Low-Default Portfolio/One-Class Classification: A Literature Review.

Author: Kenneth Kennedy
Supervisors: Dr. Brian Mac Namee & Dr. Sarah-Jane Delany

DITAIG
School of Computing,
DIT Kevin St.,
D8

April, 2009

Abstract

Consider a bank which wishes to decide whether a credit applicant will obtain credit or not. The bank has to assess if the applicant will be able to redeem the credit. This is done by estimating the probability that the applicant will default prior to the maturity of the credit. To estimate this probability of default it is first necessary to identify criteria which separate the *"good"* from the *"bad"* creditors, such as loan amount and age or factors concerning the income of the applicant. The question then arises of how a bank identifies a sufficient number of selective criteria that posses the necessary discriminatory power. As a solution, many traditional binary classification methods have been proposed with varying degrees of success. However, a particular problem with credit scoring is that defaults are only observed for a small subsample of applicants. An imbalance exists between the ratio of non-defaulters to defaulters. This has an adverse effect on the aforementioned binary classification methods. Recently one-class classification approaches have been proposed to address the imbalance problem. The purpose of this literature review is threefold: (i)

Present the reader with an overview of credit scoring; (ii) Review existing binary classification approaches; and (iii) introduce and examine one-class classification approaches.

Contents

1 Introduction

This section provides an introduction to the literature review and describes the framework used to conduct the review. Section 1.1 discusses the purpose of the literature review from a general perspective and then more specifically from our problem domain, One-Class Classification (OCC) perspective. Section 1.2 provides an overview of the topics covered in the literature review. Section 1.3 supplies a brief explanation of the research methodology employed developing this review. Finally, Section 1.4 describes the structure and organisation of this review.

1.1 Purpose of Literature Review

The overall aim of this literature review is to provide an analysis of the present state of research on the application of OCC techniques in the financial domain, predominantly through the use of credit risk scorecards. Before attempting to review this subject area, it is necessary to discuss the purpose of a literature review in a wider context.

Hart (1998, pg.13) defined a literature review as:

> *"The selection of available documents (both published and unpublished) on the topic, which contain information, ideas, data and evidence written from a particular standpoint to fulfil certain aims or express certain points of views on the nature of the topic and how it is to be investigated, and the effective evaluation of these documents in relation to the research being proposed."*

The inclusion of the term unpublished is interesting. The standard practice of many researchers is to consider only peer reviewed publications. For instance Taylor and Procter (1998) define a literature review as:

> *"A literature review is an account of what has been published on a topic by accredited scholars and researchers."*

Hart's (1998) definition encourages a broad-based approach to the literature review. By this definition a full investigation from the origins of theories and ideas to their publication is promoted. From this definition it can be asserted that a literature review comprises of the identification, evaluation

and interpretation of work produced by researchers, scholars and practitioners involved in the chosen field of research. In contrast, Taylor and Procter (1998) advocate a refined or constricted approach that focuses on published material by recognised researchers. While such an approach offers safeguards against unsubstantiated statements it can promote the endorsement of established research which can lead to a one-sidedness or biasness. Furthermore, it can act as a hindrance for topics that receive little research attention.

Under the principles of identification, evaluation and interpretation this Hart's (1998) definition can be used to establish a number of goals or purposes that a literature review serves. Bourner (1996, p.8), Hart (1998, p.27) and Neuman (2003, p.96) provide a number of reasons for conducting a literature review which can be grouped as:

1. Identification

    - Detect *"gaps"* in the literature: By conducting a literature review, areas of on-going, current interest and, possibly, areas of relative neglect should become apparent. The literature review can assist the researcher in distinguishing between completed work and, to date, sparsely researched areas of the topic.
    - Proceed from where others have already reached: A literature review allows the researcher to build on the platform of existing knowledge and ideas. A review of the literature can identify seminal works in the researcher's area. Avoid unnecessary repetition of existing work (avoid making the same mistakes as others).
    - Identify the main methodologies and research techniques that have been used.
    - Identify other researchers working in the same fields. A researcher network is a valuable resource.
    - Identify opposing views

2. Evaluation

    - An aim of the literature review is to provide the intellectual context for the researcher's work, enabling the researcher to position their project relative to other work. Thus allowing the researcher to put their work into perspective.

- A literature review provides an understanding of the structure of the subject and helps establish the context of the topic or problem.
- Increases the breadth of the researcher's knowledge in the subject area and acquire/enhance their subject vocabulary.

3. Interpretation

    - Establish creditability by demonstrating access to previous work in an area. From this, place the research in a historical context to show familiarity with state-of-the-art developments.
    - A literature review can identify information and ideas that may be relevant to the project. Discover important variables relevant to the topic.
    - Identify relationships between ideas and practise. Relating ideas and theories to applications.

Using this context, the overall achievement of this literature review from an academic viewpoint is to produce a body of knowledge beneficial to the Dublin Institute of Technology, Kevin Street, Artificial Intelligence Group (AIG). It is intended as a broad overview of OCC within the financial sector and to provide a basis for future research. For the author, the high level aims of the exercise are to:

- Develop a familiarity with OCC and credit risk scorecards.
- Accumulate a body of knowledge on the topic in order to commence the establishment of an authority and credibility within the field of research.
- Show the development of prior research and place the current literature review in context to it.
- Integrate and summarise what is known in an area.
- Learn from others and stimulate new ideas (methodologies used, *"blind alleys"* to avoid).

1.2 Overview of the Literature Review

Classification means to identify whether an object is contained in a class or not (Stephan, 2001). In most classification problems, training data is available for all classes of instances that can occur at prediction time (Hempstalk *et al.*, 2008). In this case, the learning algorithm can use the training data for the different classes to determine decision boundaries that discriminate between these classes (Hempstalk *et al.*, 2008).

Classification models are used by financial institutions for classifying an applicant for credit into classes according to their likely loan repayment behaviour (e.g. *"default"* or *"not default"*) (Hand & Henley, 1997). The term used to describe the estimate or classification of this repayment behaviour is called the probability of default (PD).

Credit scoring (also known as credit risk scoring) is the term used to describe the process of estimating the PD. Classification models in the form of scorecards, use predictor variables (or characteristics) from credit application forms and other sources to yield estimates of the probability of defaulting (Hand & Henley, 1997).

Estimating the PD of an asset group is the responsibility of a financial institution's risk manager. The PD estimate is required both for internal risk control procedures and for regulatory compliance (Kiefer, 2008). Under the Basel II framework (Basel Committee on Banking Supervision, 2004) for capital standards, provisions are made for financial institutions to use models to assess risks and determine minimum capital requirements (Kiefer, 2008). The Basel II framework places a strong emphasis on data and relies heavily on the use of empirically derived techniques to evaluate risk (Phipps *et al.*, 2004).

Questions and concerns from the financial industry were raised in regard to Basel II and low-default portfolios (BIS, 2005). Low default portfolios can be defined as portfolios where the bank has no, or a very low, level of defaults and is therefore unable to estimate and validate the PD on the basis of a proven statistical significance (Sabato, 2006). There are a significant number of business types where sufficient default information is not available (Phipps *et al.*, 2004). These can be relatively new businesses, but they can also be mature portfolios where the firm has wide experience but very few, if any default observations (Phipps *et al.*, 2004). Examples include (Phipps *et al.*, 2004):

- Sovereign debt

- Banks (particularly in developed countries)
- Large corporate businesses
- Repossession-style business
- Niche counterparties, such as train operating companies, housing associations, local authorities etc.
- Private banking exposures
- Residential mortgage portfolios

Financial institutions refer to this as the *low default portfolio problem*, but more generally in machine learning literature it is referred to as single-sided (or OCC). Low default portfolios can present a significant obstacle in developing credit risk models (Sabato, 2006). This project will investigate techniques for addressing the low default portfolio problem. The project will begin by examining the state-of-the-art in single-sided classification and from this, developing algorithms to address the unique features of the low default portfolio problem faced by financial institutions.

1.3 Research Methodology

This section details the methodology used during the lifetime of the literature review project. The main methodology used in the research of this review involved secondary research. According to Hart (1998, p.32) planning a literature search comprises of the following steps:

1. Topic Definition
2. Scope of Topic
3. Planned Outcomes
4. Organisation of Sources
5. Searched Sources
6. Listed Sources

1.3.1 Topic Definition

This stage involved some general reading in order to gain familiarity with the topic. As an introduction the books "*Introduction to Machine Learning*" (Alpaydin, 2004) and "*Learning from Data*" (Cherkassky & Mulier, 2007) were consulted. From these sources concepts used and cited authors were noted. The end result was an initial list of terms for further searching.

1.3.2 Scope of Topic

The search time frame was established and relevant subject areas were identified. Next the leading authors and researchers in the area were identified and noted. From this the search vocabulary was defined:

- Single-sided Classification
- One-sided Classification
- Novelty Detection
- Outlier Detection
- Credit Risk Scorecards
- Credit Risk
- Scorecards
- The works of leading gurus in the field

1.3.3 Planned Outcomes

As stated previously in Section 1.1 the purpose of this literature review is to provide an analysis of the present state of research on the application of OCC in the financial domain through the use of credit risk scorecards. This involves the identification, interpretation and evaluation of literature within the defined scope of the topic.

1.3.4 Organisation of Sources

A simple numbering system was used to track all sources. An Excel spreadsheet recorded each source number, name and location. In future, the spreadsheet will be expanded to track of cross-referencing between the sources. A reference management software tool, Zotero, was employed to manage bibliographies and references. Zotero is a free open source extension for Mozilla Firefox browser that enables users to collect, manage and cite research from all types of sources from the browser.

