Chapter 1
Introduction

> 'Into thy presence we come, not by the works we have done, but by the grace and the grace alone, into thy presence we come.' (Benjamin Dube, 2007)

The rate of growth in data volumes stored by organisations continues to grow at a phenomenal rate. For many organisations, the amount of data stored in the data warehouses is in the region of many terabytes. At the extreme end, there are organizations whose data warehouse sizes are in the region of 50 terabytes or more. Data warehouses and business intelligence tools for data analysis have become a necessity in many organizations due to the ever increasing competitive nature of doing business in the information age.

Real-time data warehousing is not uncommon. Given the large volumes of data that are collected by business, government, non-government and scientific research organizations, a major challenge for data mining researchers and practitioners is how to select sufficient amounts of data for analysis, in order to meet the objectives of a data mining task. As second major challenge is design of fast methods of data analysis. The central argument of this thesis is that there is a need to employ methods of dataset selection that provide as much information as possible to the data mining algorithms. The dataset selection methods need to be coupled with fast and reliable methods of data analysis for the creation of reliable data mining models. The thesis concentrates on predictive data mining algorithms for classification tasks. Methods for feature selection, dataset selection, and model construction, are proposed and studied. It is argued and demonstrated that these methods result in the construction of reliable, high performance classification models for data mining from very large datasets.

1.1 Motivation for the research

Data mining is commonly defined as a collection of methods for the analysis of observational data (Hand et al, 2001; Smyth, 2001). The methods used in data

mining for purposes of data analysis originate mainly from the fields of Computer Science, Statistics and Operations Research. Several researchers (e.g. Giudici, 2003; Smyth, 2001; Hand, 1999) have observed that data mining lies at the interface between Computer Science and Statistics. More recently, Olafsson et al (2008) have discussed the contributions of Operations Research to data mining. Formally, Hand et al (2001) have defined data mining as follows.

> 'Data mining is the analysis of (often large) observational datasets, to find unsuspected relationships, and to summaries the data in novel ways that are both understandable and useful to the data owner.'

From the Computer Science perspective, the main contribution to the field of data mining has been algorithms from the area of machine learning. The algorithms that originate from machine learning are employed in the implementation of local and global models from observational data (Giudici, 2003; Smyth, 2001). From Statistics, the parent field for data analysis, the main contribution has been the large body of knowledge on the summarisation of data that is generated by stochastic processes, estimation of descriptive and predictive models for stochastic processes, and the evaluation of the estimated models (Giudici, 2003; Smyth, 2001). From Operations Research the most distinctive contribution has been optimisation methods that can be employed in various modeling activities and especially in the selection of the best model from a set of possible models (Olafsson et al, 2008; Osei-Bryson, 2004, 2007, 2008; Fu et al, 2003, 2006).

The research for this thesis was directed at the selection of training data from large datasets for purposes of aggregate modeling. Aggregate modeling is concerned with the creation of many base models which are then combined into one aggregate model. From a computational perspective, it can be argued that the processing time complexity of most machine learning algorithms employed in data mining is typically non-linear. This property of machine learning algorithms places a limit on the amount of data that can be processed in order to provide results within a reasonable and acceptable amount of time. The time complexity of machine learning algorithms for data mining is not the only issue to consider when faced with large data volumes. From a statistical perspective, it is not desirable to use a very large amount of data in the process of estimating one model. In the past there have been several negative comments, especially originating from the Statistics community, directed at various

research directions in data mining. In 1998, Hand (1998) made the following observation.

> '..the term data mining is … synonymous with data dredging.. and has been used to describe the process of trawling through data in the hope of identifying patterns. It has a derogatory connotation because a sufficiently exhaustive search will certainly throw up some patterns of some kind … the object of data analysis is not to model the fleeting random patterns of the moment, but to model the underlying structures which give rise to consistent and replicable patterns. ..the term data mining conveys the sense of naïve hope vainly struggling against the cold realities of chance.'

Both the computational perspective and the statistical perspective as discussed above, point to the need for data reduction. It is the author's opinion that research efforts should be directed towards the study of methods for the selection of relevant data that can be used to create models that provide a high level of predictive performance.

The problem that the work reported in this thesis aims to solve is the design of methods for training dataset selection, for purposes of creating many base models which can be combined into one aggregate model. Such an aggregate model should provide a higher level of predictive performance compared to a single model created from a single training dataset. This approach should lead to the usage of significantly large amounts of data while at the same time avoiding the computational and statistical problems highlighted above. The idea of using aggregate models is not new. As far back as 1996, Breiman (1996) proposed bootstrap aggregation as a method of improving predictive accuracy for models constructed from small datasets. At the present time, there are many research efforts directed at the design of aggregate predictive models.

1.2 Current debates and practices in data mining from large datasets

One approach that has been investigated by researchers in predictive data mining is the use of very large training datasets obtained from very large datasets. Training datasets of several millions records have been processed using very powerful machines (Chawla et al, 2001; Hall et al, 2000). The rationale behind this approach is

that when very large amounts of data are processed, then as much as possible of the information gathered about a subject area is incorporated in the model construction. An obvious disadvantage is that the model construction process takes a very long time. A second and more serious disadvantage may be explained through statistical theory. Smyth (2001), Hand et al (2001), and Hand (1998) have cautioned that when training datasets are very large it becomes very difficult to distinguish between noise and real structure in the data.

Another explanation of this disadvantage comes from the machine learning literature. Dietterich (1995) has observed that for classification problems, a predictive model which has a very high level of training accuracy is not necessarily reliable when put to practical use. The main purpose of predictive modeling is to process data in order to find relationships that can be generalized. If an inductive algorithm is used to create a predictive model from a very large amount of data it will minimize the training error. However, there is a very high risk that it will fit the predictive model to the noise in the training data by memorizing peculiarities of the training data rather than finding a general predictive rule. This phenomenon is called *overfitting* (Smyth, 2001; Dietterich, 1995). Prediction models based on very large amounts of data should therefore be treated with caution.

A second approach to predictive data mining from large datasets is to take a single sample from a very large dataset and use it for model construction. Additional samples are then taken for validation and testing (Domingos, 2001; Kohavi et al, 2004; Provost et al, 1999; John & Langley, 1996). This approach has also received much attention from theoretical research in statistical pattern recognition and machine learning, for example, Valiant (1984). The main advantage of this approach is that the training sample is typically much smaller than the large dataset, and so, is much faster to process. An obvious disadvantage is that the bulk of the data is discarded and only a small fraction of the data is used for making decisions about feature selection, model structure and model performance. A second disadvantage is that sampling results in stochasticity. If another random sample were to be taken, the selected features, model structure and measured performance may be significantly different.

A third approach to predictive data mining from large datasets is to partition a large dataset, construct a predictive model based on each partition and then combine the different models into one aggregate model (Chawla et al, 2001; Hall et al, 2000;

Chan & Stolfo, 1998). One obvious advantage of this approach is that partitioning attempts to use as much of the available data as possible. Several researchers who have studied aggregate modeling from large datasets (e.g. Chawla et al, 2001; Hall et al, 2000; Chan & Stolfo, 1998) have argued that the performance of an aggregate model normally exceeds that of a single model constructed from a single large training sample. On the other hand, other researchers (e.g. Hall et al, 2000; Ali & Pazzani, 1996) have argued that there are various domains where partitioning does not result in any performance gains and may in fact result in loss of accuracy.

The use of aggregate models has been studied by many researchers (e.g. Osei-Bryson et al, 2008; Sun & Li, 2008; Ooi et al, 2007; Neagu et al, 2006; Kim et al, 2002; Chan & Stolfo, 1998; Breiman, 1996; Krogh & Veldelsby, 1995; Kwok & Carter, 1990) even though these studies have not always been in the context of very large datasets. A large body of literature and evidence exists to support the claims that aggregate modeling often leads to improved predictive performance. Given the foregoing observations, it is the author's opinion that studies in dataset selection from large datasets should be directed towards improving the predictive accuracy of aggregate models.

1.3 Scope of the research

The title of this thesis makes reference to the term, *predictive data mining*. It is therefore important for the author to highlight the difference between predictive and non-predictive data mining.

Data mining tasks may be broadly divided into four categories, namely: exploratory data analysis (EDA), local methods for pattern detection and rule extraction, descriptive modeling, and predictive modeling (Hand et al, 2001). Exploratory data analysis is concerned with the exploration of data without any prior clearly articulated idea of what one is looking for, or any plan of what output needs to be generated. Pattern detection and rule discovery activities are concerned with the identification of regions of the instance space whose characteristics significantly differ from those of the other regions (e.g. association rule mining) or locating patterns of interest in data as is done in text mining (Hand et al, 2001). The objective of descriptive modeling is to create a model that describes the data or the process that generates the data (Hand et al, 2001). Examples of this include density estimation (estimation of the

overall probability distribution), cluster analysis (identification of naturally occurring groups in the data), segmentation (division of data into groups based on specified criteria) and, dependency modeling (description of the relationship between variables). For predictive modeling, the purpose is to create a model that may be used for the prediction of the value of the dependent variable, given the values of the independent (predictor) variables.

The term *predictive data mining* refers to data mining methods that create predictive models (Hand et al, 2001). Predictive models may be constructed to predict the values of a quantitative variable as in regression or to predict the values of a qualitative variable as in classification. The research reported in this thesis is primarily concerned with classification problems. As discussed in the last section, there is a large body of evidence to support the claim that aggregate modeling has the potential to improve classification performance. The scope of the research reported in this thesis is directed at classification methods that employ aggregate modeling.

In the data mining literature, Giudici (2003) has made a distinction between computational data mining and statistical data mining. The distinguishing characteristic between computational and statistical data mining is that while statistical data mining methods assume a specific probability distribution for the process that generates the data, computational data mining methods make no specific assumptions about the probability distribution for the data generating process. However for computational data mining and machine learning, there is the (not always stated) assumption that the data generating process is governed by a fixed but unknown probability distribution (Mitchell, 1997). The research reported in this thesis is aimed at computational data mining.

1.4 The claims of the thesis

The central argument of this thesis is that it is possible for predictive data mining to systematically select many dataset samples and employ different approaches (different from current practice) to feature selection, training dataset selection, and model construction. When a large amount of information in the large dataset is utilised in the modeling process, the resulting models should have a high level of predictive performance and should be reliable.

Ngwenyama (2007) has identified seven categories of scientific research claims. The first four claims identified by Ngwenyama (2007) are: (1) a scientific problem that has been solved (2) a general contribution to science (3) extension of a body of knowledge and (4) appropriateness of the research methodology. Ngwenyama (2007) has used the argumentation model by the philosopher Toulmin (Toulmin et al, 1979; Toulmin, 1958) to analyse the four categories of scientific research claims. In Toulmin's argumentation model (Toulmin et al, 1979; Toulmin, 1958) *claims* are supported by *data* (observations / evidence) and *warrants*. The *data* (observations / evidence) are the grounds on which the claim stands. *Warrants* consist of general rules of inference and existing theories that serve as bridges or connections between the *data* (observations / evidence) and the *claims*. *Warrants* are supported by *backings* which are the known authoritative sources from which the *warrants* are drawn. The claims of this thesis are presented in terms of Ngwenyama's (2007) categorisation and Toulmin's (1958) argumentation model. The scientific problem that has been solved and the general contributions to science are presented in this section. The extensions to the body of knowledge and the research paradigm are presented in the next two sections.

The first *claim* that is made in this thesis is that aggregate classification models based on One-versus-All (OVA) modeling (Ooi et al, 2007; Rifkin & Klautau, 2004) and positive-Versus-negative (pVn) modeling can be used to increase the amount of relevant data in the training datasets. Increasing training data through OVA and pVn modeling results in improved predictive performance compared to the use of a single model. OVA modeling involves the decomposition of a *k*-class prediction task into *k* 2-class prediction tasks. pVn modeling involves the decomposition of a *k*-class prediction task into *j (j<k)* prediction tasks. OVA and pVn aggregate models differ from the aggregate models commonly discussed in the literature (e.g. Osei-Bryson et al, 2008; Kim et al, 2002; Chan & Stolfo, 1998; Breiman, 1996; Krogh & Veldelsby, 1995; Kwok & Carter, 1990; Hansen & Salamon, 1990). Firstly, the aggregate models discussed in the literature cited above do not employ problem decomposition. Secondly, the training datasets used for the base models that constitute such aggregate models generally re-use the small amount of available data. The methods proposed in this thesis for the implementation of OVA and pVn aggregate models do not re-use training data, but rather, use a different training dataset for each base model. These methods result in high coverage of the instance space while at the same time avoiding the problems of data dredging and overfitting. Traditionally, data

dredging and overfitting are associated with the usage of large training datasets for single models. High coverage of the instance space provides more information for the prediction task which in turn results in high predictive performance.

