening. *Definiens IN-SIGHT Webinar Series.* URL: http://definiens.com/binary_data/473_070425_definiens_webinar.pdf

[2] Ainscow, E.K. (2007b) Statistical Techniques for Handling High Content Screening Data. *European Pharmaceutical Review*, **5**.

[3] Bensmail, H. and Celeux, G. (1996) Regularized Gaussian Discriminant Analysis Through Eigenvalue Decomposition. *J. Am. Stat. Assoc.*, **91**, 1743-1748.

[4] Biernacki, G. and Govaert, G. (1999) Choosing Models in Model-Based Clustering and Discriminant Analysis. *J. Stat. Comput. Sim.*, **64**, 49-71.

[5] Breiman, L. (1994) Bagging Predictors. *Technical Report No. 421*. Department of Statistics, University of California.

[6] Breiman, L. (2001) Random Forests. *Mach. Learn.*, **45**, 5-32.

[7] Breiman, L., Friedman, J.H., Olshen, R.A and Stone, C.J. (1984) *Classification and Regression Trees*, California: Wadsworth International Group.

[8] Brentnall, A.R., Crowder, M. J. and Hand, D. J. (2008) A statistical model for the temporal pattern of individual automotive machine withdrawals. *Applied Statistics*, **57**, 43-59.

[9] Campbell, N.A. (1978) The influence function as an aid to outlier detection in discriminant analysis. *Applied Statistics*, **27**, 251-258.

[10] Celeux, G. and Govaert, G. (1992) A Classification EM algorithm for clustering and two stochastic versions. *Computation Statistics & Data Analysis*, **14**, 315-332.

[11] Chambers, J.M. and Hastie, T.J. (1993) *Statistical Models in S*, New York: Chapman & Hall.

[12] Clemons, P.A. (2004) Complex phenotypic assays in high-throughput screening. *Curr. Opin. Chem. Biol.*, **8**, 334-338.

[13] Cooke, E.L., Ainscow, E.K., Hargreaves, A., Sullivan, E., Alcock, P., Ellston, J., Peters, S., Major, J. Wannop, J., Allen, H., Plant, D., Coundhry, S., Hicks, R., McCall, E., Shaw, J. and Ronco, L. (2003) G-Protein-Coupled Receptor High-Throughput Screen Using Norak Transfluor® Technology and the IN CELL Analyser 3000. *Proceedings of the SBS 9th Annual Conference*.

[14] Cox, T.F. and Cox, M.A. (2001) *Multidimensional Scaling*, Boco Raton: Chapman & Hall/CRC.

[15] Dasarathy, B.V. and Sheela, B.V. (1978) A Composite Classifier System Design: Concepts and Methodology. *Proceedings of the IEEE*, **67**, 708-713.

[16] Dean, N., Murphy, T.B. and Downey, G. (2006) Using Unlabelled Data to Update Classification Rules with Applications in Food Authenticity Studies. *Applied Statistics*, **55**, 1-14.

[17] Dubuisson, B. and Masson, M. (1993) A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition*, 26, 155-165.

[18] Everitt, B. (2005) *An R and S-PLUS® Companion to Multivariate Analysis*, London: Springer.

[19] Everitt, B.S., Landau, S. and Leese, M. (2001) *Cluster Analysis (Fourth Edition)*, London: Arnold.

[20] Fisher, R.A. (1936) The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.*, **7**, 179-188.

[21] Fisher, R.A. (1938) The Statistical Utilization of Multiple Measurements. *Ann. Eugen.*, **8**, 376-386.

[22] Fraley, C. and Raftery, A. (2007) mclust: Model-Based Clustering / Normal Mixture Modelling. R package version 3.1-1. http://www.stat.washington.edu/mclust

[23] Fraley, C. and Raferty, A.E. (2002) Model-Based Clustering, Discriminant Analysis and Density Estimation. *J. Am. Stat. Assoc.*, **97**.

[24] Gagarin, A. Makarenkov, V. and Zentilli, P. (2006) Using Clustering Techniques to Improve Hit Selection in High-Throughput Screening. *J. Biomol. Screen.*, **11**, 903-914.

[25] Guyon, I. and Elisseeff, A. (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**, 1157-1182.

[26] Hand, D.J. (1997) *Construction and Assessment of Classification Rules*, Chichester: John Wiley & Sons.

[27] Hand, D.J. (2006) Classifier Technology and the Illusion of Progress. *Stat. Sci.,* **21**, 1-14.

[28] Hartley, H.O. and Rao, J.N.K. (1968) Classification and Estimation in Analysis of Variance Problems. Review of International Statistics Institute, **36**, 141-147.

[29] Hastie, T. and Tibshirani, R. (1996) Discriminant Analysis by Gaussian Mixtures. *J. R. Statist. Soc. B*, **58**, 155-176.

[30] Hastie, T. and Tibshirani, R. (S original). Leisch, F., Hornik, K. and Ripley, B.D. (R port) (2006) mda: Mixture and flexible discriminant analysis. R package version 0.3-2.

[31] Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.