1.3.5 Searched Sources

A list of likely relevant sources of information to be searched was created. This list included:

- Anthologies
- Conference papers
- Journals
- Lectures
- Reports
- Textbooks
- Theses

1.3.6 Listed Sources

Working through the searched sources listed above, the following segments were searched:

- Abstracts
- Bibliographies
- Dictionaries
- Indexes

The search started with general sources and then moved to more specific sources such as Indexes. Notes on ideas and concepts to follow up were recorded. The above terms were used with Google Scholar and hard copies. All of this research was conducted primarily in Dublin and made extensive use of library facilities at the Dublin Institute of Technology, and Trinity College Dublin.

1.4 Organisation of the Review

The remaining sections of this literature review are organized as follows:

Section 2 introduces the proposed OCC application domain of the research, credit risk scorecards. Credit risk is examined in Section 2.1. The concept of risk and the ability to measure it is essential to many problems in business and economics. Section 2.2 establishes credit scoring as a form of measuring credit risk. Section 2.3 offers a distinction between consumer and corporate credit risk scoring. In Section 2.4 existing approaches to credit risk scoring are examined. Section 2.6 outlines the impact of recent regulation, Basel 2, and the increased importance of credit risk measurement.

Before reviewing OCC it is necessary to consider classification. Section 3 begins by defining classification in the context of machine learning. Section 3.2 offers an example of the traditional classification task. Following this the theory and theoretical framework of classification is discussed in Section 3.3 and Section 3.4. The evaluation techniques of classification algorithms are covered in Section 3.5. Classification methods used in the credit scoring domain are highlighted in Section 3.6. The notion of unbalanced data and its consequences to classification are explored in Section 3.7.

An overview of OCC is presented in Section 4. The question of whether or not the training of one-class classifiers is strictly confined to one class is posed and discussed in Section 4.1. The additional problems and challenges faced by One-class classifiers are reviewed in Section 4.2. Section 4.3 examines the common characteristics and goals of one-class classifiers. An overview of current OCC methods is offered in Section 4.4.

Finally, Section 5 offers a conclusion to the literature review including proposals for future research. A full bibliography is included at the end of the literature review.

1.5 Conclusion

This Section described the framework used for the literature review project. Section 1.1 discussed the purpose of the literature review from a general perspective and then more specifically from our problem domain, OCC, perspective. A definition and reasons (identification, evaluation and interpretation) for conducting a literature review were furnished. Section 1.2 provided an overview of the topics covered in the literature review and the low default portfolio problem was formally introduced. Section 1.3 supplied a brief explanation of the research methodology employed developing this review. The scope of the topic and the organisation of research sources were recorded. Finally, Section 1.4 described the structure and organisation of the literature review.

2 Credit Risk Scorecards

The purpose of this Section is to provide the reader with a broad understanding of credit risk scorecards. Section 2.1 examines credit risk and touches on its importance to society. Section 2.2 reviews credit scoring by offering a definition of the term and outlining its history. The types of credit scoring (Application and Behavioural scoring) are distinguished. The industry practice of reject inference is also highlighted. Section 2.3 offers a distinction between consumer and corporate credit risk scoring. The effects of legislative requirements such as the Basel accords are outlined in Section 2.4.

2.1 Credit Risk

As of December 24th 2008, the value of bank loans secured on real estate in the US stood at $3,760.4 billion (Federal Reserve, 2008). On the same date, consumer loans in the personal sector totaled $882 billion. In total these two figures accounted for 65% of the total loans and leases ($7123.1 billion) estimated to be held by commercial banks in the United States (Federal Reserve, 2008).[1]

As of November 2008, financial institutions in the Eurozone had outstanding loans to households, non-financial corporations and government totaling €18,295.5 billion (ECB, 2008).[2] Lending for house purchases in the Eurozone stood at €3,523.2 billion (ECB, 2008) or 20% of the former figure. Lending for house purchases in Ireland accounted for €116.5 billion or 0.64% of this total figure (ECB, 2008).[3] These figures clearly illustrate our dependence on credit and the importance played by credit suppliers. In order to ensure continued functioning and viability of these financial institutions, and arguably present day society, it is necessary to ensure proper safeguards are in place to limit the exposure to known risks. Financial institutions label this type of risk as Credit Risk. Regrdless of the size and nature of its operations, the most significant threat faced by a business is counterparts' credit risk (Wang *et al.*, 2005). Particularly for credit-granting institutions, such as mortgage

[1] Data from the Federal Reserve Board, H8, *Assets and Liabilaties of Commercial Banks in the United States.*

[2] Data from the European Central Bank, *Aggregated balance sheet of euro area monetary financial institutions, excluding the Eurosystem.*

[3] Data from the European Central Bank, *Aggregated balance sheet of euro area monetary financial institutions, excluding the Eurosystem: ire/Ireland.*

lenders and some retailers, the ability to discriminate faithful customers from bad ones is crucial (Wang *et al.*, 2005). Exposure to credit risk is the leading source of problems in banks throughout the world (BIS, 1999).

The Bank of International Settlements (BIS) provide a definition of the term credit risk:

> *"Credit risk is most simply defined as the potential that a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms."* (BIS, 2000)

The credit risk presented by a counterparty is assessed by a lender through both the lender's and the borrower's circumstances and the lender's view of the likely future economic scenarios (Thomas *et al.*, 2002). In 2008, year end, the United Kingdom (UK) Council of Mortgage Lenders (CML) estimate that 1.8% (or 210,000) of all UK mortgages are in arrears of three months or more.[4] For 2009 this figure is forecasted to rise to 4.41% (or 500,000). In total, CML expects 75,000 home repossessions for 2009. The long-term success of any financial institution is dependent on its ability to effectively manage credit risk as part of an overall approach to risk management (BIS, 2000). The credit risk inherent in the entire portfolio as well as the risk in individual credits or transactions must be managed not just by legal constraints but also as a basic operational requirement (BIS, 2000). Financial institutions are required to identify, measure, monitor, and control credit risk. Furthermore it is necessary to determine that a financial institution holds adequate capital against these risks and that they are adequately compensated for risks incurred (BIS, 2000). For most financial institutions, loans are traditionally considered the largest and most obvious source of credit risk (BIS, 2000). Additional sources of credit risk increasingly faced by financial institutions in various financial instruments other than loans include acceptances, interbank transactions, trade financing, foreign exchange transactions, financial futures, swaps, bonds, equities, options, the extension of commitments and guarantees, and the settlement of transactions (BIS, 2000). According to the BIS (2000) one of the main causes of *"serious banking problems continues to be directly related to lax credit standards for borrowers and counterparties, poor portfolio risk management"*. From this statement it is clear that there is a need for interpretable, consistent and accurate credit risk scoring.

[4]Figures supplied by CML Market Commentry accessed at http://www.cml.org.uk/cml/publications/marketcommentary/109

2.2 Credit Scoring

Thomas *et al.* (2002) define credit scoring as:

> *"...the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit".*

The basic premise of credit scoring is the determination of how likely applicants are to default with their loan repayments. Typically, this is achieved through the application of predictive statistical models. A financial institution maintains a record of customers to whom it has granted credit. Each customer has certain attributes or characteristics such as: income, employment status, outstanding loans. Credit scoring models are developed to categorise applicants for credit as either *accept* or *reject* with respect to the applicant's characteristics such as age, income and marital condition (Huang *et al.*, 2007). An objective of a credit scoring model is to perform a systematic analysis of this data and identify behavioural patterns and credit characteristics. A credit scoring model captures the relationship between historical information and future credit performance of customers and new customers (Zhang *et al.*, 2007). Effective credit risk assessment is now recognised as a crucial factor to gaining a competitive advantage which can help financial institutions to grant credit to credit worthy customers and reject non-creditworthy customers.

In order to identify potential loan defaulters, a default must first be defined. A default definition is set either by company policy or regulation. One such definition is provided in the Basel II accord (BIS, 2006). This accord contains recommendations on banking laws and regulations issued by the Basel Committee on Banking Supervision and is examined in greater detail in Section 2.4.1. Basel II (section 452 BIS, 2006) states that a default is considered to have occurred when:

- *"The bank considers that the obligor is unlikely to pay its credit obligations to the banking group in full, without recourse by the bank to actions such as realising security (if held)".*

- *"The obligor is past due more than 90 days on any material credit obligation to the banking group. Overdrafts will be considered as being past due once the customer has breached an advised limit or been advised of a limit smaller than current outstandings".*

Under a *current-status definition*, a case tests positive only if the condition holds true at the end of a period, whereas a *worst-ever definition*, tests positive if the condition holds true at any point over the period. Basel II requires that banks use a *worst-ever definition*, covering a one-year period. Basel II does not distinguish between low-default or non-low-default banking portfolios. Rather, there is a continuum between these two extremes. A portfolio is closer to the low-default portfolio end of this continuum when a banks internal data systems include fewer loss events (Basel Committee Newsletter No. 6, September 2005).

2.2.1 History of Credit Scoring

Fisher (1936) introduced the concept of discriminating between groups of a population in statistics. His work was on the use of a technique called *linear discriminant analysis* to classify different species of irises. Although Fisher's work focused on the sciences, it provided the basis for the predictive statistics used in a multitude of other disciplines. It is generally regarded that the history of credit scoring began in 1941 when Durand (1941), using linear discriminant analysis, published a study that distinguished between good loans and bad loans made by 37 firms (Crook *et al.*, 2007). In the same decade (1940s), banks granted credit or mail order firms sent merchandise based on a judgemental decision performed by a credit analyst (Thomas, 2000). Typically the decision of the credit analyst was evaluated on the basis of the 5Cs:

1. Character (reputation of the person or family)
2. Capital (leverage - what amount is being requested)
3. Collateral (security)
4. Capacity (earnings volatility)
5. Cycle conditions (macroeconomic or market conditions)

Traditional expert systems specify no weighting scheme and are inconsistent and subjective in their assessments. Over time a set of rules were developed and from this it was possible to develop statistically derived models for making lending decisions (Thomas, 2000). In the early 1950s Bill Fair and Earl Issac formed the first consultancy dealing with finance house

retailers and mail order firms and in 1958 they developed their first application risk scorecards for American Investments. Due to entrenched attitudes within the banking industry the initial take-up of credit scoring was slow, however, before long many financial institutions began to realise the potential, and adopted credit scoring as part of the decision process (Lewis, 1992; Anderson, 2007). The usefulness of credit scoring was enhanced when financial institutions used credit scoring models in the late 1960s with the introduction of the credit card (Thomas, 2000). Due to the high level of applications it was necessary to automate the lending decision (Thomas, 2000). Furthermore, many of the banks to which the credit cards were licensed were experiencing large losses. Through the use of credit scoring banks found default rates dropped by as much as 50% when compared to the judgemental scheme (Thomas, 2000). Despite this success, some people remained unconvinced about the total reliance on statistical models, which removed any human element from the decision process (Anderson, 2007). The first fully automated implementation of credit scoring was performed in 1972 by Fair Isaac for Wells Fargo (Anderson, 2007).