The second *claim* of this thesis is that the performance of aggregate models can be improved when the training samples for the base models are purposefully designed to reduce the bias and variance components of the prediction error. The bias component of the prediction error reflects the level of error in the estimation process of the model. The variance component reflects the sensitivity of the model to the training sample used to estimate the model (Friedman, 1997; Geman et al, 1992).

The *warrants* and *backing* for the first and second *claim* are as follows: Based on statistical theory a random / stochastic process can be studied using many small samples of the data generated by the process in order to establish the underlying structure of that process. Secondly, theories have been formulated in machine learning and statistical pattern recognition to explain how prediction errors arise. Based on these theories, it is possible to select training datasets in such a way that the chances of error are significantly reduced. There have been various research efforts that use several samples in model construction and feature selection. Breiman (1996) has studied the use of many bootstrap samples from small datasets to implement classifier committees. Freund and Schapire (1997) have studied boosting through the sequential creation of many small training samples, where each successive training sample consists of a larger number of training instances that are difficult to predict correctly. Studies have been reported on dataset selection methods which are guided by information on the characteristics of the instance space (Chan & Stolfo, 1998; Kubat & Matwin, 1997). All the above studies have demonstrated that purposeful training dataset selection for base models can result in major improvements in the predictive performance of aggregate models.

The third *claim* of this thesis is that the use of many (relatively) small samples to measure correlations between the variables for the prediction task leads to a more reliable selection of the relevant features for the prediction task. The fourth *claim* of this thesis is that, when the domain-specific definitions of the strength of association between variables are incorporated into the feature selection decisions, good subsets of predictive features will be selected for the prediction task.

The *warrants* and *backing* for the third and fourth *claims* are as follows. Statistical theory tells that, when the correlation between two random variables is measured using one sample then if the sample is small, a small or large correlation coefficient could be purely due to chance (Smyth, 2001). On the other hand if the sample is large, a small correlation coefficient may appear to be statistically significant even though it has no practical significance (Cohen, 1988). For purposes of measuring the correlations between the predictive variables and the class variable, Bi et al (2003) have studied the use of many bootstrap samples for micro-array datasets, in order to achieve reliable feature subset selection. Even though the studies by Bi et al (2003) have been conducted on small datasets, the results of their studies indicate that there are benefits in using many small samples to establish feature relevance for prediction tasks. Research has been conducted on the incorporation of user preferences in algorithms for predictive modeling. Osei-Bryson (2004) has proposed the incorporation of user preferences in decision tree selection. Ooi et al (2007) and Yu and Liu (2004) have proposed the incorporation of user-specified preferences in feature selection methods. The foregoing observations provide motivation for the incorporation of domain-specific definitions of feature relevance into feature selection algorithms.

The fifth and final *claim* of this thesis is that research into aggregate model construction methods using different methods of sample composition and feature selection should lead to useful theories for the improvement of aggregate model performance. When the data available for model construction is small, as was typically the case in the past, statisticians invented effective methods of model construction, validation and testing (Mitchell, 1997; Cohen, 1995). Bootstrap sampling for example, is useful for purposes of creating several large samples which have the same statistical properties as the small sample from which they are generated (Cohen, 1995).

In this thesis the author further argues that, since at the present time very large amounts of data are available for data mining, it is productive to investigate (new) ways of predictive model construction coupled with new ways of dataset selection. It is the author's opinion that the following issues have not been sufficiently studied by researchers:

(1) The use of many samples drawn from very large datasets for purposes of feature selection.

(2) The use of sampling in conjunction with partitioning for purposes of dataset selection and aggregate model construction.
(3) The design of training dataset samples aimed at reducing bias and variance in the prediction error without the need to re-use training data.

The author further argues that when very large amounts of data are available, data mining researchers have at their disposal a great opportunity to conduct empirical studies of the factors, and the relationships between the factors that affect various aspects of predictive model design and construction. In the data mining literature, there seems to be a scarcity of clearly articulated theoretical models based on empirical studies that can help to explain the relationships between the factors that determine: (1) the quality of selected feature subsets, (2) the quality of selected dataset samples and, (3) the predictive performance of aggregate models. It should be pointed out however that for aggregate model construction, several researchers have conducted studies on various factors that affect aggregate model performance in the context of small datasets. Examples of these studies are Kwok and Carter (1990), Ali and Pazzani (1996), Breiman (1996), and Ho (1998).

The investigations of this thesis were directed at dataset selection methods from large datasets for purposes of aggregate model implementation. The main research question for the thesis was as follows:

What methods of dataset selection can be used to obtain as much information as possible from large datasets while at the same time using training datasets of small sizes to create predictive models that have a high level of predictive performance?

The investigation of the answers to the above question was conducted using the design science research paradigm which is described briefly in the following section and in detail in chapter 4. The design science research paradigm enabled the author to generate experimental *evidence* (*data*) to support the *claims* presented in this section.

1.5 Research paradigm

The research paradigm used for this research is design science research as described by March and Smith (1995), Hevner et al (2004), Vaishnavi and Kuechler

(2004/5), and Manson (2006). Design science research involves two distinct steps. In the first step, an artifact is created. In the second step, an analysis of the usage and performance of the artifact is conducted. The purpose of the analysis is to understand, explain, and possibly improve on one or more aspects of the artifact (Vaishnavi & Kuechler, 2004/5).

In the context of information systems, artifacts may be models (abstractions and representations), methods (algorithms and practices) and instantiations (implemented and prototype systems) (Hevner et al, 2004). Manson (2006) has summarised these views by observing that design science research is a process of using knowledge to design and create useful artifacts, and then using rigorous methods to analyse why, or why not, a particular artifact is effective. Scientific research is about generating knowledge. A design science research effort should therefore make a contribution to the knowledge base of the field. More specifically, the contributions of design science research could be:
(1) Constructs. These are the components of the conceptual vocabulary of the domain.
(2) Models. These are propositions expressing the relationships between the constructs / concepts of the research domain.
(3) Methods. This is the 'how-to' knowledge. It is specified in the form of steps used to perform a given task.
(4) Instantiations. This is the operationalisation of the constructs, models and methods to demonstrate that the models and methods can be implemented in a working system.
(5) Better theories.

Design science research was found to be appropriate for this thesis because the central argument is based on the development of methods for feature and training dataset selection as well as the design and creation of predictive models.

1.6 Research contributions

It was stated in the last section that design science research should make a contribution to the knowledge base of the field. The *claims* of the research contributions of this thesis to the knowledge base of predictive data mining are summarised in this section in terms of the expectations of design science research

outputs. Two additional components in Toulmin's (1958) argumentation model are *qualifiers* and *rebuttals*. *Qualifiers* are used to limit the strength of a *claim* and *rebuttals* provide an elaboration for the *qualifiers*. A detailed discussion of the *claims* of the research contributions and, the *qualifiers* and *rebuttals* identified by the author are presented in chapter 11 of this thesis.

1.6.1 Methods and instantiations

Methods for feature selection from large datasets were studied. The studies involved testing methods of reliable feature selection that involve the use of robust measures of correlation, the use of many samples to measure correlations, and the use of statistical tests, such as the t-test and fake variables, for the validation of selected features. Arising from these studies, recommendations are given in this thesis on how to conduct reliable ranking of predictive features when large datasets are available.

A new search algorithm for feature subset selection is proposed. This algorithm uses the domain-specific knowledge of the meanings of the terms *strong correlation* and *weak correlation* in order to select the best subset of features for a list of ranked features. It is claimed in this thesis that the proposed method makes better decisions compared to two feature subset selection algorithms proposed in the literature, namely: Correlation-based Feature Selection (Hall, 1999, 2000) and Differential Prioritisation (Ooi et al, 2007).

The implementation of One-versus-All (OVA) aggregate classification models in the presence of large datasets was studied. A new method of determining composition of the training dataset for each base model is proposed. A new method of aggregate model implementation, named positive-Vs-negative (pVn) classification is proposed. An algorithm is proposed for the determination of the classes to be included in each base model. A method of determining the sample composition for the training dataset of each base pVn model is proposed. An algorithm for combining base model predictions and resolving conflicting predictions is proposed.

1.6.2 Constructs, models and better theories

Theoretical models are propositions expressing the relationships between the constructs / concepts of the research domain. For feature selection, a model was created to combine the work of various researchers. This model was extended by the author to explain how the definition of feature relevance, the methods used to measure correlations, and the number of dataset samples used, all combine to affect the quality of selected feature subsets. For aggregate model construction, the work of Ho (1998), Freund and Schapire (1997), Ali and Pazzani (1996), Breiman (1996), Kwok and Carter (1990), and Hansen and Salamon (1990), was used as a basis to construct a theoretical model that explains the relationships between the factors that affect aggregate model performance. This model was extended by the author to explain how dataset partitioning methods, learning task complexity, overlap between learning tasks, overlap between training instances, and the quality of the selected features affect the performance of aggregate models. The experimental results were used to demonstrate the relationships between the various factors that affect predictive model performance.

1.7 Overview of the thesis

Chapters 2, 3 and 4 provide the background to the research. Chapter 2 provides a discussion of the dataset selection problem for predictive data mining. The chapter provides a background to this problem, giving examples of several application domains where very large datasets are to be found. A review of literature on current methods of selecting training set data from very large datasets for purposes of classifier construction is given. Theoretical methods as well as empirical methods are discussed. The discussion of this chapter also covers single model and aggregate model construction, since the problems of dataset selection and model construction are related. Chapter 3 provides an overview of the feature selection problem for classification tasks in predictive data mining. A review of the available methods for feature selection from small datasets is provided. The weaknesses of these methods are also highlighted. Robust measures of correlation are discussed briefly. In chapter 4, the research questions, the central argument of the thesis, and the research paradigm and research methods, are discussed in detail.

Chapters 5, 6, 7, 8 and 9 provide the details of the empirical studies that were conducted. Further details of the experimental results are provided in the appendices. In chapter 5, the experimental results on feature subset selection are presented. The experimental results demonstrate that the use of many samples results in more reliable feature selection. The results also demonstrate that the use of domain-specific knowledge will lead to better feature subset selection when heuristic subset feature selection is employed. Based on the experimental results of chapter 5 and the existing literature, a theoretical model for the factors that influence the quality of feature selection is proposed in chapter 10.

Chapter 6 provides a discussion of the methods that were used in the experiments for aggregate model design, training dataset design and selection, partitioning and sampling, and base model design and aggregation. The studies to evaluate the performance of the proposed methods are presented in chapters 7, 8 and 9.

Chapter 7 provides a discussion of the empirical study of the use of OVA modeling. It is demonstrated that the use of OVA base models where each base model uses a different training set of the same size as a single model can lead to significant improvements in predictive performance. It is further demonstrated that, by establishing the nature of the instance space and then determining which regions of the instance space to take samples from for each OVA base model, a level of predictive accuracy that is higher than that of a single *k*-class model can be obtained. Based on the experimental results of chapter 7 and the existing literature, a theoretical model for the factors that influence the performance of aggregate models is proposed in chapter 10.

Chapter 8 presents a discussion of the new method of aggregate model implementation called positive-Vs-negative (pVn) classification, as well as the proposed methods for determining the class and sample composition for each pVn base model. Experimental results of the studies to demonstrate the performance of pVn modeling are presented. The experimental results demonstrate that, for the datasets used in the experiments, pVn aggregate modeling provides a high level of predictive accuracy. The experimental results of chapter 8 are used in chapter 10 to enhance the theoretical model for the factors that influence the performance of aggregate models. Chapter 9 provides an in-depth analysis of the OVA and pVn aggregate models operating under different conditions.