[32] Huang, K. and Murphy, R.F. (2004) Boosting Accuracy of Automated Classification of Fluorescence Microscope Images for Location Proteomics. *BMC Bioinformatics*, **5**.

[33] Hubert, M., Rousseeuw, P.J. and Van Aelst, S. (2005) Multivariate Outlier Detection and Robustness. In Roa, C.R., Wegman, E.J. and Salka, J.L. (editors) *Handbook of Statistics 24: Data Mining and Data Visualization*. Elsevier B.V.

[34] Kelly, M.G., Hand, D.J. and Adams, N.M. (1999) The Impact of Changing Populations on Classifier Performance. *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[35] Kenny, B.A., Bushfield, M., Parry-Smith, D.J., Fogarty S. and Treherne, J.M. (1998) The Application of High-Throughput Screening to Novel Lead Discovery. *Prog. Drug Res.*, **51**, 245-69.

[36] Kuncheva, L.I. (2004) *Combining Pattern Classifiers: Methods and Algorithms*, New Jersey: John Wiley & Sons.

[37] Liaw, A. and Wiener, M. (2002) Classification and Regression by randomForest. *R News*, **2**, 18-22.

[38] Loo, L., Wu, L.F. and Altschuler, S.J. (2007) Image-based multivariate profiling of drug responses from single cells, *Nature Methods*, **4**, 445-453.

[39] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis*, London: Academic Press Limited.

[40] McLachlan, G.J. (1975) Iterative Reclassification Procedure for Constructing an Asymptotically Optimal Rule of Allocation in Discriminant Analysis. Journal of the American Statistical Association, **70**, 365-369.

[41] McLachlan, G.J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*, New York: John Wiley & Sons.

[42] Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (editors) (1994) *Machine Learning, Neural and Statistical Classification*,
URL: http://www.maths.leeds.ac.uk/~charles/statlog/index.html

[43] Mills, L. (2004) Analysis of High Content Cell Biology Measurements. *MSc Dissertation*, Department of Probability and Statistics, University of Sheffield.

[44] Molina, R., Pérez de la Blanca, N. and Taylor, C.C. (1994) Modern Statistical Techniques. In Michie, D. Spiegelhalter, D. J. and Taylor, C. C. (editors) *Machine Learning, Neural and Statistical Classification*,
URL: http://www.maths.leeds.ac.uk/~charles/statlog/index.html

[45] O'Brien, P.J., Irwin, W., Diaz, D., Howard-Cofield, E., Krejsa, C.M., Slaughter, M.R., Gao, B., Kaludercic, N., Angeline, A., Bernardi, P., Brain, P. and Hougham, C. (2006) High concordance of drug-induced human hepatotoxicity with in vitro cytotoxicity measured in a novel cell-based model using high content screening. *Arch. Toxicol.*, **80**, 580-604.

[46] Pearlman, Z.E., Slack, M.D., Feng, Y., Mitchison, T.J., Wu, L.F. and Altschuler, S.J. (2004a) Multidimensional Drug Profiling by Automated Microscopy. *Science*, 306, 1194-1198.

[47] Pearlman, Z.E., Slack, M.D., Feng, Y., Mitchison, T.J., Wu, L.F. and Altschuler, S.J. (2004b) Multidimensional Drug Profiling by Automated Microscopy (Supporting Online Material),
URL: http://www.sciencemag.org/cgi/data/306/5699/1194/DC1/2.

[48] R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
URL: http://www.R-project.org

[49] Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*, Cambridge University Press

[50] Ripley, B.D. (2007) tree: Classification and regression trees. R package version 1.0-26.

[51] Rousseeuw, P.J. (1984) Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871-881.

[52] Rousseeuw, P.J. and Van Driessen, K. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212-223.

[53] Shi, T. and Horvath, S. (2006) Unsupervised Learning with Random Forest Predictors. *Journal of Computational and Graphical Statistics*, **15**, 118-138.

[54] Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. and Feuston, B.P. (2003) Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modelling. *J. Chem. Inf. Comput. Sci.*, **43**, 1947-1958.

[55] Venables, W.N. and Ripley B.D. (2002) *Modern Applied Statistics with S. Fourth Edition*. New York: Springer.

[56] Weiss, S.M. and Kapouleas, I. (1989) An Empirical Comparison of Pattern Recognition, Neural Nets, and Machine Learning Classification Methods. *Proceedings of the International Joint Conference on Artificial Intelligence*.

[57] Willet, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today*, **11**, 1046-1053.

[58] Young, I.T. (1977) Proof without Prejudice: Use of the Kolmogorov-Smirnov Test for the Analysis of Histograms from Flow Systems and Other Sources. *J. Histochem. Cytochem*., **25**, 935-941.

[59] Zhou, X., Chen, X., Liu, K.L., Lyman, S., King, R., and Wong, S. (2007) Time-Lapse Cell Cycle Quantitative Data Analysis Using Gaussian Mixture Models. In *Life Science Data Mining*, Singapore: World Scientific.

[60] Zhou, X. and Wong, S.T.C. (2006) Informatics Challenges of High-Throughput Microscopy. *IEEE Signal Processing Magazine*, **23**, 63-72.