In the 1980s after the success of credit scoring with credit cards banks began using credit scoring for other products like personal loans, and later with home loans and small business loans (Thomas, 2000). Today, practically all major banks use credit scoring in partnership with specialised consultancies that provide credit scoring services and powerful software to score applicants, monitor their performance and help manage their accounts (Crook *et al.*, 2007). Figure 1 illustrates the idea of a very basic credit risk scorecard. Each characteristic (e.g. age) has one or more attributes (e.g. 18 - 24, 25 - 40 etc.) and each attribute is assigned a score. If the sum of the scores exceeds the cut-off threshold, an application for credit is accepted otherwise it is rejected. In some instances when the score is close to the cut-off threshold it is referred to an expert. There are a number of ways of selecting a cutoff, refer to Beranek and Taylor (1976) for an overview.

Legislative events have also served to act as a catalyst for credit scoring growth. The first such legislation was passed in the US in 1975. The Equal Credit Opportunity Acts outlaws discrimination on the basis of:

> *"race, colour, religion, national origin, sex or marital status, or age (provided the applicant has the capacity to contract); all or part of the applicant's income derives from any public assistance program; the applicant has in good faith exercised any right*

under the Consumer Credit Protection Act" (Federal Trade Commission, 1998).

Under this act the refusal to grant credit had to be made by empirical derivation and a statistical basis. The act ensures that *"all consumers are given an equal chance to obtain credit"* (Federal Trade Commission, 1998). Other such ground breaking legislative acts include the Basel Accords which are examined later.

Time at Address	< 6 Months	6 - 18 Months	> 18 Months
	100	200	300
Own House	Yes	No	
	240	90	
Propertry Location	Urban	Suburban	Rural
	65	75	70
Income (Monthly)	< €1500	€1500 - €3500	> €3500
	210	235	250
Age	18 - 24	25 - 40	> 40
	60	90	100

Figure 1: Credit Risk Scorecard

Credit scores come under a variety of different labels, that are dependent on: (i) the information source; (ii) the task being performed; or (iii) what is being measured (Anderson, 2007). The most common labels are: *Application scoring*, *Behavioural scoring*, *Collections scoring*, *Customer scoring* and *Bureau scoring*. Collections scoring predicts the probability of a loan that recently fell into arrears will remain so until a specified default period (usually 30-90 days). Customer scoring are scorecards that have been developed using the customer's characteristics on all lenders products in order to estimate the probability of default on all or some of the loans (Thomas *et al.*, 2001). Bureau scoring is a score provided by a credit bureau, usually a bankruptcy predictor that summarises the data held by them (Anderson, 2007). Application scoring and Behavioural scoring are the two most common scoring techniques and are described in detail in the following sections.

2.2.2 Application Scoring

Using information obtained from the credit applicant, application scoring generates one overall score measuring the creditworthiness of the applicant. The application scores provide a key indicator in deciding whether new credit should be granted or not. The information that forms the basis for these scoring techniques includes both the applicant's application form details and the information held by a credit reference agency on the applicant (Thomas, 2000). There is also in most cases a mine of information on previous applicants - their application form details and their subsequent performance (Thomas, 2000). Robust and accurate classification models can be built on this available data. Classification accuracy is of benefit both to the creditor (increased profit or reduced loss) and to the applicant (avoid over commitment) (Hand & Henley, 1997).

2.2.3 Behavioural Scoring

Behavioural scoring, uses the current and most recent performance of the consumer as a way of updating the assessment of consumer credit risk (Thomas *et al.*, 2001). This replaces the first snapshot used in Application scoring with a description of the dynamics of the consumer's recent performance, but the second snapshot still remains (Thomas *et al.*, 2001).

A sample of customers is chosen so that data is available on their performance either side of an arbitrarily chosen observation point. The period prior to the observation time is called the *performance* or *observation period* and is usually 6 to 12 months in length (Thomas *et al.*, 2001). Typical performance data would be average, maximum and minimum levels of balance, credit turnover, and debit turnover. Some of the characteristics are indicators of delinquent behaviour; number of missed payments, times over overdraft or credit limit. Others characteristics may reflect difficulty in money management such as the number of cash advances using a credit card (Thomas *et al.*, 2002).

The period after the observation point is the *outcome period*, which is usually taken as 12 months, and the customer, is classified as a good or a bad depending on their status at the end of this outcome period (Thomas *et al.*, 2001). A common definition is to classify a bad to be someone who is 90 days overdue at this point.

One of the disadvantages of behavioural scoring is the need for two years

worth of history to build a scorecard. Consequently the population that the scorecard is then applied to may be quite different from that it was built on. One way used to reduce this is to take a shorter observation period and/or performance period of six months.

2.2.4 Reject Inference

A financial institution possesses the application form details on those customers it rejected for credit but no knowledge on how they would have preformed (Thomas, 2000). Credit risk scorecard models constructed on the known performance of customer characteristics are referred to as the *"known Good/Bad sample"* (Siddiqi, 2005). Application scorecards developed to predict the behaviour of all applicants, using a model based on only previously approved applicants can be inaccurate (Siddiqi, 2005). The *accept population* is biased and not representative of the *reject population* (Thomas, 2000). Reject inference is a process whereby the performance of previously rejected applicants is analysed to estimate their behaviour (i.e. assign a class). Crook and Banasik (2004) define reject inferences as:

> *"Reject inference techniques attempt to incorporate characteristics of rejected applicants into the process of calibrating a scorecard based primarily on the repayment behaviour of accepted applicants."*

Reject inference enables accurate and realistic performance forecasts for all applicants. The concept of *"reject inference"* has been widely adopted by the industry. Reject inference involves predicting the unknown and will always carry a degree of uncertainty. Attempting to impute whether rejected customers will be good or bad has been the subject of considerable debate (Thomas, 2000). Hand and Henley (1993) conclude that the question of whether it is possible to impute if a customer will be "*good*" or "*bad*" cannot be overcome, unless particular relationships are assumed between the distributions of the goods and bads which holds true for both the accepted and rejected population. One solution is to accept all applicants for a short period of time and to use that group as a training sample (Thomas, 2000). However, financial institutions are constrained by the cost of default and cannot accept all applicants, and so use versions of reject inference (Thomas, 2000). Hand and Henley (1993,1994) and Reichert et al. (1983) conclude that reject inference cannot work unless additional assumptions about the data were made,

i.e. assuming particular forms for the distribution of the good and bad risks. Crook and Banasik (2004) offer a review of reject inference techniques.

2.3 Corporate Credit Ratings and Consumer Credit Scoring

This literature review is concerned with methods used to predict the probability of default for individuals, namely consumer credit risk. For clarity, it is worthwhile to consider the corporate sector and the corporate credit rating. Many of the under-lying techniques used in consumer credit risk can also be employed in corporate credit ratings.

Corporate credit ratings are used extensively by bond investors, debt issuers, and government officials as a *"surrogate measure of riskiness of the companies and bonds"* (Huang *et al.*, 2004). They are important guides of risk premiums and directly affect the marketability of bonds (Huang *et al.*, 2004). Two basic types of corporate credit ratings exist. The first, more frequently studied, is often referred to as *"bond rating"* or *"issue credit rating"*. These ratings attempt to inform the public of the probability of an investor being paid the promised principal and interest payments associated with a bond issue (Huang *et al.*, 2004). The second corporate credit rating is referred to as *"counter party credit rating"*, *"default rating"* or *"issuer credit rating"*. It is an evaluation of current opinion of an issuer's overall capacity to pay its financial obligations (Huang *et al.*, 2004). It focuses on the issuer's ability and willingness to meet its financial commitments on a timely basis (Huang *et al.*, 2004).

A credit rating agency offers an opinion as to the credit worthiness of lending to large quoted companies. This evaluation is typically indicated by an alphabetic ordinal scale, for example AAA, AA,...,CCC labels. Higher ratings mean less risk and lower ratings imply more risk in the opinion of the credit rating agency.

The details of the methods used to construct a scale are proprietary (Crook *et al.*, 2007). Typically, the company requesting a credit rating submits a package containing information such as: annual reports, latest quarterly profits, balance sheets and other such specialised information in order to perform quantitative analysis. The rating agency then assigns financial analysts to conduct research on the competitive environment, regulations and internal factors such as managerial quality and strategies (Crook *et al.*,

2007). Prior to making a rating public, senior management of the rated firm are consulted and the proposed rating is discussed.

Ratings however, are by far the most common reference on individual credit risk in the industry, for practical and regulatory reasons. In general, the corporate credit rating process involves a subjective assessment of quantitative and qualitative factors (fundamental analysis) of a particular company as well as market level variables (Huang *et al.*, 2004). In contrast, consumer credit scoring is based largely on a statistical-based approach such as discriminant analysis which classifies a population into clearly distinguishable groups (e.g. *"good"* and *"bad"*). Consumer credit scores are actual estimates of the probability of default of an applicant for credit. The usefulness of consumer credit score depends in part on the sample size, the proportions of good and bad debtors it contains, and the classification model used to distinguish between them.