Chapter 10 presents the recommendations for dataset selection based on the experimental results for the thesis. Chapters 11 and 12 provide discussions and conclusions for the thesis as well as suggestions for future work. Chapter 11 provides a discussion of the contributions of this thesis to the knowledge base of the field of predictive data mining using aggregate classification models. The discussion of the contributions is presented in terms of the outputs of design science research. Chapter 12 provides conclusions for the thesis as well as suggestions for future work.

Chapter 2

Dataset Selection and Modeling from Large Datasets

This chapter provides a discussion of the dataset selection problem for predictive data mining. The discussion provides a background to the dataset selection problem, giving examples of application domains where very large datasets are to be found. A review of the literature on current methods of selecting training set data from very large datasets for purposes of classification modeling is given. Theoretical methods as well as empirical methods are discussed. Since dataset selection and model construction are intimately linked, the discussion in this chapter also addresses single model and aggregate model construction. The strengths and shortcomings of the theoretical and empirical methods are highlighted. The chapter ends with a discussion of research directions that, in the author's opinion, are useful to pursue in order to effectively answer the research question which was presented in chapter 1.

This chapter is organised as follows: Section 2.1 provides motivation for the dataset selection problem with examples of four application domains for data mining. Sections 2.2 and 2.3 respectively introduce the classification modeling problem and dataset selection problem. Sections 2.4 and 2.5 respectively provide a review of theoretically based and empirically based methods for training dataset selection for single model construction. Section 2.6 gives a discussion of existing methods for training dataset selection for multiple model construction. Conceptual views of classification modeling and the sources of classification error are respectively discussed in sections 2.7 and 2.8. The limitations of current training dataset selection methods and the proposed methods of training dataset selection are respectively presented in sections 2.9 and 2.10. Section 2.11 concludes the chapter.

2.1 The need for dataset selection

Modern data warehouses store very large volumes of data. In many areas where data mining is applied, very large amounts of data are collected. There are many application areas where data mining from large datasets is applied. These areas

include scientific applications (Fayyad et al, 1996), forensic data mining for purposes of predicting telephone fraud (Hand, 1999), credit card fraud (Chan & Stolfo, 1998), computer network intrusion detection (Lee & Stolfo, 2000), web usage mining for analysing and predicting customer purchases behaviour (Theusinger & Huber, 2000; Kohavi et al, 2004), and customer relationship management (Rygielski et al, 2002; Kohavi et al, 2000; Berry & Linoff, 2000). This section provides examples of application areas where very large datasets for data mining are encountered. Customer Relationship Management (CRM) is discussed in section 2.1.1. Web usage mining and electronic commerce are discussed in section 2.1.2. Forensic data mining is discussed in section 2.1.3. Scientific applications of data mining are discussed in section 2.1.4.

2.1.1 Customer Relationship Management - CRM

Customer Relationship Management (CRM) (Giudici, 2003; Rygielski et al, 2002; Bose, 2002; Berry & Linoff, 2000) is a collection of business activities specifically aimed at maintaining good relationships with the business customers. CRM involves the formulation and implementation of strategies to encourage customer loyalty in order for a business to obtain as much value as possible from the customers. Statistically driven CRM (Giudici, 2003) involves the collection, storage and analysis of data about customer interactions with a business in order to obtain a better understanding of customer behaviour. A better understanding of customer behaviour enables businesses to provide better services and product offerings to the customers (Giudici, 2003; Rygielski et al, 2002; Bose, 2002).

Rygielski et al (2002) have argued that, in order for a business to succeed with CRM, the business needs to capture and analyse massive amounts of customer data, analyse the data and transform the analysis results into actionable information. Rygielski et al (2002) have also argued that the analysis of customer data using predictive data mining, especially to extract rules, is an essential component of CRM for the modern business. The use of electronic commerce has made it much easier to collect massive amounts of data about customer purchasing behaviour in data warehouses. The availability of large volumes of data on customer purchasing activities has given rise to research interest in the area of web usage mining for e-commerce. Typical usage of data mining for CRM includes the analysis of customer

attrition, churn, propensity to purchase and customer lifetime value (Giudici, 2003; Rygielski et al, 2002).

2.1.2 Web usage mining and electronic commerce

For electronic commerce, since data collection is an automated process, data volumes can grow very rapidly. One interesting application area which has emerged for e-commerce data is clickstream analysis (Kohavi et al, 2004; Theusinger & Huber, 2000). Clickstream analysis is used to study user navigation patterns at a website. The study of user navigation patterns at a website can expose structural or usability problems for a website, which in turn provide useful information for improving the website design. Such a study will also identify which click sequences lead to purchases (Theusinger & Huber, 2000). Kohavi et al (2004) have observed that websites that have 30 million page views per day will need to store in the region of 10 billion records of clickstream data each year. Linden et al (2003) have reported that Amazon.com™ conducts electronic trading with more than 29 million customers per month and stocks several million catalogue items at any given time. The collection of large amounts of web navigation and purchases data creates major challenges for clickstream analysis, for e-traders such as Amazon.com™. Web usage mining applications make explicit the fact that it may be practically impossible to process all of the available data for real-life e-commerce applications of data mining.

2.1.3 Forensic data mining

Forensic data mining involves processing large amounts of data in order to identify criminal activities such as credit card fraud (Chan & Stolfo, 1998; Hand, 1999) and computer network intrusion (Lee & Stolfo, 2000). Chan and Stolfo (1998) have reported studies conducted on data for credit card transactions. Chan and Stolfo (1998) have observed that, for the credit card fraud detection domain, there may typically be millions of transactions occurring every day. Hand (1999) has reported that 350 million transactions are recorded annually by UK's largest credit company. Hand (1999) has further discussed the need for real-time data analysis for fraud detection and has argued that, since banking transactions happen all the time, models created, say weeks or months after the fact are useless. There is a need to

constantly create new and up-to-date models. Hand (1999) has further reported that by 1999 AT&T™ was recording 200 million call detail records per day. Phua et al (2005) have reported that descriptive modeling (e.g. cluster analysis), predictive modeling (classification and regression), and pattern detection and rule extraction (e.g. association rules) are all data mining methods that are commonly employed in fraud detection. Scalability of these methods is therefore a serious issue for fraud detection, and dataset selection becomes a necessity.

For modern computer networks large volumes of data are collected and stored in server log files to record all user connections to each server in the network. The users who access the network servers may be authentic users, or may be malicious criminal entities. The data stored in the server log files may be used to create predictive models that are used as network intrusion detection systems (IDS) (Lee & Stolfo, 2000; Stolfo et al, 2000). Lee et al (2000) have observed that the volumes of data stored in server log files are typically huge, as computer networks can experience several million connections on some days due to denial-of-service attacks.

2.1.4 Scientific applications of data mining

Fayyad et al (1996) have presented various case studies of the application of data mining to scientific data. Fayyad et al (1996) have observed that the main challenge for the application of data mining to scientific data that is automatically collected by scientific instruments is that these instruments can easily generate terabytes of data at rates as high as several gigabytes per hour. One interesting example is the Palomar Observatory Sky Survey that was conducted over a period of six years (Fayyad et al, 1996). The data collected consisted of 3TB of image data containing 2 billion sky objects. The basic problem here was to create a survey catalogue recording the (predictive) features of each object with its class: star or galaxy. Fayyad et al (1996) have stated that the problem was solved using decision tree learning with multiple trees, and rule extraction with statistical optimisation.

A second interesting example of the application of data mining to scientific data is the analysis of geoscience data for purposes of earthquake detection. Stolorz and Dean (1996) have discussed the Quakefinder system which detects and measures tectonic activity in the earth's crust by examining satellite data. The Quakefinder system

processes massive datasets on a 256-node Cray™ T3D parallel supercomputer to ensure fast turnaround of results for scientists. It is generally not possible for a predictive data mining algorithm to process all of the data for scientific applications where data is automatically collected by measuring instruments. Supercomputers are however used in order to process as much of the data as possible.

2.2 Classification modeling from very large datasets

Classification modeling is the process of creating a model which predicts the values of a qualitative variable called the class variable. There are two approaches that have been proposed in the literature for the construction of predictive classification models from very large datasets. The first approach to modeling is concerned with constructing one model using a single sample whose performance is estimated to be as good as that of a model that would be obtained from the whole dataset. The second approach to modeling is concerned with the partitioning of a large dataset into many small subsets which can be efficiently processed, possibly in parallel, creating a base model from each subset of data, and then combining the base models into an aggregate model. The predictive performance of the aggregate model is expected to be at least as good and in several cases superior to that of a single model. Aggregate model construction methods are generally concerned with increasing accuracy compared with the use of a single predictive model. Several methods for aggregate model construction are directly concerned with the parallel processing of the dataset using massively parallel machines in order to ensure that all the data, or as much as of the data as possible, is used in model construction. This section provides a formal definition of the classification problem and the terminology for classification modeling. Methods for single model construction from large datasets as well as the methods for aggregate model construction from small datasets and from very large datasets are discussed. The terminology for classification modelling is presented in section 2.2.1. The classification modeling problem is discussed in section 2.2.2. Single model and aggregate construction are respectively discussed in sections 2.2.3 and 2.2.4. Serial and parallel aggregation, and model testing are respectively discussed in sections 2.2.5 and 2.2.6.

2.2.1 Terminology for classification modeling

A dataset for predictive modeling may be described as an *N x (d+1)* data matrix. In the data matrix each row represents *(d+1)* measurements on a real-life object so that the *N* rows in the data matrix represent N real-life objects (Hand et al, 2001). The rows of the data matrix are commonly called *patterns* (Liu & Motoda, 1998), *examples* (Mitchell, 1997), *instances* or *cases* (Hand et al, 2001). The columns of the data matrix are commonly called *variables*, *features* or *attributes* (Hand et al, 2001; Mitchell, 1997). For predictive modeling the first *d* columns are called the *predictor variables* or *features* and the $(d+1)^{st}$ column is called the *predicted variable*. Specific to classification modeling, the *predicted variable* is called the *class variable*. The *d*-dimensional space defined by the *variables* is commonly called the *measurement space* (Hand, 1997) or *instance space* (Mitchell, 1997). Within this *d*-dimensional space, each object (instance) corresponds to one point and the object has an associated class label specified by the $(d+1)^{st}$ column (*class variable*). In this thesis the term *instance* is used to refer to the objects, the term *feature* is used to refer to a predictor variable, the term *class variable* has the usual meaning and, the term *variable* is used to refer to a random variable in the generic sense. The term *instance space* is used to refer to the *d*-dimensional space defined by the predictor variables.

The variables for the data matrix may be quantitative or qualitative (Giudici, 2003; Hand et al, 2001). A *quantitative variable* has numeric values that are either discrete or continuous. The values of a *quantitative discrete variable* have a finite number of levels. The values of a *quantitative continuous variable* come from the domain of real numbers. A *qualitative variable* has values that are either nominal or ordinal. The values of a *qualitative nominal variable* have a finite number of categories which do not possess an ordering. The values of a *qualitative ordinal variable* have a finite number of categories which possess an ordering (Giudici, 2003; Hand et al, 2001). The term *categorical variable* is also used in the literature to refer to a *qualitative variable* (Giudici, 2003; Hand et al, 2001). The terms *quantitative variable*, *quantitative feature*, *qualitative variable,* and *qualitative feature* were adopted for this thesis.

In the literature on machine learning and data mining, various names are used to refer to predictive models for classification. A predictive classification model that is created from a single training sample using a single classification algorithm is called a classifier. When several classifiers are created from one or more training datasets

for purposes of combining them into one predictive model, these classifiers are called *base classifiers* or *base models*. A classifier that is created by combining several *base classifiers* is referred to using various terms. Breiman (1996) has used the term *aggregation* to refer to the process of combining classifier predictions, and the term *aggregate predictor* to refer to the model that results when several classifiers are combined into one model. The terms *ensemble* and *ensemble classifier* have been used by Hansen and Salamon (1990) to refer to combinations of artificial neural networks and are very commonly used in the current machine learning and data mining literature. The term *committee of classifiers,* originating from work on query by committee, has also been used to refer to *ensemble classifiers*. The term *multiple model* is also commonly used (Sun & Li, 2008; Ali & Pazzani, 1996; Kwok & Carter, 1990). In this thesis a decision was made to use the terms *single model, base model,* and *aggregate model.* The term *single model* is used to refer to a classifier created by one algorithm from a single training dataset. The term *aggregate model* is used to refer to a classification model that is created by combining several *base models*. The terms *single model* and *aggregate model* were chosen as it was felt that they capture more precisely, and clearly contrast the structures of the models to which they refer.