2.4 Basel Capital Accords

In 1988, prompted by several international bank failures that highlighted the need for a common way to manage risk across countries, The Basel I Capital Accord (Basel I) was published. It was the work of the Basel Committee on Banking Supervision, appointed by the Bank for International Settlements. The purpose of Basel I was to improve the soundness and stability of the international banking system and provide a "*level playing field*" for lenders from different countries. This was done, in part, by standardising and increasing capital reserves held for credit risk. Basel I set minimum capital requirements, or capital-adequacy ratios, for banks by using a very simplistic approach. Prior to Basel I capital-adequacy ratios were initially set as flat percentages of banks' assets, typically somewhere between 4% and 10% (Anderson, 2007). However, this failed to recognise the riskiness of banks' assets, which varied according to the nature of their portfolios (Anderson, 2007). Basel I requires banks to assign adequacy-ratios or risk-weights to their exposures into four broad classes: (i) 0% - sovereign debt of OECD countries (S); (ii) 20% - other banks and public sector institutions in OECD countries (O); (iii) 50% - resedential loans (R); and (iv) 100% - all other loans (U). The risk-weighted assets (RWA) is obtained by adding the total lending in each class:

$$RWA = (0\% * S) + (20\% * O) + (50\% * R) + (100\% * U) \tag{1}$$

Basel-I requires banks to keep their capital ratios above 8%, where the capital ratio is defined as:

$$Capital\ Ratio = \frac{Total\ Capital}{RWA} \tag{2}$$

For instance, a bank that only has exposures to corporate loans (class iv) will be allowed to issue loans for a maximum of 12.5 times its total capital. Basel I was originally intended for internationally active banks in G10 countries, however it was adopted as a global standard by over 120 countries. Clementi(2000) observed that Basel I increased financial stability by providing: (i) a framework for determining the riskiness of assets; (ii) a definition of capital-weighted assets; and (iii) a minimum capital-adequacy ratio. Although Basel I had the desired effect of stabilising the declining trend in banks' solvency ratios by increasing capital reserves, one of it's main shortcomings was that it did not provide any further risk differentiation within its defined broad asset categories. Refer to Stephanou & Mendoza (2005, p3-4) for further details on shortcomings of Basel I.

2.4.1 Basel II

Following the publication of three consultative papers (CP1,CP2 and CP3) between 1999 and 2003, the Basel Committee members agreed in June 2004 on a revised capital adequacy framework (Basel II). In the European Union (EU) all deposit takers had to implement Basel II no later than January 2008. The US delayed this date to January 2009 due to concerns about how it will impact the competitiveness of smaller banks. Basel II consists of three pillars: (i) minimum capital requirement; (ii) supervisory review process and the role of bank supervisours; and (iii) market discipline through enhanced disclosure.

Aside from credit risk two other types of risk have been introduced: market risk and operational risk. Equation (2) can now be expanded to:

$$Capital\ Ratio = \frac{Total\ Capital}{RWA + Market\ Risk + Operational\ Risk} \tag{3}$$

Pillar 1 requires lenders to assess their market and operational risk and provide capital to cover such risk. Operational risk is defined as *"the risk of loss resulting from inadequate or failed processes,people, and systems, or from external events"* (Basel II). Market risk refers to the risk of adverse movement

in prices of traded securities (Anderson, 2007). Regarding credit risk, Pillar 1 provides for the calculation of risk weights used to determine a basic minimum capital figure. There are two approaches to calculate this figure. The simplest is the *standardised* approach, which provides set risk weights for 11 asset classes (e.g. sovereign and central banks, private-sector banks, residential property, etc.) and requires the weight on others to be determined by the public credit rating assigned to the particular asset by the rating agencies (e.g. Standard & Poor's, Moody's Corporation, etc.). Lenders can choose the more sophisticated *internal ratings based* (IRB) approach. There are two approaches within this model; *foundation* and *advanced*. These approaches allow lenders to develop their own risk models to determine appropriate minimum capital. Credit risk is estimated using four parameters: probability-of-default (PD); loss-given-default (LGD); exposure-at-default (EAD); and maturity (M).

The estimated loss (EL) is the amount that the lender expects to lose, based upon available data (Anderson, 2007). The EL can be calculated using the following:

$$\$EL = PD\% * \$EAD * LGD\% * f(M) \tag{4}$$

PD% is a borrower risk rating or the probability that a borrower will default in the horizon of one year. This figure is related to individual economic and environmental circumstances. $EAD is a monetary value related to the outstanding balance, agreed loan limit, the lender's target limits, and loan product characteristics. It is the expected total exposure or outstanding loans to the borrower at the time of default. LGD% is the proportion of the EAD that the lender expects to lose in the event of default, which is heavily influenced by collateral and other security. The expected percentage of the exposure which the bank will be unable to recover. f(M) is an adjustment that is a function of the remaining loan term or repayment schedule.

Banks operating under the *advanced* variant of the IRB approach will be responsible for providing all four of these parameters themselves, based on their own internal models. Banks operating under the *foundation* variant of the IRB approach will be responsible only for providing the PD parameter, with the other three parameters to be set externally, by the Basel committee.

Under Pillar 2, lenders are required to assess risks to their business not captured in Pillar 1, for which additional capital may be required. Pillar 3 requires lenders to publish information on their approach to risk management and is designed to raise standards through greater transparency.

Basel II poses significant challenges, and requires substantial investments in information technology and risk-assessment capabilities, both for initial compliance and ongoing improvements thereafter (Anderson, 2007). Some of the requirements under Pillars 2 (e.g. legal mandate to impose higher capital requirements) and 3 (e.g. confidentiality rules) are currently beyond the working parameters of many supervisory authorities and require changes to a country's legal and judicial framework (Stephanou & Mendoza, 2005).

Credit scoring's first significant boost came from the Equal Credit Opportunity Act in 1975 and is now receiving another boost from Basel II (Anderson, 2007). Basel II provides the basis for internal ratings for retail banking. Part of the initial attraction with Basel II was the potential of lower capital requirements due to improved risk assessment, however many financial organistations view compliance as a type of market-legitimacy that facilitates lower borrowing costs in international finance markets (Anderson, 2007).

2.5 Conclusion

This Section offered a brief snapshot of credit risk. Section 2.1 examined credit risk and part of its importance to society. Section 2.2 reviewed credit scoring by offering a definition of the term and how it has evolved from the late 1930's. The biasing practise of reject inference was introduced. Application scoring and behavioural scoring were also discussed. The emphasis of future research will be with regard to Application scoring. Section 2.3 distinguished between consumer and corporate credit risk scoring. The legislative requirements of the Basel capital accords were explained in Section 2.4. The next Section, Classification, examines the mechanics of credit scoring and offers an overview of popular credit scoring algorithms along with measures used to assess their usefulness.

3 Classification

This section will examine classification. First, in Section 3.1, the term is defined. In Section 3.2 the theoretical requirements of a classifier are discussed. Following on from this, Section 3.3 discusses the theory behind the workings of a classifier. Section 3.4 looks at evaluation techniques used for assessing the usefulness of a classifier. Credit scoring methods that are used throughout industry are reviewed in Section 3.5. Section 3.6 discusses class imbalance and maps out the approach to OCC.

Throughout this section two-class classification is considered the traditional or conventional form of classification, as a multi-class problem (i.e. more than two categories of output exist) can be treated as a two-class problem by isolating one class as the target class and combining the other classes as the non-target class. Another point to establish about this review is that classification is a form of supervised learning. Supervised learning is a *learning by examples* approach. Classification models are developed based on training examples which consist of input and output vectors. Unsupervised learning involves inferring patterns based on the input when no specific output values are supplied. In this scenario, historical data on customer loans would not contain information on whether or not the loan was repaid. This is outside the scope of this review. However semi-supervised learning, a mixture of supervised and unsupervised learning, is briefly discussed.

3.1 Definitions

A rather simple yet precise definition of classification is supplied by Russell & Norvig (2002, p.353):

> *"Classification - checking whether an object belongs to a category."*

Batista *et al.*,(2000) expand on this by stating:

> *"Supervised learning is the process of automatically creating a classification model from a set of instances, called a training set, which belong to a set of classes. Once a model is created, it can be used to automatically predict the class of other unclassified instance."*

This definition highlights the function and the means of operation of a classifier. The following section looks at the theory of classification.

3.2 Theoretical Framework

The problem of predictive learning involves estimating an unknown dependency from known observations (or training samples) (Cherkassky & Mulier, 1999). Estimating this dependency is the key to predicting future or unseen data. At present, there is no one widely accepted theoretical framework for predictive learning (Cherkassky & Mulier, 1999). Furthermore, the terminology used to describe the approaches to predictive learning varies. For instance the following terms cover the same topic; statistical learning, predictive learning, empirical learning. Part of this stems from the fact that the use of classification encompasses a number of fields of study; statistics, engineering, signal processing, biological developments and computer science.

The lack of semantic cohesion extends to efforts to distinguish between approaches for estimating learning models from data. At present we consider the Cherkassky & Mulier, (1999) differentiation of three main approaches as follows:

1. Classical (Parametric) Statistical Estimation: Using this approach the parametric form of the dependency is known (up to the value of its parameters). The parameter values are then estimated using the training data. This approach assumes a strong *a priori* knowledge about the unknown dependency. In practice it is difficult to extend the parametric approach to high-dimensional settings as a large number of training examples are necessary for accurate estimation.

2. Empirical Nonlinear Method: Examples of this type of method include artificial neural networks and flexible statistical methods that were developed in the 1980's to address the shortcomings of the parametric approach. These methods use nonlinear models based on the available data, without relying on strong assumptions about the unknown dependency. These models lack an underlying unified mathematical theory.

3. Statistical Learning Theory: This approach was developed in the late 1960's (Vapnik & Chervonenkis, 1968). *"It is a theory for nonparametric (distribution free) dependency estimation with finite data. The theory is based on the theoretical analysis of the empirical risk minimisation (ERM) inductive principle"* (Cherkassky & Mulier, 1999).

Another more general approach to differentiation worth considering is to divide learning models into parametric and non-parametric categories. The parametric category consists of classification techniques that use a verification-based approach, in which the user makes assumptions or hypothesises about the underlying data and then employs tools to verify these assumptions. The non-parametric approach is a discovery-based approach in which learning algorithms expose patterns in the data. These methods do not assume a certain form of the underlying data. The methods employ computational power to search and iterate through the data until the model achieves a good fit to the data. The non-parametric approach is best suited to learning problems where there is little knowledge about the statistical properties of the data.

3.3 Theory

According to Dietterich (2000): A standard two-class classification program is given training examples of the form: $\{(x_1, y_1) \ldots (x_i, y_i)\}$ for an unknown function $y = f(x)$.

The training examples, x_i, are typically described by vectors in the form $\langle x_{i1}, x_{i2}, \ldots x_{in} \rangle$ whose components are a set of discrete- or real-valued features such as income, savings, existing loans etc. These are also called the features of x_i.

For classification, the y values are drawn from a discrete set of values, in the case of two-class classification: $y_i \in \{-1, +1\}$

Instances for which $f(x) = 1$ are called positive members or members of the target concept. Conversely, instances for which $f(x) = -1$ are called negative examples or non-target members of the concept (Mitchell *et al.*, 1990, p.23). Given a set X^{tr} of training examples the learner must hypothesise, or estimate, f. Generally, classification algorithms search a very large space of possible hypotheses to determine a hypothesis that best fits the observed data and any prior knowledge held by the learner (Mithchell *et al.*, 1990, p.14). The symbol H is used to denote the set of all possible hypotheses that the learner considers for identifying the target concept (Mitchell *et al.*, 1990, p.23). This is also known as the hypothesis class or version space. H is determined by the user's model selection, represented by $f(\bullet)$. The model is usually preselected; examples of learning models include linear classifiers, nave Bayes, k-nearest neighbour, neural networks or support vector classifiers. Each hypothesis, h, in the hypothesis class, H, is instantiated by $f(x;w)$ where $f(\bullet)$ is the model, x is the input and w are the parameters (Alpaydin,

2004). A learning algorithm finds the optimal values of the parameters or weights based on the training set. The goal of the learning algorithm is to find a particular model such that *f(x;w) = f(x)* for all of x in X^{tr}. This process is called **Model Selection** which Hastie *et al.* (2001) define as *"Estimating the performance of different models in order to choose the (approximate) best one"*. Poggio (1990) also described this as a representation problem.

A specific hypothesis found to estimate the target function well over a large training set is also expected to perform similarly over other unseen instances. This selected hypothesis is called the inductive learning hypothesis, or inductive learner for short. Many forms of classification are ill-posed, due to a lack of knowledge about the underlying dependency and the finiteness of available data (Mulier, 1990; Cherkassky & Mulier, 1999). The available data is not sufficient to find a unique solution (Alpaydin, 2004). It is necessary to make assumptions about the underlying data in order to create a unique solution. These assumptions are called the inductive bias of the learning algorithm.

Theoretically, the best approach for both the Model Selection and Assessment processes is to split the dataset into three parts: a training set, a validation set, and a test set. The training set is used to fit the models, the validation set is used to estimate prediction error for model selection. The test set is used for assessment of the generalisation error of the inductive learner. The test set should be a hold-out set which is unused during the training and validation process. A typical split is 50% for training, and 25% each for validation and testing (Hastie *et al.*, 2001). Following the Model Selection process it is necessary to estimate the prediction error, or generalisation error, of the inductive learner (Hastie *et al.*, 2001). This is also known as the **Model Assessment** process (Hastie *et al.*, 2001).

Generalisation describes how well a model trained on training data predicts the correct output for previously unseen instances. The complexity of the hypothesis should match the complexity of the function underlying the data. According to Mitchell *et al.* (1993), a hypothesis overfits the training set if there exists some other hypothesis that fits the training set less well but performs better over the entire distribution of instances. Conversely, underfitting occurs when the hypothesis is too simple for the function of the underlying data. This is known as the bias/variance dilemma or decomposition. As bias decreases, the model becomes more flexible and variance increases. The hypothesis overfits the data and risks learning any associated noise. Noise can be defined as *"any unwanted anomaly in the data"*

(Alpaydin, 2004). Noise can be introduced in a number of ways:

- Imprecision in recording the input attributes
- Teacher noise - incorrect labelling of data
- Hidden/latent attributes - unobserved or unaccounted for attributes that affect the label of an instance.

Most learning algorithms handle noise by fitting the maximum likelihood or least squares error of noisy data (Kulkarni, 1995; Rasmussen, 1996). According to Dietterich (2003) in all supervised learning algorithms there is a triple-trade off between:

- Classifier complexity
- Amount of training data available
- Classifier accuracy on new instances (generalisation)

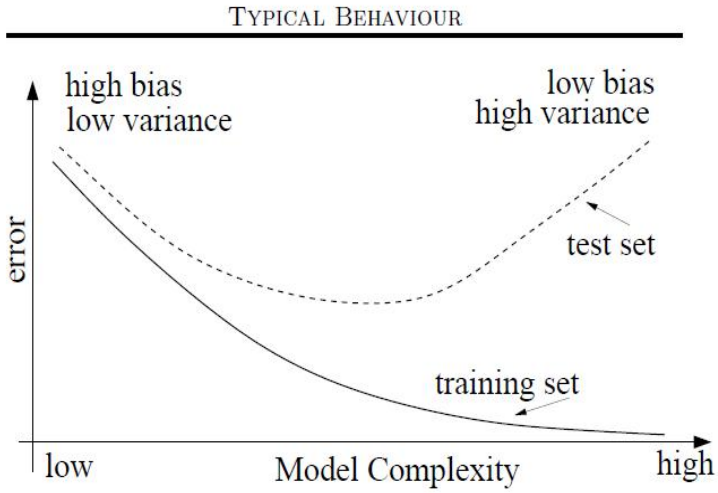

Figure 2: Model Complexity (Hastie *et al.*, p. 38)

Dietterich (2003) observed that as the number of training instances increased, the generalisation error decreased. As model complexity increased, the generalisation error decreased at first but subsequently began to increase.

This phenomenon is referred to the curse of dimensionality (Bellman, 1961). Hastie *et al.*, (2001) graphed the typical behaviour of the test and training error as model complexity varies. This is shown in Figure 2. The training error decreases as the variance increases, due to a closer fit to the data. However, if the model adapts too closely to the training data the test error increases. If the model is not complex enough, it will underfit and result in poor generalisation.

3.3.1 The Loss Function

In general, more than one model, *f(.)*, can fit a given training set X^{tr}. To find the best fitting model it is necessary to define a merit function that measures the agreement between the data and the model (Liano, 1996). A loss function *L(.)*, also known as an error function, or approximation error, is the sum of losses over the individual instances (Alpaydin, 2004). The loss function, *L(f, w,* X^{tr}*)* defines the optimal parameters w for the function f on a given training set X^{tr} (Tax, 2001, p4).

Different definitions of the loss function can be employed. A straight forward approach is to employ the 0-1-loss. This counts the number of incorrectly classified objects. The most common error for real valued functions $f(x_i;\mathrm{w}) \in [-1,1]$ are the mean squared error (MSE) (Thomas, 2000; Tax 2001):

$$\epsilon_{MSE}(f\,(x_i;w), y_i) = (f\,(x_i;w) - y_i)^2 \tag{5}$$

and the cross entropy (where labels should be rescaled to positive values $y_i = 0,1$):

$$\epsilon_{ce}(f\,(x_i;w), y_i) = f\,(x_i;w)^{y_i}(1 - f(x_i;w))^{1-y_i} \tag{6}$$

By minimising the error ϵ on the training set, an optimal set of parameters w is specified. This in turn translates to a good classification for the training set. However, an acceptable training error does not necessarily translate to an acceptable test error. The optimal parameters w^* of the function f are the parameters that result in the smallest average error over all possible samples (Tax, 2001; Huang *et al.*, 2007):

$$w^* = argmin_w \epsilon_{true}(f,w,X) \tag{7}$$

Where the true error, ϵ_{true}, is defined as (Tax, 2001):

$$\epsilon_{true}(f,w,X) = \int \epsilon(f(x;w), y)p(x,y)dxdy \tag{8}$$

p(x,y) represents the *true* data distribution.

One of the main concerns is the connection between the training and test set instances. A key assumption is the assumption of stationarity; the training and test instances are selected randomly and independently from the same population of instances with the same probability distribution (Russell & Norvig, 2002). However, the cumulative distribution of *f(x)* is unknown and the only available information about this distribution is in the finite training sample X^{tr} (Cherkassky & Mulier, 2007). It is necessary to adopt an induction principle to approximate the true error (Vapnik, 1995). ϵ_{true} is often approximated by the empirical error on the training set:

$$\epsilon_{emp}(f, w, X^{tr}) = \frac{1}{N}\Sigma\epsilon(f(x_i; w), y_i) \tag{9}$$

This error gives an approximation of the true error, which is accurate when the distribution of the training data mirrors the true data distribution and the sample size is very large (Tax, 2001).

Cherkassky and Mulier (2007) summarise that in order to form a model from data, any learning process requires the following:

1. A wide and flexible set of approximating functions $f(x; w)$.

2. A priori knowledge used to impose constraints on a potential of a function from the class (1) to be a solution. Generally, such a priori knowledge provides, implicitly or explicitly, ordering of the functions according to some measure of their flexibility to fit the data.