In the literature on ensemble classification the terms *complementary classifiers* and *complementarity* are used to refer to base classifiers which make uncorrelated errors (e.g. Martínez-Muñoz et al, 2009). Base model *diversity* is a property that is related to *complementarity*. The term *syntactic diversity* is also used in the literature to refer to base model *diversity* (e.g. Ho, 1998; Ali & Pazzani, 1996; Krogh & Vedelsby, 1995; Kwok & Carter, 1990; Hansen & Salamon, 1990). S*yntactic diversity* refers to the level of structural differences between the base models that constitute an aggregate model. Martínez-Muñoz et al (2009) have observed that base model *diversity* is a necessary but not sufficient condition for *complementarity*. The term *syntactic diversity* is used in this thesis to refer to base model *diversity*. The term *competence* is used in the literature (e.g. Ali & Pazzani, 1996) to refer to the high predictive performance or high predictive expertise of base models. The terms *competence* and *high expertise* are used synonymously in this thesis.

In machine learning literature, the terms *generalisation error* and *generalisation accuracy* are used to refer to the error and accuracy rates of a classifier on data that was not used in training the classifier (Mitchell, 1997). In statistics and data mining literature the terms *prediction error* and *prediction accuracy* are used to refer to the error and accuracy of a predictive model. In this thesis the terms *prediction error* and

prediction accuracy were adopted. The term *predictive performance* is used to generally refer to various measures of performance including *prediction error* and *prediction accuracy*. Performance measures for classification models are presented in chapter 4.

The term *bias* appears in machine learning and statistics literature with different meanings. In statistics literature the term *bias* refers to *estimation bias* which is the error in the estimation of a parameter or a model (Mitchell, 1997). In machine learning literature the term *bias* has been adopted with the same meaning as used in statistics (Mitchell, 1997; Geman et al, 1992). In machine learning the terms *inductive bias* and *preference bias* refer to the set of methods used by an inductive algorithm to select a hypothesis (model) from the set of all possible hypotheses (models) in the hypothesis space (model space) (Mitchell, 1997). In this thesis the term *bias* is used with the statistical meaning and the term *inductive bias* is used with the machine learning meaning. The term *search bias* is used to refer to the preferences of a heuristic search procedure.

2.2.2 The classification modeling problem

This research is specifically concerned with classification modeling. Classification modeling is the process of creating a model to be used for the prediction of the values of a qualitative variable, given the values of the predictive features. For applied data mining, classification modeling is part of a whole process which involves business understanding, data understanding and preparation, model creation, model assessment and deployment. The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a process model that has been widely adopted for applied data mining (Shearer, 2000). CRISP-DM provides recommendations for the phases to be conducted for data mining projects. Within CRISP-DM the two phases that are directly related to predictive modeling are *data preparation* and *modeling*. For predictive classification modeling, these two phases involve (among others) the following activities: (1) data selection (2) data construction (e.g. creation of the class variable) (3) feature selection (4) model construction (5) estimation of model performance (Shearer, 2000).

It has been illustrated by the examples of the last section that for many application areas, data already exists in large quantities. Data selection is concerned with the

selection of instances and features that have some relevance to the prediction task. Feature selection is concerned with the selection of the most useful features for the prediction task. Classification modeling often requires the construction of a class variable using information derived from other variables based on the objectives of the classification task. Classification modeling involves the estimation of a mapping m (or hypothesis h) from an instance $\boldsymbol{x} = (x_1,...,x_d)$ in the *d*-dimensional instance space to the values of the class variable which consists of classes $\{c_1,...,c_k\}$ (Hand et al, 2001). The two conceptual views of classification are discussed later in this chapter.

2.2.3 Single model construction

Methods for single model construction from large datasets are motivated by the learning curve. Several researchers have argued that the empirical estimation of training and predictive accuracy achievable from a given large dataset and a given learning algorithm may be done using learning curves (Provost et al, 1999; John & Langley 1996; Catlett 1991). A learning curve shows the relationship between sample size (x axis) and the accuracy of the model (y axis) produced by an inductive algorithm. Learning curves typically have three sections as shown in figure 2.1. The leftmost section has a steep slope, the middle section has a more gentle slope, while the rightmost section is a plateau (Provost et al, 1999; Catlett 1991). These three properties of the learning curve have been used as justification that a single model constructed from a large sample should provide a sufficient level of predictive accuracy (Provost at al, 1999; John & Langley, 1996; Catlett, 1991).

John and Langley (1996), Provost et al (1999) and others have conducted empirical studies and devised methods for establishing the sample size $n_{\min}$ needed to obtain maximum accuracy for a given dataset and algorithm. Extrapolation of learning curves (ELC) is one method that has been used to fit learning curves (Frey & Fisher 1999). For ELC, training sets of increasing size are used to fit a parametric learning curve, which is an estimate of the algorithm's accuracy as a function of training set size.

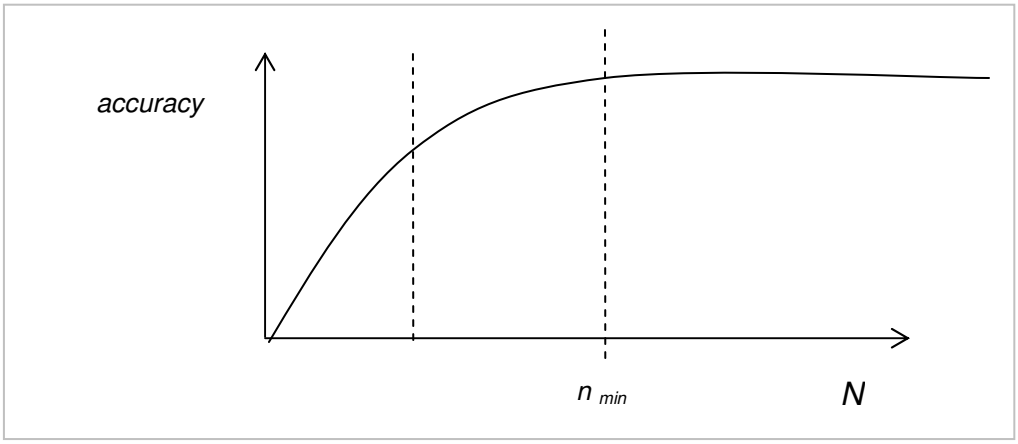

Figure 2.1 A typical learning curve

2.2.4 Aggregate model construction

The idea of using an aggregate model originates from the work of Breiman (1996) on bagging predictors. Breiman (1996) has demonstrated that, by creating classifiers from many bootstrap samples of a small dataset, prediction performance may be greatly improved. Bootstrap samples are created by using sampling with replacement in order to create many training datasets each with the same size as the original dataset. Hand et al (2001) have observed that model aggregation has conceptual similarities with Bayesian model-averaging. For Bayesian model-averaging all models in the model space are used in order to maximise predictive accuracy. The vote of each model is weighted by the posterior probability of that model, given the training data (Domingos, 2000b; Ali & Pazzani, 1996). Since the generation of all models is intractable, all implementations of aggregate modeling have to be approximations, and bagging predictors are an example of such an approximation (Ali & Pazzani, 1996).

Chawla et al (2001) have proposed a method of improving classifier accuracy by partitioning a large dataset, constructing a base model with all the data from each partition, and combining the base models into an aggregate model. Chawla et al (2001) have concluded that such a strategy leads to a higher level of predictive performance compared to the use of a single model constructed from the whole dataset. Chawla et al (2001) have argued that bagging is not suitable for very large datasets. In their experiments with various ways of partitioning a dataset, Chawla et al (2001) have concluded that disjoint partitioning results in the best performance. It should be highlighted that Chawla et al (2001) used a supercomputer with a

massively parallel architecture and it took ten hours to create an aggregate model for a 3.6 million record dataset with 304 features.

Hall et al (2000) have conducted experiments that are fairly similar to those of Chawla et al (2001), using the same architecture as that used by Chawla et al (2001). The main difference in the studies is that Hall et al (2000) have used four very large datasets (1.6, 3.2, 6.4 and 51 million instances) in their experiments compared to Chawla et al (2001) who have used one very large dataset (3.6 million instances). Hall et al (2000) have observed that for different datasets, different amounts of partitioning provide different levels of accuracy. Hall et al (2000) have reported that accuracy will actually decrease when partitioning is applied to very large datasets where very small classes are present in the data. Partitioning of such datasets causes the very small classes to appear as noise. The main conclusion made by Hall et al (2000) is that the use of disjoint partitions of a very large dataset may result in a model with the same accuracy as that obtained without any partitioning. Hall et al (2000) have further concluded that the use of overlapping subsets, in a manner similar to bagging, may provide an increased level of accuracy.

Ali and Pazzani (1996) have studied the use of aggregate models on data originating from many different domains. The objective of Ali and Pazzani's (1996) study has been to explain why there is a significant variation in prediction error reduction from domain to domain when aggregate models are used. Ali and Pazzani (1996) have tested twenty nine (29) datasets and found that aggregate models provide significant prediction error reduction on only half of these datasets. Ali and Pazzani (1996) have made four main conclusions from their study. The first conclusion is that aggregate models are better at reducing prediction error on domains for which the prediction error is already very low, than on domains that have noisy data. The second conclusion is that aggregate models improve prediction performance in those domains with many irrelevant features. The third conclusion is that as the number of irrelevant features increases, the performance of aggregate models decreases. The fourth conclusion is that when the prediction errors made by the base models are strongly correlated, the aggregate model does not provide any prediction performance improvements.

Several authors (e.g. Ho, 1998; Krogh & Vedelsby, 1995; Kwok & Carter, 1990; Hansen & Salamon, 1990) have argued that when aggregate models exhibit syntactic diversity, then major improvements in prediction performance should be

realised. On the other hand, Ali and Pazzani (1996) have argued that the accurate models that can be learned for several domains are syntactically similar, so that increasing syntactic diversity does not result in improvements. Ali and Pazzani (1996) have further argued that in order to minimize aggregate model prediction error, it is necessary to balance increased diversity with competence, that is, ensure that the base models are all competent, and have a very high level of training accuracy.

Ho (1998) has discussed the use of decision forests for the improvement of decision tree accuracy. For a decision forest, an aggregate model is constructed through random sampling of the feature space. Each classification tree that is constructed is capable of (an expert in) classification of instances that reside in the instance space defined by that subset of features which has been randomly selected. The combined performance of the decision forest is then higher than that of a single decision tree that is created to predict in the instance space defined by all the features of the dataset. The experiments conducted by Ho (1998) on feature space partitioning have been based on small datasets. Ho's (1998) method however shows promise for a divide-and-conquer approach for very large datasets of high dimensionality. The method demonstrates that syntactic diversity can be achieved through variation of the feature space for each base model.

Chan and Stolfo (1998) have proposed a method of aggregate model construction that addresses the problem of handling large two-class datasets with skewed class distributions. Chan and Stolfo (1998) have compared their method to that of using a single model and have concluded that their method provides superior performance. A more detailed discussion of Chan and Stolfo's (1998) method is given in section 2.6 where the methods that combine dataset sampling and partitioning are discussed.

Boosting (Freund & Schapire, 1997) is a method of aggregate model construction which combines training set selection with aggregate model creation. For boosting, a sequence of base models is created, with each base model in the sequence having a higher level of competence at the classification of 'difficult' instances. In this context a 'difficult' training instance is one that cannot be classified correctly by all preceding base models in the sequence. A more detailed discussion of boosting is provided in section 2.6 of this chapter.