3. An inductive principle to act as a general prescription for combining a priori knowledge (2) with available training data in order to produce an estimate of (unknown) true dependency. Empirical Risk Minimisation (ERM) is an example of an inductive principle.

4. A learning method that constructs a (computational) implementation of an inductive principle for a given class of approximating functions. An important issue for learning methods is an optimisation procedure used for estimating the optimal parameters w^*.

3.4 Evaluation Techniques

Once a classification model has been built, it is necessary to measure its performance. In this subsection different measures for assessing the quality of learning algorithms are presented. Validation of an classification model involves assessing its discriminatory power, which is the ability to separate the distributions of observed goods and bads over the characteristics. When 0-1-loss is used, all errors are equally graded; this allows the use of a confusion matrix, also known as a misclassification matrix or a 2x2 contingency table. Table 1 presents a confusion matrix which classifies all possible situations of classifying an object in two-class classification.

Table 1: Confusion Matrix, *EI* = Error Type I, *EII* = Error Type II

-	Object classed as 1	Object classed as -1
Object from 1	True Positive (TP)	False Negative (FN) *EI*
Object from -1	False Positive (FP), *EII*	True Negative (TN)

In credit scoring a confusion matrix is created by: (i) choosing a cut-off score; (ii) marking all accounts below the cut-off score as expected bad, and all those above as expected good. The correctly classified cases are the true positives and true negatives. If labels do not correspond they are labelled false positive and false negative. In credit scoring, the cost of the two types of error are very different. Classifying a good as bad (Error Type I) results in a loss of profit, L. Whereas classifying a bad as good (Error Type II) means an expected default, D which is often considerably higher than L (Thomas *et al.*, 2002).

In credit scoring, the following measurements are used as statistical measures of classifier predictiveness. In practice most of the measures are for comparative purposes and are used in conjunction with strategic considerations when selecting the final preferred classifier (Siddiqi, 2005).

3.4.1 Common Measures

Accuracy is a widely used measure, suitable only for balanced data sets. This measure is unsuitable for imbalanced data sets for a number of reasons: (i) The measure assumes equal misclassification costs for false positive and false negative predictions, i.e. D = L; (ii) Another implicit assumption of the

use of Accuracy is that the class distribution among examples is presumed constant over time and relatively balanced (Provost et al, 1999. P43:19). Accuracy is measured as:

$$\frac{\|TP\| + \|TN\|}{\|TP\| + \|FN\| + \|FP\| + \|TN\|} \tag{10}$$

Recall or **Sensitivity** or the **True Positive Rate** is the portion of actual good matches that have been classified correctly. It is measured as:

$$\frac{\|TP\|}{\|TP\| + \|FN\|} \tag{11}$$

Specificity or the **True Negative Rate** is the portion of actual bads classified as bads and is measured as:

$$\frac{\|TN\|}{\|TN\| + \|FP\|} \tag{12}$$

Average Class Accuracy is the average of the individual class accuracy. It is measured as:

$$\frac{Sensitivity + Specificity}{2} \tag{13}$$

Precision is the number of true positives divided by the total number of instances labelled as positive. It is measured as:

$$\frac{\|TP\|}{\|TP\| + \|FP\|} \tag{14}$$

False Positive Rate is measured as (1 - Specificity):

$$\frac{\|FP\|}{\|TN\| + \|FP\|} \tag{15}$$

F-Measure combines both Precision and Recall and is measured as:

$$\frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{16}$$

3.4.2 Receiver Operating Characteristic Curve

A Receiver Operating Characteristic (**ROC**) curve is method for visualising, ranking and selecting classifiers based on their performance (Fawcett, 2004). A ROC curve is a 2-dimensional graphical illustration of the sensitivity (True Positive Rate) on the Y-axis versus 1-specitivity (False Positive Rate) on the X-axis for various values of the classification threshold. A ROC curve illustrates the behaviour of a classifier without regard to class distribution or error cost. In a credit scoring context, the X-axis depicts the percentage of bads predicated to be good (or alternatively, the percentage of goods predicted as bad) whereas the Y-axis displays the percentage of goods predicted as good (or alternatively, the percentage of bads predicted as bad). A ROC curve has properties that make them useful for domains with skewed class distribution and unequal classification error costs (Fawcett, 2004). Consider the confusion matrix in Table 1. The class distribution - the proportion of good instances to bad instances - is the relationship of the left column (goods) to the right column (bads). As with Accuracy - any performance metric that uses values from both columns will be inherently sensitive to class skews (Fawcett, 2004). As a class distribution changes measures such as accuracy will change as well, even if the fundamental classifier performance does not (Fawcett, 2004). ROC graphs do not depend on class distributions as they are based upon the True Positive Rate and False Positive Rate, in which each metric is a strict columnar ratio.

Broadly speaking, a classifier yields an instance probability or score that represents the degree to which an instance is a member of a class. Such a classifier can be used with a threshold or cutoff point to produce a discrete (binary) classifier: if the classifier output is above the cutoff the classifier produces a 0(good), else a 1(bad) (Fawcett, 2004). Using the true positive rate and false positive rate for each cutoff/threshold value, a different point on the ROC graph is produced. Combining these points results in the ROC curve.

Figure 3 provides an example of 3 ROC curves. The lower left point (0,0) represents the strategy of never issuing a positive classification; such a classifier commits no false positive errors but also gains no true positives (Fawcett, 2004). The upper right point (1,1) represents the opposite strategy of unconditionally issuing positive classifications (Fawcett, 2004). The point (0,1) represents the perfect classification. Each ROC curve passes through

the points (0,0) and (1,1). Curve A , the ratio of goods to bads is the same for all score ranges. This is no better than classifying randomly given the known ratio of good to bads in the entire population (Thomas *et al.*, 2002). The further from the diagonal curve ((0,0) to (1,1)) the ROC curve is, the better the scorecard. ROC curve C is always further from the diagonal than ROC curve B and is therefore considered a better classifier at all cut off scores. Lowering this threshold corresponds to moving from the *conservative* to the *liberal* area of the graph.

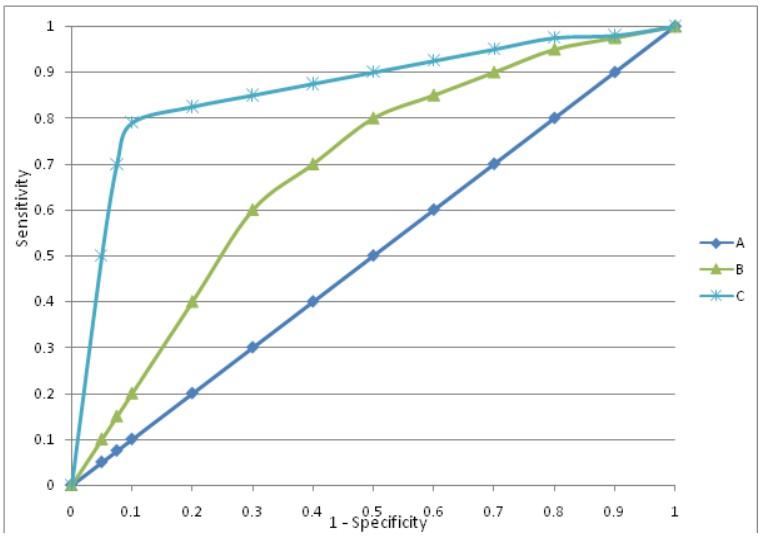

Figure 3: The receiver operating characteristic curve (ROC)

Informally, classification points on the left handside of the graph can be thought of as *"conservative"*: they make positive classifications only with strong evidence and therefore make few false positive errors, but they have low true positive rates as well (Fawcett, 2004). Classification points on the upper right handside of the ROC graph are considered *"liberal"*: they make positive classifications with weak evidence so they classify nearly all positives correctly, but they often have a high false positive rate (Fawcett, 2004). In credit scoring, the ROC curve is also known as the Lorenz curve.

3.4.3 Area Under the Curve

ROC curves of different classifiers may however intersect making a performance comparison less obvious (Baesen, 2002). This problem can be ad-

dressed by calculating the area under the ROC curve (AUC). According to Baesen (2002) "*The AUC provides a simple figure-of-merit for the performance of the constructed classifier*". Many methods have been suggested to compute the AUC (Baesens - 64,111). The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2004). This is equivalent to the Wilcoxon test of ranks (Hanley & McNeil, 1982). The AUC is closely related to the Gini coefficient (Breiman *et al.*, 1984), which is:

$$Gini + 1 = 2 * AUC \tag{17}$$

Fawcett (2004) cautions that when making conclusions about classifier superiority care should be taken when averaging the output of multiple ROC curves. A measure of variance is required to compare multiple classifiers. Vertical averaging takes vertical samples of the ROC curves for fixed FP Rates and averages the corresponding TP Rates. Sometimes when the FP Rate is not under the direct control of the researcher it may be preferable to average the ROC scores by controlling the threshold variable. This is known as Threshold averaging.

3.4.4 Other Measures

Kolmogorov-Smirnov (**KS**) is commonly used statistic in credit scoring (Anderson, 2007). It measures the maximum deviation between the cumulative distribution of target and non-target data. Separation is measured at one point only, and not on the entire score range. A KS curve (also known as fish-eye graph) is a data visualisation tool used to illustrate scorecard effectiveness (Anderson, 2007).

The following methods are used to assess scorecard strength.

Lift/Concentration Curve is calculated by (positive predicted value) / (% of positives in the sample). A lift curve can be plotted if the classifier prediction can be expressed in the form of ranking based on the predicted class probability. The lift curve then plots cumulative true positive coverage (y-axis) against the rank-ordered examples (x-axis) (Ye, 2004). A random ranking results in a straight diagonal line on this line. A lift curve of a model is usually above this line, the higher the better.

Cost Ratio is the ratio of the cost of misclassifying non-target data as target data to the cost of misclassifying target data as non-target data. Alternatively it can be expressed as the ratio of the cost of false negative to false positive.