2.2.5 Serial and parallel model aggregation

In general all aggregate models consist of two components. The first component is the set of base models. The second component is the combination algorithm. A combination algorithm may perform *parallel combination* or *serial combination* of the predictions of the base models. The methods for aggregate model construction which were discussed in the last section employ a *parallel combination* algorithm. The method of *parallel combination* consists of two steps. In the first step, all the base models make their individual predictions. In the second step, the combination algorithm selects that prediction with the strongest supporting evidence. Kittler (1998) has observed that base model combination methods for parallel aggregation fall into two categories. The first category involves discrete classification (Fawcett, 2004, 2006) where only the class labels for the classes predicted by the base models are available. For this category, a voting scheme based on the majority rule (Breiman, 1996; Hansen & Salamon, 1990) is appropriate for the combination of base model predictions. The majority rule is implemented by selecting that class which is predicted by the majority of base models.

The second category involves probabilistic classification (Fawcett, 2004, 2006) where probabilistic scores for each class are provided by the base models. Given the base models $M_1,..,M_A$, and the classes $c_1,..,c_k$, let the probabilistic scores assigned to a query instance $\mathbf{x}_q$ by models $M_1,..,M_A$ for classes $c_1,..,c_k$, be denoted by the values $CONF(c_i,M_j)$, $i=1,..,k, j=1,..,A$. Kittler (1998) has discussed four different rules that can be used to combine the $CONF(c_i,M_j)$ scores in order to select the winning class. The *product rule* involves the multiplication of the scores for each class to obtain the combined class score $conf_i$ for each class, where $conf_i$ is defined as (Kittler, 1998)

$$conf_i = \prod_{j=1}^{A} CONF(c_i, M_j) \tag{2.1}$$

and selecting the class with the largest value of $conf_i$ defined as (Kittler, 1998)

$$conf_i^* = max\{conf_1,..,conf_k\} \tag{2.2}$$

The *sum rule* involves the summation of the scores for each class to obtain the combined class score $conf_i$ defined as (Kittler, 1998)

$$conf_i = \sum_{j=1}^{A} CONF\ (c_i, M_j) \tag{2.3}$$

and selecting the class with the largest value of $conf_i$ as defined by equation (2.2). The *max rule* involves the selection of the class with the largest score defined as (Kittler, 1998)

$$conf_i = \max_{j=1}^{A} CONF\ (c_i, M_j) \tag{2.4}$$

and selecting the class with the largest value of $conf_i$ as defined by equation (2.2). The *min rule* involves the selection of the class with the smallest score defined as (Kittler, 1998)

$$conf_i = \min_{j=1}^{A} CONF\ (c_i, M_j) \tag{2.5}$$

and selecting the class with the largest value of $conf_i$ as defined by equation (2.2). Ho (1998), and Kwok and Carter (1990) have implemented the sum rule for decision tree base models by computing the arithmetic mean of the scores for each class and selecting that class with the largest arithmetic mean score. Berry and Linoff (2000: pg 217) have provided an illustrative example of how the product rule may be implemented.

More recently, a second method of base model combination called *serial combination* has been proposed (Sun & Li, 2008; Neagu, 2006; Kim et al, 2002). *Serial combination* is a multi-step process. In the first step the base models are arranged in a series. In order to classify a new instance, the instance is passed to the first base model in the series. If the base model makes a *'credible prediction',* then the process stops otherwise the instance is passed to the next base model in the series. In general, if a base model makes a 'credible prediction' the process stops otherwise the instance is passed to the next base model in the series (Sun & Li, 2008).The meaning of a *'credible prediction'* may be defined and implemented in a variety of ways. Sun & Li (2008) have used the following definition and implementation. For each class, the base model that has the highest predictive accuracy on that class is

identified. When a base model predicts a class that it is best at predicting, the base model has made a '*credible prediction'*, otherwise the prediction is considered to be '*not credible'*. Sun and Li (2008) have demonstrated that their method of serial combination produces an aggregate model whose performance on each class is as good as the performance of the best base model on the class.

For the research reported in this thesis, the method of *parallel combination* was studied. In chapter 10, a comparison is made between the advantages of serial combination and the advantages of the methods proposed in this thesis.

2.2.6 Model testing

Traditionally, the three methods of model testing in machine learning and statistical pattern recognition are, the hold-out method, *K*-fold cross validation, and the bootstrap method (Mitchell, 1997; Moore & Lee, 1994). These methods of model testing were designed for model construction from small datasets, and primarily address the problem of data shortage. For the hold-out method the available data is split into a training set and a test set (hold-out set). The test set is used to estimate the predictive accuracy. The test set may be ½, ⅓, or ¼ of the available data. For, *K-fold* cross validation, the available dataset consisting of *n* instances is divided into *K* subsets of equal size. For each of the *K* subsets, the remaining *K-1* subsets are combined into the training set, and the remaining subset is used to estimate the error. For $K << n$, the entire process is typically iterated many times (e.g. 100) and the results are averaged. When $K = n$, the leave-one-out (LOO) method is obtained. For the bootstrapping method, a training set of size *n* is chosen randomly with replacement, which means that each item may appear more than once in the training set. Only those items that do not appear in the training set are used for the test set and only once each. This process is iterated many times (e.g. 200) and the error rates are averaged (Moore & Lee, 1994).

Testing models in the presence of large volumes of data continues to be done using either K-fold cross validation or the hold-out method. *K*-fold cross validation is used to establish the accuracy on the training data. When only one model is being considered, the hold-out method is used to create two datasets, one for training and one for measuring the predictive accuracy of the final model. When several models are constructed with the objective of selecting the best one, the hold-out method is

used to create three datasets, one for training, one for validation, and one for measuring the predictive accuracy of the final model that is selected. The validation dataset is used to determine which of the many models has the best predictive performance. Given the stochastic behaviour of predictive models, several samples taken from the validation and test datasets are used for both the validation and testing steps and the results are averaged. Several researchers have argued that predictive accuracy should not be the only measure of model performance (Osei-Bryson, 2004, 2007; Giudici, 2003; Hand, 1997). Various measures of classification model performance are discussed in chapter 4.

2.3 The dataset selection problem

It was stated in section 2.2.2 that data preparation is one of the phases of the CRISP-DM model for applied data mining (Shearer, 2000). Within CRISP-DM, data preparation involves three steps namely data selection, data construction, and feature selection, among others. Data selection is concerned with the identification and selection of sufficient quantities of good quality data that is relevant to the data mining goals (Shearer, 2000). Data records (instances) as well as relevant attributes (features) are identified and selected during this step. Data construction is concerned with the creation of any necessary new features, for example, the class variable for classification (Shearer, 2000).

The data selected during the data preparation phase as prescribed in the CRISP-DM model is commonly pre-processed further when the modeling task is to create predictive models. Firstly, it is important to select training data so that overfitting of predictive models is avoided (Smyth, 2001; Dietterich, 1995). This is accomplished through data reduction. Hand et al (2001) have advised that one approach to reducing the amount of training data when the objective of data mining is to create models, is through sampling from the very large dataset. A second approach that is suggested by Hand et al (2001) is the use of sufficient statistics. Hand et al (2001) have provided least squares regression as an example of modeling where the use of sufficient statistics is enough to estimate the regression coefficients. For least squares regression, the sufficient statistics are the sum for each variable, sum of squared values for each variable, and sum of products for the values of the regression variables. Note that regression models are predictive. For classification, there are algorithms for which the usage of sufficient statistics seems feasible. The

Naïve Bayes classifier (Mitchell, 1997) is characterised by two types of probabilities: the probability of the class and the probability of a variable value given the class. For the creation of a Naïve Bayes classifier, the data records could be replaced by the probability values.

Secondly, pre-processing may be done to make the data suitable for a classification algorithm. For example, artificial neural networks (Engelbrecht, 2002; Bishop, 1995) require normalised data, and K-nearest neighbour algorithms (Cover & Hart, 1995) perform best with normalised data. Thirdly, pre-processing may also be done to increase the likelihood that the classification algorithm will produce a classification model with high predictive performance. This third type of pre-processing involves selecting the most relevant training data for the classification task (e.g. Blum & Langley, 1997), or altering the probability distribution of the training data when data has a skewed class distribution (e.g. Chan & Stolfo, 1998; Kubat & Matwin, 1997). Fourthly, pre-processing is done to further select the most relevant features for the prediction task.

The dataset selection problem addressed in this thesis was concerned with the selection of relevant features and relevant training data for the construction of many base models that make up an aggregate model. The use of aggregate models was studied for purposes of increasing the amount of (relevant) training data while at the same time avoiding the problem of overfitting. Training dataset selection was directed at classification algorithms for which data appears in raw form (at the instance level) to the algorithm. The next two sections provide a discussion of dataset selection methods that have been found appropriate for the modeling methods discussed in the last section, and for which training data must be presented to the algorithm at the instance level as opposed to a summarised (aggregated) level. Feature selection methods are discussed in chapter 3.

2.4 Theoretical methods for single sample selection

Predictive data mining has its roots in the fields of machine learning and statistical pattern recognition. The purpose of this section is to discuss the theories of machine learning and statistical pattern recognition which have been proposed for purposes of characterising the behaviour of algorithms that create predictive classification models through a process of induction from supplied example data. These theories may be

used to estimate a sufficient sample size, or sample complexity, for achieving a given level of accuracy for a single predictive classification model. The important lessons to be learned from the theories on sample complexity, as well as the weaknesses of these theories, are highlighted in this section. The probably approximately correct learning theory is presented in section 2.4.1. The theory on the Hoeffding-Chernoff bounds is discussed in section 2.4.2.

2.4.1 Probably Approximately Correct (PAC) learning

The probably approximately correct (PAC) theoretical model of learning proposed by Valiant (1984) and discussed by Mitchell (1997) has been designed for purposes of characterising algorithms that learn target concepts by generating a hypothesis *h* from a set *H* of all possible hypotheses that belong to some concept class. The learning algorithms use training instances drawn at random according to some unknown, but fixed, probability distribution. PAC is concerned with the identification of classes of hypotheses that can and cannot be learned from a polynomial number of instances. Within the PAC theory various measures of hypothesis space complexity have been proposed for purposes of establishing bounds for the number of training instances required for achieving a given level of accuracy for inductive learning algorithms. Within the PAC framework, a learning algorithm that finds the hypothesis $h \in H$ with the minimum training error is called an *agnostic* (or robust) learner. For a hypothesis space *H*, it is guaranteed with probability $(1-\delta)$, that an agnostic learner will output a hypothesis $h \in H$, which has a prediction error rate of at most ε. This guarantee will hold provided that *n*, the size of the training sample used to generate *h*, conforms to (Mitchell, 1997)

$$n \geq \frac{1}{2\varepsilon^2}\left(\ln\frac{1}{\delta} + \ln \mid H \mid\right) \tag{2.6}$$

Equation (2.6) is applicable to classes of hypotheses for which /*H*/, the size of the hypothesis space, is finite. One major problem with the sample complexity estimates based on equation (2.6) is that the size of the hypothesis space is not always easy to estimate. As an example, for decision trees the hypothesis space is the set of all possible decision trees that can be created from the given dataset. A second problem is that the instances in the training sample are assumed to be independent and

identically distributed, a requirement that is extremely difficult to satisfy. A third problem is that the hypothesis space may be infinite in size. For infinite hypothesis spaces, a useful measure of the complexity of *H* is its Vapnik-Chervonenkis dimension, *VC(H)* (Vapnik & Chervonenkis, 1971). *VC(H)* is the size of the largest subset of instances that can be shattered (split in all possible ways) by *H.* An alternative upper bound for the sample complexity *n,* under the PAC model is given by (Mitchell, 1997)

$$n \geq \frac{1}{\varepsilon}\left(4\log_2(2/\delta) + 8VC(H)\log_2(13/\varepsilon)\right) \tag{2.7}$$

One major problem with the sample complexity estimates based on equation (2.7) is that it is not always easy to estimate the *VC* dimension for a given classification algorithm. Additionally, the *VC* dimension might be infinite, as is the case for a fully grown decision tree. In artificial neural network learning, however, the application of the *VC* dimension has been used successfully. A general criticism of the use of equations (2.6) and (2.7) is that they provide a training sample size estimation which is usually excessively large.

2.4.2 The Hoeffding-Chernoff bounds

The Hoeffding-Chernoff theorems (Hoeffding, 1963) have been proposed by several researchers (e.g. Watanabe, 2005; Domingo et al, 2002; Kiniven & Manila, 1993) as an alternative method for training sample size estimation. Kiniven and Manila (1993) have discussed the use of concentration bounds (Hoeffding-Chernoff bounds) for determining sufficient sample sizes for a specified level of accuracy, when determining the truth of universal sentences expressed as first order logic formulae. Toivonen (1996) has discussed the use of these bounds for sample size estimation in association rule mining. The major criticism of the usage of the Hoeffding-Chernoff bounds is similar to that of PAC estimates. The sample sizes they estimate are usually excessively large.

Watanabe (2005) and Domingo et al (2002) have proposed an adaptive sampling scheme, which incorporates the use of the sample size bounds stated by the Hoeffding-Chernoff theorems. Watanabe (2005) and Domingo et al (2002) have argued that the methods they have proposed preserve the theoretical guarantees

(level of accuracy and confidence in the level of accuracy) of the theorems while at the same time providing good and practical estimates of sample sizes.

2.5 Empirical methods for single sample selection

For the empirical estimation of a sufficient training sample size, three approaches have been reported in the literature. A sufficient training sample size is one which provides a level of predictive accuracy that is comparable to processing the whole dataset. The first approach to empirical sample size estimation involves taking progressively larger samples from a large dataset until the sufficient sample size has been reached (Provost et al 1999; John & Langley 1996). The second approach is based on the assumption that a sample that has statistical similarity to the whole dataset is a sufficient sample. Statistical similarity is measured in terms of the descriptive statistics for the dataset variables (Lutu & Engelbrecht 2006; John & Langley 1996). The third approach to the empirical estimation of sufficient sample sizes is to select samples based on the characteristics of the instance space (Palmer & Faloutsos, 2000; Kubat & Matwin, 1997). The three approaches are discussed in this section. Dynamic sampling and progressive sampling methods are discussed in section 2.5.1 and 2.5.2 respectively. Static sample size estimation is presented in section 2.5.3. Density-biased sampling and one-sided sampling are respectively discussed in sections 2.5.4 and 2.5.5.

2.5.1 The Dynamic Sampling method

John and Langley (1996) have proposed a method they call dynamic sampling, which combines database sampling with the estimation of classifier accuracy. The method is most efficiently applied to classification algorithms which are incremental, for example, Naïve Bayes and artificial neural network algorithms such as backpropagation. John and Langley (1996) have defined the concept of *'probably close enough' (PCE),* which they use for determining when a training sample size provides an accuracy that is probably good enough. *'Good enough'* in this context means that there is a small probability δ that the classification algorithm could do better by using the entire dataset. The smallest sample size *n*, is chosen from a dataset of size *N*, so that $P_r(accuracy(N) - accuracy(n) > \varepsilon) \leq \delta$, where

accuracy(n) is the accuracy after processing a sample of size *n*, and ε is a parameter that describes what '*close enough'* means.

Dynamic sampling works by gradually increasing the sample size *n* until the PCE condition is satisfied. *accuracy(n)* is estimated by taking a new sample from the database, classifying all instances in the sample and measuring the accuracy. *accuracy(N)* is estimated using the method of Extrapolation of Learning Curves (ELC). In their study, John and Langley (1996) have compared the accuracy of static and dynamic sampling for the Naïve Bayes classifier, and have concluded that the use of dynamic sampling results in the selection of a single sample which provides a level of accuracy that is very close to that obtained when the whole large dataset is used for classifier construction.

2.5.2 The progressive sampling method

Provost et al (1999) have proposed progressive sampling as an alternative method for the empirical estimation of sufficient training sample sizes. Provost et al (1999) have addressed the issue of convergence, where convergence means that a learning algorithm has reached its plateau of accuracy. In order to detect convergence, Provost et al (1999) have defined the notion of a sampling schedule as a sequence $\{n_0, n_{1,}..n_i\}$ of sample sizes to be provided to an inductive algorithm. Provost et al (1999) have argued that schedules where the sample size n_i increases geometrically as $\{n_0, a.n_0, a^2 n_0, ..., a^i n_0\}$ are asymptotically optimal. Progressive sampling is similar to the adaptive sampling method of John and Langley (1996), except that a non-linear increment for the sample size is used. Provost et al (1999) have handled the problem of convergence detection by using a method called *Linear Regression with Local Sampling (LRLS).* *LRLS* fits a linear regression line in the neighbourhood of n_i, the size of the last training sample obtained. If the slope of the line is sufficiently close to zero, then convergence is detected. Provost et al (1999) have reported experimental results which show that geometric progressive sampling far outperforms dynamic sampling.

2.5.3 Static sample size estimation

John and Langley (1996) and Provost et al (1999) have made a distinction between static and dynamic sampling for data mining. For static sampling, n_{min}, the smallest sample size needed to achieve maximum accuracy, is determined on the basis of a sample's statistical similarity to the whole large dataset. Statistical similarity is measured in terms of the descriptive statistics for the dataset variables. Lutu and Engelbrecht (2006) have studied the selection of samples based on statistical validity and concluded that statistical validity is not a sufficient test for dataset selection. Lutu and Engelbrecht (2006) have concluded that there is a statistically significant performance difference between small statistically valid samples and large statistically valid samples. One important difference they have identified is information content as measured using the entropy function.

2.5.4 Density-biased sampling

Palmer and Faloutsos (2000) have proposed density biased sampling as a suitable method for sampling from large datasets in which clusters of differing sizes occur. Palmer and Faloutsos (2000) have argued that for such datasets, uniform sampling fails to represent small clusters (small groups) of interesting instances in the instance space. For density biased sampling, the aim is to sample so that within each cluster, instances are selected uniformly to obtain a training sample that is density preserving and biased by cluster size. Density preserving in this context means that the expected sum of weights of the sampled instances for each cluster is proportional to the cluster's size. The method of density-biased sampling is used to select instances to be included in the dataset based on the density of the various regions of the instance space. The purpose is to ensure that all regions of the instance space are equally represented in the selected dataset.

2.5.5 One-sided sampling

One-sided sampling is a training sample selection method that has been proposed by Kubat and Matwin (1997) for the selection of training instances based on the class distributions in the different regions of the instance space. Kubat and Matwin (1997) have argued that one-sided sampling is suitable for datasets with skewed class

distributions. For datasets with skewed class distribution, Kubat and Matwin (1997) have argued that the training datasets should be selected based on where the decision boundaries of the classes lie in the instance space. For 2-class problems with positive and negative instances for the concept to be learned, Kubat and Matwin (1997) have identified four types of negative instances as follows:

(1) Noisy instances. These are instances that incorrectly have the negative class label.

(2) Borderline instances. These are instances that are located very close to the decision boundary between the positive and negative class.

(3) Redundant instances. These are instances that lie far away from any decision boundary.

(4) Safe instances. All the instances that do not fall into any of the above categories are safe instances.

For the one-sided sampling method, instances that fall in categories (1), (2) and (3) above are removed from the training dataset. The rationale for one-sided sampling is that when one-sided samples are used for training, then the regions of class confusion are removed from the training data. Therefore classifiers based on discriminative classification should not experience any class confusion. Kubat and Matwin (1997) have demonstrated that this scheme produces good training datasets for the k-Nearest Neighbour and decision tree classifiers. One obvious problem with one-sided sampling is that when borderline negative instances (category 2) are removed from the training dataset, the resulting predictive model has limited information to predict instances that are located in the borderline regions. However, the results of the studies conducted by Kubat and Matwin (1997) may be used to argue that purposeful dataset selection, based on the characteristics of the instance space, may lead to the selection of training datasets that result in a higher level of predictive accuracy compared to training datasets obtained through pure random sampling.

2.6 Methods for selecting multiple training datasets

The construction of aggregate models requires the use of several training datasets. Each training dataset is used to construct one base model, and the base models are then combined into one aggregate model. For small datasets, methods such as bootstrapping and boosting have been devised for purposes of increasing the

number of training instances available for base model creation. Breiman (1996) has investigated the use of bootstrap sampling of small datasets in order to create the training datasets for the base models. Traditionally, boosting has been used in statistical modeling to improve model performance. Boosting involves the use of several variations of one training dataset to create several base models (Giudici, 2003; Freund & Schapire, 1997). For large datasets, partitioning and sampling have been used to create training datasets for base models. Chawla et al (2001) have investigated the partitioning of a large dataset in order to create several training datasets for the base models. Chan and Stolfo (1998) have investigated combining dataset partitioning with sampling in order to create the base models. In this section, the methods proposed in the literature, for obtaining multiple training datasets (samples) for aggregate model construction are discussed. The methods for bootstrap sampling and boosting of small datasets are presented in section 2.6.1. Partitioning of large datasets and the methods for combining partitioning and sampling from large datasets are respectively discussed in sections 2.6.2 and 2.6.3.

2.6.1 Bootstrap sampling and boosting of small datasets

For small datasets, Breiman (1996) has proposed the use of bootstrap sampling (Cohen, 1995) in order to create the required number of training datasets. Bootstrap samples are created by using sampling with replacement in order to create many training datasets each with the same size as the original dataset. Breiman (1996) has recommended that at least 30 training datasets should be generated and used to create the base models of an aggregate model when bootstrap sampling is applied to a small dataset.

Boosting is a statistical approach to model construction which aims to direct the largest effort of model construction towards the more difficult aspects of the process to be modeled. Giudici (2003) has observed that early versions of boosting fitted models on several versions of the training dataset, where the observations with the poorest fit received the largest weight. For classification modeling, Adaboost (Schapire, 2003; Freund & Schapire, 1997) is a boosting algorithm which creates many base classifiers that are finally combined into one prediction model. At each iteration of Adaboost, the training instances that are misclassified by the most recently created base classifier are assigned larger weights in the training set for the next base classifier. For classification, this means that the instances are replicated in

the next training set in proportion to the assigned weights. The rationale behind training dataset selection by Adaboost is to increase the representation of the instances that come from those regions of the instance space that are very difficult to model and predict.

The method of bootstrap sampling is commonly used in statistics to create larger datasets that have the same statistical properties as the small dataset from which the bootstrap sample is obtained (Cohen, 1995). In the context of aggregate model creation, bootstrap sampling provides a large amount of data for purposes of creating the base models. When large amounts of data are available, bootstrap sampling obviously becomes unnecessary. Boosting, as implemented in Adaboost, aims to increase coverage of the difficult regions of the instance space when there is a shortage of data, as is the case for small datasets. The studies reported in this thesis demonstrate that, first of all, the use of aggregate models as is done in bootstrap aggregation also provides performance improvements over single models when large amounts of data are available. Secondly, when large amounts of data are available, it is possible to increase coverage of the difficult regions of the instance space without using the methods of Adaboost, and without using all of the available data.

2.6.2 Partitioning of large datasets

For very large datasets, the training datasets are typically obtained by dividing the large dataset into several partitions. The most common approach to dataset partitioning for data mining is to use horizontal partitioning. For horizontal partitioning, a criterion is applied to assign each instance of the dataset to one of P partitions. The partitioning criteria that have been studied include disjoint partitioning and overlapped partitioning. For disjoint partitioning every instance in the dataset (of size N) appears in exactly one partition (Chawla et al, 2001; Hall et al, 2000). The original dataset is divided into PT partitions each of size (N/PT) so that each instance appears in exactly one partition (Chawla et al, 2001).

For overlapped partitioning an instance may appear in more than one partition (Chawla et al, 2001; Hall et al, 2000; Breiman, 1996). Each partition is created independently of the others using either random sampling with replacement or random sampling without replacement. Randomly selected instances are added to the partition until the partition is of size (N/PT). If sampling is done with replacement

some replication of instances within each partition and across the *PT* partitions will occur. If sampling is done without replacement, replication of instances within partitions does not occur (Chawla et al, 2001).