Hand (2005 p10) questions the merit of any measure of scorecard performance which uses the distributions of scores. For instance the aim of an application model is to divide applicants into those accepted and the rest. In this scenario, it is only the numbers of cases above and below the cut-off that is relevant. The distance between the score and the cut-off point is not.

3.5 Credit Scoring Classification Algorithms

Dinh and Kleimeier (2007) refer to the process of selecting an appropriate classification algorithm for credit scoring as the estimation method and state: *"The development of the CSM* [**credit scoring model**] *starts with the decision about the basic form of the model, i.e. its estimation method via decision trees, linear probability models, logit or probit regression models, or multiple discriminant analyses"*. The basic model can take one of a number of forms. The most popular of which are discussed below.

3.5.1 Statistical Methods

Statistical methods can be divided into two approaches: parametric and non-parametric. For the parametric approach, an assumption is made that the sample is drawn from some distribution that obeys a known model, such as Gaussian, and that this model is valid over the entire input space (Alpaydin, 2004). There are at a minimum four parametric approaches to developing multivariate credit-scoring systems (Altman & Saunders, 1998); linear regression, discriminant analysis, logistic regression and the probit model. In non-parametric estimation the sole assumption is that similar inputs have similar outputs (Alpaydin, 2004). Popular non-parametric approaches include; decision trees, linear programming, neural networks and k-nearest neighbours. A limited discussion on the non-parametric approaches is included. At a further date this will be expanded. For the present however, the common parametric approaches are discussed below.

3.5.2 Linear Regression

The purpose of regression is to write the numeric output (the dependent variable) as a function of the input (the independent variable) (Alpaydin, 2004). Assume a vector of inputs $x = (x_1, x_2, \ldots, x_p)$ and a real valued output y to predict. The linear regression model has the form:

$$f(x) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j \tag{18}$$

The linear model either assumes that the regression function $E(y\,|x)$ is linear, or that the linear model is a reasonable approximation (Hastie, 2002). β_j's are unknown parameters that are estimated from a set of training data $(x_1, y_1 \ldots x_N, y_N)$. Each $x_i = (x_{i1}, x_{i2}, \ldots x_{ip})^T$ is a vector of feature measurements for the ith case or instance. The most common estimation method in which to estimate the coefficients $\beta = (\beta_0, \beta 1, \ldots, \beta p)^T$ is the *least squares*:

$$\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \tag{19}$$

The biggest problem associated with linear regression is the number of assumptions that it makes (Anderson, 2007). Some of these assumptions include:

- Linearity. x and the mean of y are related in a straight-line fashion.
- Equal Variance. The variability of y around its mean is the same at every x.
- Normality. Usually it is assumed that the distribution of the error term ϵ_i is normal.
- Independence. Usually it is assumed that ϵ_i and $\epsilon_{i'}$ are independent for $i \neq i'$. That is, the residuals for two different observations on y do not "*travel together*" once their corresponding x's are taken into account.

In credit scoring, where there is a discrete output, linear regression is referred to as linear probability modelling (LPM). Ordinary linear regression has been used in credit scoring. Orgler (1970) used regression analysis in a model for commercial loans and Orgler (1971) used regression analysis to

construct a score card for evaluating outstanding loans, behavioural scoring (Hand & Henley, 1995). He found that the behavioural characteristics were more predictive of future loan quality than are the application characteristics (Hand & Henley, 1995).

3.5.3 Logistic Regression

From a credit scoring point-of-view linear regression has one obvious flaw, the right hand side of the equation can take any value from -∞ to +∞. In credit scoring only target or non-target is required (1 or 0). **Logistic regression** is a commonly used algorithm for developing credit risk scorecards. The dependent variable can be reduced to a binary outcome (target/non-target). Linear regression is used in cases where the dependent variable is continuous. Logistic regression uses a set of predictor variables to predict the probability of an outcome. This is done using a process called *maximum likelihood estimation* (MLE), which: (i) transforms the dependent variable into a log function, (ii) estimates the coefficients β; and (iii) determines changes to the coefficients to maximise the log likelihood. The end result is a regression formula of the form:

$$In\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + \epsilon \tag{20}$$

Where p_i is the probability that case or instance i is good. This implies that the probability of case i being a good is:

$$p_i = \frac{e^z}{1+e^z} \tag{21}$$

where, from (17):

$$z = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + \epsilon \tag{22}$$

Traditionally the predictor variables for each individual were used to predict a value for p_i which is compared with some critical cut off value or threshold and a decision is made (Crook *et al.*, 2007). p_i can be used in other ways. For example, p_i can be used to determine the number of cheques in a cheque book, the interest rate for a loan product or a salary multiple for a credit limit.

Logistic regression does not require linear relationships between the independent factor or covariates and the dependent, as does OLS regression,

but it does assume a linear relationship between the independents and the log odds (logit) of the dependent variable. The dependent variable is a categorical target variable that has exactly two categories (i.e., a binary or dichotomous variable) or a continuous target variable that has values in the range 0.0 to 1.0 representing probability values or proportions.

Previously, the primary disadvantage associated with Logistic regression was its computational intensiveness (Anderson, 2007). One way to reduce this computational intensiveness is to control how variable are entered into the model. Variable can be entered in the order specified by the researcher or logistic regression can test the fit of the model after each coefficient is added or deleted, called stepwise regression. There are three types of stepwise logistic regression techniques that can be used to reduce the computational intensity of model selection (Siddiqi, 2005):

1. Forward Selection selects the best fitting one feature based on the individual predictive power of each feature. It then incrementally adds features until no remaining feature have a probability score less than some significant value.

2. Backward Elimination is the opposite to Forward Selection. Features that are considered least significant are sequentially eliminated until all remaining features have a probability score under some value.

3. Stepwise is a combination of the two above techniques. Features are added and removed dynamically until a best combination is achieved.

Improvements in computer hardware have made the problem of computational intensiveness less of an issue. Wiginton (1980) was one of the earliest authors to describe the results of using logistic regression in credit scoring. Though Wiginton (1980) was not overly impressed with its performance logistic regression is now the main approach used for developing credit-scoring models (Thomas, 2000; Anderson, 2007). This is in part due to the fairly robust estimate of the actual probability provided by logistic regression based on the available information. Lawrence *et al.* (1992) use the logit model to predict the probability of default on mobile home loans. They found that payment history is by far the most important predictor of default (Altman & Saunders, 1997).

3.5.4 Discriminant Analysis

Using Discriminant Analysis, Fisher (1936) sought to identify which linear combination of variables best separates the two groups to be classified (Thomas, 2000). Discriminant analysis in its simplest form requires an analysis of a set of variables to maximise the variance between classes while minimising the variance within the class among these variables (Altman & Saunders, 1997). Discriminant analysis uses Bayes' theorem to compute the posterior probability:

$$p(xy) = \frac{p(x\,|y)p(y)}{p(x)} \tag{23}$$

Two popular forms of discriminant analysis include:

1. Linear Discriminant Analysis (LDA)

2. Quadratic Discriminant Analysis (QDA)

If the distributions of the probabilities $p(x|\,y) = 1$ and $p(x|\,y) = 0$ are multivariate normal with a common covariance matrix then LDA is used. If the covariance matrices of the populations of the goods and the bads are different then the analysis results in a quadratic discriminant function. The Bartlett test and Levene test calculate the equality of the covariance matrices.

LDA and QDA are popular classification techniques that have been successfully applied in various settings (Baesens, 2003). In a majority of the cases reported in the literature LDA appears more robust than QDA (Titterington, 1992). A major reason attributed to this is that in QDA it is necessary to estimate more parameters from the same sample and these estimates may be poor for small data sets with many inputs (Beasens, 2003)

3.5.5 Mathematical Programming

Linear programming is a well known form of Mathematical programming. Mangasarian (1965) first identified linear programming as a means of solving classification problems comprising of two groups separable by separating hyperplane. Baesens (2003, p.14) states:

> *"Linear programming (LP) is probably one of the most commonly used techniques for credit scoring in the industry nowadays".*

Typically a pre-specified cut-off point of threshold is used to separate instances which are assigned a score using weights. To take into account misclassifications, a positive slack variable is entered. A popular linear programming formulation can be represented as (Baesen, 2002):

$$min_{w,\xi} \sum_{i=1}^{N} \xi_i \tag{24}$$

subject to

$$\begin{cases} w^T x_i \geq c - \xi_i, & y_i = +1 \\ w^T x_i \leq c + \xi_i, & y_i = -1 \\ \xi_i \geq 0, & i = 1 \ldots, N, \end{cases} \tag{25}$$

Where ξ represents the vector of ξ_i values. The first set of inequalities attempt to separate the goods from the bads by assigning them a score $w^T x_i$ which is higher than the prespecified cut-off c. Similarly the second set of inequalities attempt to separate the bads from the goods by assigning them a score $w^T x_i$ which is lower than the prespecified cut-off c. In order to account for misclassifications, the positive slack variables ξ_i are entered. The aim is then to minimize the misclassifications by minimizing the sum of the slack variables ξ_i. Linear programming methods can model domain knowledge on a priori bias by including additional constraints (Baesans, 2003). This allows linear programming to easily include a selected bias into scorecard development. For example, X_1 is the binary variable of being under 30 years of age or not and X_2 is the binary variable of being over 65 years of age or not. If one know a piori that the under 30s have a larger impact on the score than over 65s then for the score for under 30s to be higher than over 65s the constraint $w_1 \geq w_2$ is simply added to equation (25). LP methods have been shown to empirically underperform statistical regression models.

3.5.6 Neural Networks

A neural network (or an artificial neural network, ANN) (Haykin, 1999) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information (Tsai & Wu, 2008). ANNs can be used to simulate the non-linear relationship in complicated data (Zhang *et al.*, 2007). The information processing system of an ANN is composed of a large number of highly interconnected processing elements (neurones) working in unison to solve specific problems (Tsai & Wu, 2008).