2.6.3 Combining dataset sampling and partitioning

Chan and Stolfo (1998) have reported experiments conducted on data for credit card transactions for purposes of identifying fraudulent transactions. Data for credit card transactions typically has a skewed class distribution with the fraudulent transactions having a representation in the range of 1% to 5% of the whole dataset (Chan & Stolfo 1998). In their studies, Chan and Stolfo (1998) have addressed the problem of creating training datasets with balanced class distributions and then creating base models from each training dataset. In order to create the training datasets, they have proceeded as follows. First, the whole dataset is partitioned according to the two classes *{normal, fraudulent}* to create two partitions *NORMAL* and *FRAUDULENT*. Since *fraudulent* is the minority class and the objective of partitioning is to balance the class distributions, the *NORMAL* partition is further divided into smaller partitions $NORMAL_1,...,NORMAL_J$. The training datasets for the base classifiers are then constructed by combining each of the small partitions $NORMAL_1...,NORMAL_J$ with the partition *FRAUDULENT.* In other words, each of the training datasets has all the minority class instances and *(1/J)*th of the majority class instances. Chan & Stolfo (1998) have concluded that compared to simple random sampling, this method of constructing training datasets results in better predictive performance for datasets with skewed class distributions.

2.7 Conceptual views of classification modeling

There are two well accepted (conceptual) views of classification, namely: discriminative classification and probabilistic classification (Hand et al, 2001). It is important to briefly discuss these views of classification modeling in order to establish the extent to which methods of data selection from large datasets attempt, or should attempt to satisfy the objectives of these views. Sections 2.7.1 and 2.7.2 respectively provide a discussion of discriminative and probabilistic classification modeling. A

concise definition of decision boundaries for classification that was adopted for the experiments of this thesis is given in section 2.7.3. Section 2.7.4 provides a discussion of training dataset selection methods aimed at supporting the objectives of classification modeling.

2.7.1 Discriminative classification

For discriminative classification (Hand et al, 2001), a classification model provides a mapping, m, from an instance $\boldsymbol{x} = (x_1,...,x_d)$ in the *d*-dimensional instance space to a set of classes $\{c_1,...,c_k\}$. The *d*-dimensional instance space is viewed as consisting of regions with labels for each of the *k* classes. The mapping, m, defines the various regions of the instance space. For each class c_i, the union of all the regions with that class label is called the *decision region* for the class. The mapping may also be interpreted as a definition of the *decision boundaries* between the *decision regions.* For real life classification problems, the classes are usually not perfectly separable in the *d*-dimensional instance space so that there are regions of class confusion for the mapping, m. Discriminative models handle the problem of class confusion by assigning a probability for each class to each decision region in the instance space. In the process of classification, a new instance ***x*** is assigned to the most probable class for the region in which it falls. The classification modeling problem may therefore be defined as a process of estimating the *decision boundaries* as closely as possible, with the objective of minimizing class confusion in each decision region. Examples of classifiers that follow this approach are decision trees for classification (Quinlan, 1993; Quinlan, 1986; Breiman et al, 1984), artificial neural networks (Engelbrecht, 2002; Bishop 1995), and K Nearest Neighbour (Cover & Hart 1967).

2.7.2 Probabilistic classification

Probabilistic models for classification are based on the assumption that, for all instances $\boldsymbol{x} = (x_1,...,x_d)$ belonging to class c_k, there is a probability distribution or density function governing the characteristics of the class c_k. For example, the probability distribution functions for a multivariate dataset with quantitative features might be multivariate normal with estimated means and variances for the features

(Hand et al, 2001). If the means associated with the different classes are far enough apart and the variances are small, then the classes will be well separated. In practice, the appropriate functional forms for describing the probability distributions for the classes are not known. However, it is possible to estimate from the data the prior probabilities $p_r(c_i)$ for each class, and the posterior probabilities $P_r(c_i \mid (x_1,...,x_d))$ of instance ***x*** **=** $(x_1,...,x_d)$ belonging to class c_i. The posterior probabilities $P_r(c_i \mid (x_1,...,x_d))$ can be viewed as carving the instance space into at least *k* decision regions and at the same time defining the decision boundaries for the classes. An examples of a modeling method based on probabilistic classification is the Naïve Bayes classifier.

One distinguishing characteristic between *discriminative classification modeling* and *probabilistic classification modeling* is that *probabilistic models* are created by computing the prior and posterior probabilities that determine whether an instance belongs to a given class. On the other hand, for *discriminative modeling* probabilities are used when the most likely class must be assigned to an instance **x**. For *probabilistic classification*, the training datasets should have the same probability distributions as the parent dataset, but for *discriminative classification* this limitation does not hold.

2.7.3 Definition of decision boundaries and class confusion regions

One of the training dataset selection methods proposed in this thesis is based on the identification of decision boundaries for classification and those regions where predictive models confuse one class for another class (confusion regions). From a probabilistic view of classification modeling, Hand et al (2001) have defined a decision boundary between two classes c_i and c_j as a 'contour' or 'surface' in the instance space which has

$$p_r(c_i, \mathbf{x}) = P_r(c_j, \mathbf{x}) = 0.5 \qquad (2.8)$$

where $P_r(c_i, \mathbf{x})$ is the prior probability that instance ***x*** has the class label c_i and $P_r(c_j, \mathbf{x})$ is the prior probability that instance ***x*** has the class label c_j. The 'contour' defined by equation (2.8) is depicted in figure 2.2 as the bold curve.

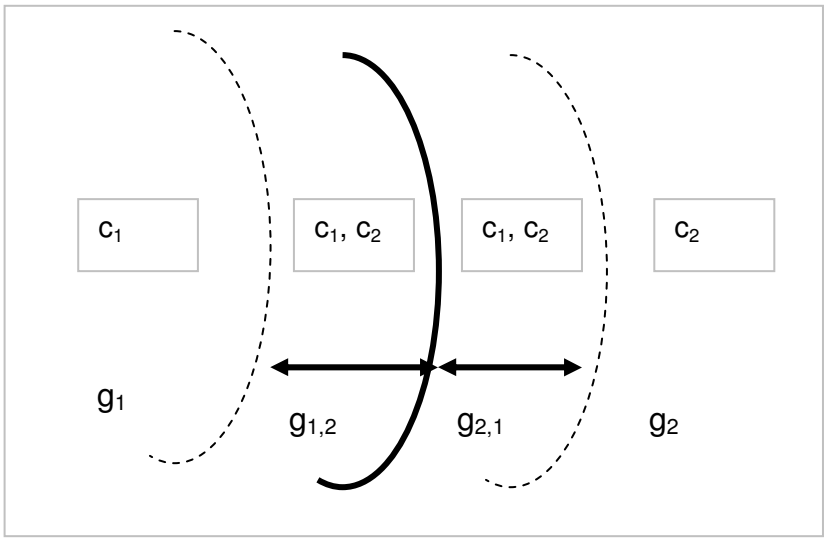

Figure 2.2: Confusion region for two classes

Based on Hand et al's (2001) definition of a decision boundary, the confusion region for classes c_i and c_j was formulated by the author as follows: On either side of the decision boundary there are two regions $g_{1,2}$ and $g_{2,1}$ where the two classes c_i and c_j occur together as depicted in figure 2.2. The region $g_{1,2}$ is characterised by three inequalities: $P_r(c_i, \mathbf{x}) > \mathrm{P_r}(c_j, \mathbf{x})$, $0 < \mathrm{P_r}(c_j, \mathbf{x}) < 0.5$ and $0.5 < P_r(c_i, \mathbf{x}) < 1.0$. The region $g_{2,1}$ is characterised by the three inequalities: $P_r(c_j, \mathbf{x}) > \mathrm{P_r}(c_i, \mathbf{x})$, $0 < \mathrm{P_r}(c_i, \mathbf{x}) < 0.5$ and $0.5 < P_r(c_j, \mathbf{x}) < 1.0$. The confusion region for classes c_i and c_j is composed of the regions $g_{1,2}$ and $g_{2,1}$. Regions g_1 and g_2 in figure 2.2 represent the instance space regions where there is no class confusion between the two classes.

2.7.4 Selection of training data to support the objectives of classification

The dataset selection methods based on theoretical bounds such as the PAC theorems (Valiant, 1984) and Hoeffding-Chernoff theorems (Hoeffding, 1963) directly

support the objectives of probabilistic classification. The dataset selection methods that employ the learning curve to empirically estimate the sufficient sample size (Lutu & Engelbrecht, 2006; Provost et al 1999; John & Langley, 1996) also support the objectives of probabilistic classification. These dataset selection methods attempt to obtain the minimum amount of data selected randomly across the instance space. The selected data enables a classification algorithm to create a predictive model based on data that reflects the natural probability distributions of the classes and variable values.

Density biased sampling (Palmer & Faloutsos, 2000) and one-sided sampling (Kubat & Matwin, 1997) directly support the objectives of discriminative classification for single model construction. These methods have the primary objective of ensuring that those regions in the instance space where prediction is difficult are sufficiently represented in the training datasets. Breiman's (1996) method of bootstrapping a dataset to create many training datasets, Freund and Schapire's (1997) method of boosting with many training datasets, and Chan and Stolfo's (1998) method of partitioning and sampling also support discriminative classification. All these methods attempt to establish the decision boundaries for the classes by using as many training datasets as possible. Additionally, Freund and Schapire's (1997) boosting method attempts to create the highest possible coverage of the decision boundary regions. Partitioning methods that process the whole dataset (Chawla et al, 2001; Hall et al, 2000) do not appear to be directed at any specific view of classification. There is, perhaps, the un-stated assumption that the large dataset is still a very large sample of the real-world data that could be collected for the application domain.

2.8 Sources of classification error

It is important to briefly examine the sources of error in predictive classification modeling. Surely if the reasons why errors arise are known, it becomes possible to design data selection methods that have the potential to produce training datasets which minimize the prediction errors. This section provides a discussion of the components of prediction error and factors that influence these components. A discussion of how training datasets can be selected to reduce prediction errors is also provided. The components of prediction error and factors that influence these errors are respectively discussed in sections 2.8.1 and 2.8.2. Methods for selecting

training data for purposes of reducing predictive classification errors are discussed in section 2.8.3.

2.8.1 Bias, variance and intrinsic errors in classification

For statistical regression modeling and artificial neural network modeling where the objective function to be minimized is the *mean squared error*, the prediction error has been decomposed into three components, namely *bias, variance* and *intrinsic error* (Giudici, 2003; Geman et al, 1992). For classification modeling in machine learning where the objective function to be minimized is the *0-1 loss* function, the prediction error has been decomposed into the same three components (van der Putten & van Someren, 2004; James, 2003; Domingos, 2000a; Friedman, 1997; Breiman, 1996; Kohavi & Wolpert, 1996; Dietterich & Kong, 1995). A prediction error has a cost (penalty) of 1 and a correct prediction has a cost of 0 for the *0-1 loss* function.

The *bia*s of a predictive model reflects how closely, on average, the (estimated) predictive model is able to approximate the target function. *Bias* reflects the error in the estimation process for the model and is due to the algorithm or inference method as well as the domain for the modeling task (van der Putten & van Someren, 2004; Giudici, 2003; Hand et al, 2001; Friedman, 1997). The *variance* reflects the sensitivity of the (estimated) predictive model to the training sample that is used to create the model. Low variance means that the (estimated) model is more stable to the variations introduced by sampling to obtain the training data (Giudici, 2003; Hand et al, 2001; Friedman, 1997). The phenomenon of *overfitting* which is discussed in the next section is also responsible for the *variance* error (van der Putten & van Someren, 2004). A simple model will have small variance, but large *bias*. A very complex model will have small bias, but large variance (Giudici, 2003).

The third component of the prediction error is called *intrinsic error* (van der Putten & van Someren, 2004; Friedman, 1997; Kohavi & Wolpert, 1996). For a given training dataset and classification algorithm, there exists a hypothetical least-error rate classifier known as the *Bayes optimal classifier* with an error rate known as *Bayes optimal error rate* (Mitchell, 1997; Breiman et al, 1984). The Bayes optimal classifier combines predictions of all possible models (hypotheses) weighted by their posterior probabilities in order to calculate the most probable prediction for a new instance (Mitchell, 1997). *Bayes optimal error rate* is the *intrinsic error* component of the

prediction error and is an irreducible component of the prediction error (van der Putten & van Someren, 2004; Friedman, 1997; Kohavi & Wolpert, 1996).

2.8.2 Factors that influence the components of prediction error

Figure 2.3 shows the components of prediction error, the factors that cause these prediction errors and the relationships between the components and the factors, as discussed in section 2.8.1. *Variance error* is caused by sampling variation in the training datasets as well as *overfitting* of models to training data. For purposes of dataset selection from large datasets it is useful to establish how *variance errors* can be reduced through the avoidance of *overfitting.* A predictive model which has a high level of predictive accuracy on the training data and a low predictive accuracy on the test data is called an *overfitted* model (Mitchell, 1997; Hand, 1997; Dietterich, 1995). The causes of *overfitting* and their relationship to *variance error* are depicted in figure 2.3. *Overfitting* arises due to one or a combination of the following reasons. Firstly, when a large number of model parameters is used in the model, the functional form (or structure) of the model becomes very complex.

For classification, examples of model parameters are the nodes of a classification tree and the nodes and connections of an artificial neural network (Engelbrecht, 2002; Hand, 1997). Secondly, when the size of the training dataset is too small and/or does not provide a representative sample for the estimation of the target function then model parameters cannot be accurately estimated (Mitchell, 1997). Thirdly, when the size of the training dataset is very large, it becomes very difficult to distinguish between noise and real structure in the data (Hand et al, 2001; Smyth, 2001; Cohen, 1995). The model is then fitted to the noise and phantom structure in the data. The first two causes of overfitting as discussed above occur most commonly when small datasets are used for training, and it could be argued that these causes of overfitting could be removed by using sufficiently large training datasets. However several researchers have cautioned against the use of very large training datasets (Hand et al, 2001; Smyth, 2001; Hand, 1998).

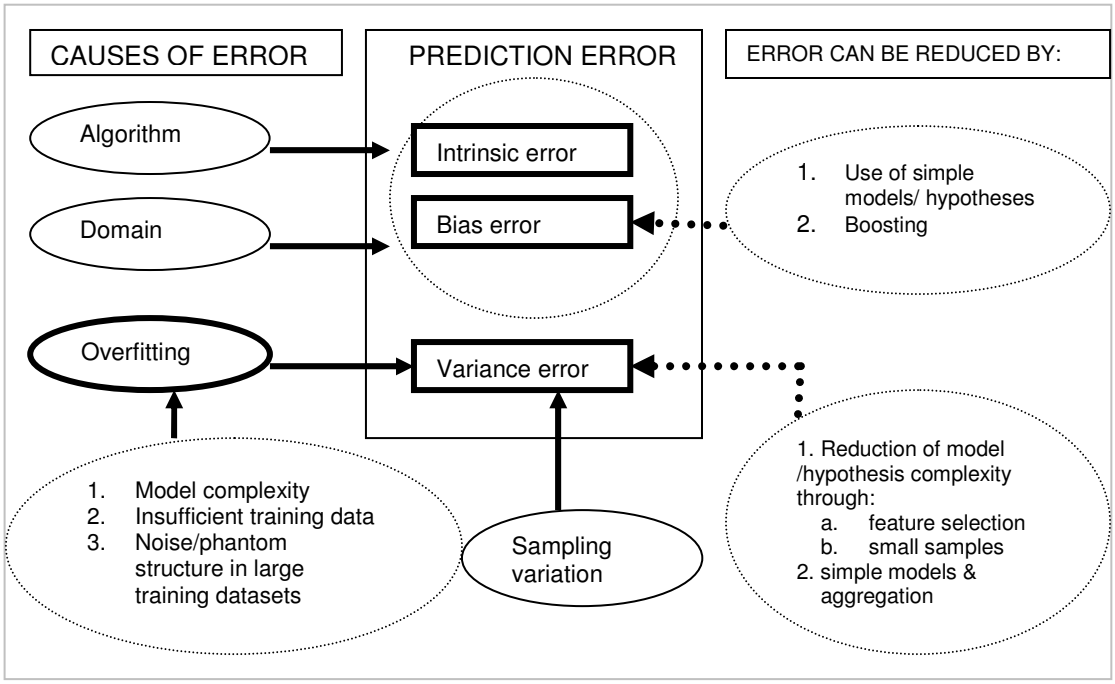

Figure 2.3: Components of prediction error and factors that influence prediction error

In statistical data analysis, the terms *massive search* and *data dredging* refer to the practice of processing as much data as possible in order to uncover evidence in support of a hypothesis (Hand et al, 2001; Smyth, 2001; Hand, 1998). The following quote is taken from Hand et al (2001):

> In the 1960s, as computers were increasingly applied to data analysis problems, it was noted that if you searched long enough, you could always find some model to fit a dataset arbitrarily well.

Smyth (2001) has warned against problems of *massive search* as practiced, for example, in association rule mining. Smyth (2001) has argued that even on purely random data where each item's values are generated randomly and independently of other items, a massive search for item associations will 'discover' significant associations between the items. These observations can also be extended to predictive modeling. The main problem here is that it becomes more difficult to distinguish between noise and real structure in the data when datasets are very large (Smyth, 2001; Hand et al, 2001; Cohen, 1995). It is argued in this thesis that one of the objectives of training dataset selection from large datasets should be to minimize the effects of noise and phantom structure in the modeling process. This in turn will lead to a reduction in the *variance* component of the prediction error.

2.8.3 Selection of training data to reduce classification error

Figure 2.3 also depicts the methods for prediction error reduction as reported in the literature. Van der Putten and van Someren (2004) have argued that *variance error* can be reduced through the use of methods that select the best predictive features. Methods for feature selection are discussed in chapter 3. The impact of *overfitting* due to noise and/or phantom structure can be reduced through the use of relatively small samples from a dataset. Cohen (1995) has advised that sampling reduces the effects of noise. The use of relatively small training datasets should lead to the reduction of variance error as long as the samples provide good coverage of the instance space. Several researchers have conducted studies to demonstrate that aggregate models based on bagging (bootstrap aggregation) (Breiman, 1996) and boosting (Freund & Schapire, 1997) achieve variance reduction (Friedman, 1997; Kohavi & Wolpert, 1996; Dietterich & Kong, 1995). Dietterich and Kong (1995) have also demonstrated that bias reduction can be achieved through the use of simple models plus increased representation of decision boundary instances as is done for boosting algorithms.

2.9 The limitations of current methods of dataset selection

The dataset selection methods discussed in this chapter for selecting training data from large datasets may be divided into three categories. The methods in the first category select and use all of the data in the belief that maximum accuracy will be achieved by processing all the data (Chawla et al, 2001; Hall et al, 2000). For the implementation of these methods, partitioning has been used in order to achieve parallel execution and fast computation of classification algorithms on massively parallel supercomputers. One obvious problem with this approach is that overfitting will occur when millions of records are used to create a predictive model. A second problem is that there is no clear explanation in the reported studies on how this approach is expected to reduce prediction error. It is the author's opinion that the objective here is to provide a very high coverage of the instance space. However, given the caution by Smyth (2001) concerning chance structure in very large datasets, one is led to conclude that high coverage of the instance space has its limits.

The methods in the second category select a subset of the data with the expectation that there is a minimum sample size, n_{min}, beyond which no further gains in predictive accuracy are possible (Lutu & Engelbrecht, 2006; Provost et al, 1998; John & Langley, 1996; Toivonen, 1996). The rationale behind this approach is that small training sets are preferred when it is prohibitively expensive to process very large datasets in reasonable time. This approach works well for single model construction. However, given the strong evidence of the superior performance of aggregate models, there is a need in the field of data mining to direct more research effort towards training dataset selection for aggregate model construction.

The methods in the third category attempt to create training datasets with balanced class distributions (Chan & Stolfo, 1998). These methods support training dataset selection for aggregate model construction. Additionally, the methods are aimed at solving the specific problem of creating predictive models from large datasets with skewed class distributions.

2.10 Proposed approach to selection of training data from very large datasets

It is the author's opinion that when large amounts of data are available, it is productive to use as much data as possible, while at the same time avoiding the problems of overfitting and the modeling of chance or phantom structure in the large datasets. The discussions in this chapter have revealed that, by reducing the bias and variance components of the prediction error, a good predictive model is obtained. This assertion is strongly supported by the success of bootstrap aggregation (Breiman, 1996) and boosting (Freund & Schapire, 1997) for small datasets. These methods are known to reduce the bias and variance components of the prediction error (van der Putten & van Someren, 2004; Friedman, 1997; Kohavi & Wolpert, 1996; Dietterich & Kong, 1995). Additionally, at the present time, there are many research efforts being undertaken in the area of aggregate model construction. These research efforts are largely motivated by the success of bootstrap aggregation. This section provides a discussion of the training dataset selection approach that was studied for this thesis, for purposes of achieving bias and variance reduction. The proposed methods for variance reduction are discussed in section 2.10.1. The proposed methods for bias reduction are discussed in section 2.10.2.

2.10.1 Variance reduction methods

Variance reduction can be achieved through at least four methods. The first method of variance reduction involves the use of many training datasets to create base models for aggregation through voting. For large datasets, this can be achieved by obtaining many randomly selected training samples from the large dataset. Each training sample is then used to create one base model. The second method of variance reduction is to provide as much coverage as possible of the decision boundary regions, as is done in boosting. For large datasets, this can be achieved by ensuring that the training datasets have many instances drawn from the decision boundary regions. The third method of variance reduction is through the avoidance of overfitting. For large datasets, the use of relatively small randomly selected training samples results in the reduction of the amount of noise (incorrect data values) and the effects of chance structure in the data. The fourth method of variance reduction is to select a good set of predictive features (van der Putten & van Someren, 2004).

The combination of the above four methods, namely: selection of many training datasets for the base models, provision of high coverage of the decision boundary regions, and the usage of relatively small training samples for the base models and, feature selection should lead to a significant reduction of the variance component of the prediction error. This approach to dataset selection was adopted for this thesis. For this proposed approach, productive usage of large amounts of data is achieved by ensuring that each of the training datasets for the base models is taken from a different region of the instance space. This approach should result in the usage of large amounts of data in the training process, without creating the problems of overfitting.

2.10.2 Bias reduction methods

Bias reduction can be achieved through at least three methods. The first method of bias reduction is through sampling to reduce the effects of noise in the training data. The second method of bias reduction is through making improvements to the algorithm for purposes of reducing bias. The third method of reducing bias is due to Dietterich and Kong (1995). Dietterich and Kong (1995) have argued that the

decomposition of a *k*-class problem into a number of 2-class problems whose solution is then converted back (combined) into the *k*-class solution, results in the correction for bias errors in the classification algorithm (Dietterich & Bakiri, 1995). Two of the three methods discussed above were combined for the proposed methods of training dataset selection. The two methods used for bias reduction were boosting of training datasets and decomposition of *k*-class problems into 2-class problems and *j*-class ($j < k$) problems.

2.11 Conclusions

The need for dataset selection has been made explicit, using examples of several application domains where data is collected in massive quantities. The examples have covered both business and scientific application areas. Methods for predictive modeling for classification using very large datasets have been discussed. These include the use of a single model and the use of aggregate models for prediction. The discussion has revealed that the methods available for aggregate model construction may result in an increase in prediction performance, but this is not guaranteed for every domain. Methods for training dataset selection have been discussed. The methods include single sample selection to obtain one dataset for training, dataset partitioning, and, a combination of partitioning and sampling to obtain several training datasets for base models. Additionally, for a given dataset, there may be other objectives, such as balancing the class distribution, which will determine the data selection method.

A discussion of the problems associated with the use of very large training datasets has been given, and reasons have been given on why it is not desirable to use very large training datasets. The various sources of classification error have been discussed. Prediction error is traditionally decomposed into two components: bias and variance. Methods of reducing bias and variance through dataset selection have been discussed. Finally, the proposed general approach to training dataset selection from large datasets in order to reduce bias and variance has been given in the last section. The next chapter presents a discussion of feature selection from very large datasets. The research methods that were used for the studies reported in this thesis are presented in chapter 4.