Neural networks learn by examples. That is, neural networks learn by experience and generalise from previous experiences to new ones which in turn facilitates decision making.

ANNs have been successfully applied to bankruptcy prediction (Min & Lee, 2005). The multilayer perceptron (MLP) is the most common type of ANN (Tsai & Wu, 2008). In credit scoring MLP is the most frequently used neural network architecture (West, 2000). A MLP consists of three layers of of units or neurones: input layer, hidden layers and output layers. As shown in Figure 4, a layer of input units is connected to a layer of hidden units which is then connected to a layer of output units. The activity of the input layer represents the raw information that is fed into the network (Tsai & Wu, 2008). The activity of each hidden unit is determined by the activities of the input units and the weights on the connections between the input and the hidden units (Tsai & Wu, 2008). The behaviour of the output units depends on the activity of the hidden units and the weights between the hidden and output units (Tsai & Wu, 2008). Backpropagation is an algorithm used to identify appropriate weights for the connections between the layers in the network (Witten & Frank, 2005). The backpropagation algorithm is the most popular learning algorithm, is adopted to perform steepest descent on the total mean squared error (MSE) (Zhang *et al.*, 2007). Given an initial weights and threshold, a set of inputs consisting of historical repayment data and loan default data are presented to a network. The MSE determines the error between the output pattern and the target pattern and adjustment to weights and threshold (Zhang *et al.*, 2007). Structural matches are found that coincide with defaulting firms and then used to determine a weighting scheme to forecast PD.

In general, ANNs has a difficulty in explaining the prediction results due to a lack of explanatory power and also suffers from difficulties with generalisation because of overfitting (Hao, 2006). This hampers its deployment in countries where it is necessary, by legislation, to provide a rejected candidate with a reason for rejection. In addition, comparable to other classification techniques, ANNs require greater time and effort to construct the optimal architecture. Desai *et al.* (1996) investigated neural networks, linear discriminant analysis and logistic regression for scoring credit decision. Their study concluded that neural networks outperformed linear discriminant analysis in classifying loan applicants into good and bad credits. Logistic regression performed comparatively to neural networks.

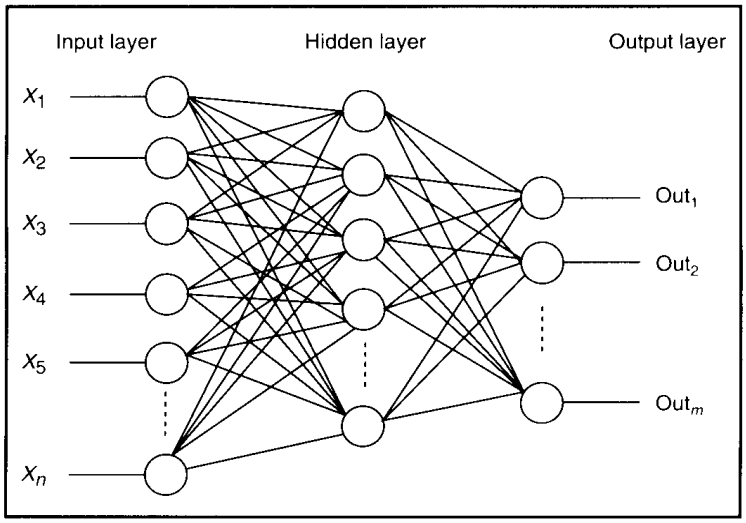

Figure 4: The three-layer neural network

3.5.7 Support Vector Machines

A support vector machine (SVM) is a machine learning technique based on Vapnik's Statistical Learning Theory (SLT) (Vapnik, 1995). SLT provides SVMs with a strong theoretical foundation and SVMs have achieved strong empirical success within the research community. A two-class SVM distinguishes between two classes in a given data set by fitting a hyperplane that maximally divides both classes (Senf *et al.*, 2002). All objects lying on one side of this optimal separating hyperplane are labelled as -1, and all objects lying on the other side are labelled as +1. The objects that lie closest to the optimal separating hyperplane are called support vectors.

This works well for data sets that are linearly separable (Senf *et al.*, 2002). In situations where the data is not linearly separable, a linear SVM may still be used when it allows for a certain amount of errors. For objects that are not easily separable even with the provision for a certain amount of errors, the objects can be projected onto a higher dimensional feature space using kernels (Senf *et al.*, 2002). This is done by introducing a slack variable, ξ, and an upper bound C for the number of errors (Senf *et al.*, 2002). The formula to be minimised takes on the commonly used form:

$$J(w, b, \xi) = \frac{1}{2}(w * w) + C \sum_{i=1}^{n} \xi_i \tag{26}$$

subject to:

$$y_i \left[w * x_i + b \right] \geq 1 - \xi_i, \quad \xi \geq 0 \tag{27}$$

Figure 5 illustrates the use of the slack variable and the use of a kernel function to identify the optimal separating hyperplane. According to Vapnik's (1995) original formulation, w represents the weight vector and b the bias. The slack variable ξ is also known as the soft margin. A kernel is a function that takes the original instances and several parameters, and increases their dimensionality (Senf *et al.*, 2002). A good choice of kernel function and corresponding parameters will allow the data to then be separable by a hyperplane. Examples of kernel functions include polynomial, radial basis function (RBF) and sigmoid. The choice of kernel function is largely application-dependent and is the most important factor in SVM applications (Huang *et al.*, 2004).

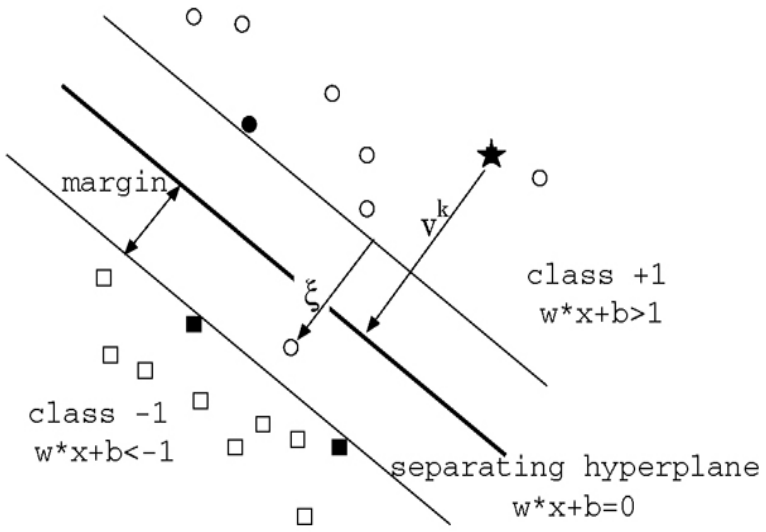

Figure 5: Two Class Support Vector Machine

Recently the SVM approach has been introduced to several financial applications such as credit rating, time series prediction, bankruptcy prediction and insurance claim fraud detection (Fan & Palaniswami, 2000; Gestel *et al.*, 2001; Tay & Cao, 2001; Viaene *et al.*, 2002; Kim, 2003; Huang, Chen *et al.*, 2004; Min & Lee, 2005). In these studies it was reported that SVMs performed comparatively to and in some cases outperformed other classifiers including Artificial Neural Networks, Case-based Reasoning, multiple

discriminant analysis and logistic regression in terms of generalisation performance.

3.6 Non-target Data - Class Imbalance

Many conventional classifiers rely on a more or less equal balance of both labelled positive and negative examples to build a classifier (Japkowicz *et al.*, 1995; Li & Liu, 2005). As previously discussed, relatively few negative examples exist in low default portfolios. In machine learning terminology the low default portfolio problem is described as class imbalance. A class imbalance problem occurs when the difference between the number of instances belonging to each class is so large that the classifier experiences difficulties with learning the concept related to the minority class.

Barandela *et al.* (2003) describe class imbalance as:

> *"A set of examples or training set (TS) is said to be imbalanced if one of the classes is represented by a very small number of cases compared to the other classes".*

However, class imbalance raises a number of issues. There is no clear consensus on what constitutes a class imbalance. Chawla (2003) states:

> "*A data set is imbalanced if the classes are not approximately equally represented".*

This differs sharply from Barandela *et al.* (2003):

> *"a very small number of cases compared to the other classes".*

In general however, we will rely on Wang & Japkowicz (2008):

> *"A data set is imbalanced if the number of instances in one class greatly outnumbers the number of instances in the other class".*

Learning from unbalanced training sets is one of the problems in supervised learning (Batista *et al.*, 2000). Often the classifier has respectable classification accuracy for the majority class, but its accuracy for the minority class is poor. The performance of the algorithm used degrades significantly if the data set is imbalanced (Japkowicz & Stephen, 2002). In very imbalanced domains, most standard classifiers will obtain higher predictive accuracies

for the majority class than that of the minority class (Wang & Japkowicz, 2008). Classes with fewer examples in the training set have a lower prior probability and a lower error cost. This raises difficulties when true error cost of the minority class is greater than is implied by the distribution of examples in the training set (Maloof, 2003).

Approaches for addressing class imbalance can be divided into two main categories (Wu & Chang, 2003; Garcia *et al.*, 2007; Wang & Japkowicz, 2008).

1. Balance the data set
2. Modify the classifier

Other areas of research on the topic of class imbalance include metrics used to measure performance. Batista et al., (2000) show that the widely used error rate and accuracy used for measuring classifier performance is misleading for unbalanced data sets. (Provost & Fawcett , 1997; Fawcett, 2006) use a Receiver Operating Characteristic (ROC) curve to measure classifier performance. Another topic that has been the focus of a number of studies is class imbalance data complexity characteristics. These studies suggest that poor classifier performance on class imbalance data sets can also be attributed to factors such as the size of data set (Orriols & Bernardo, 2005), distribution of data within each class (Japkowicz, 2001) and small disjuncts (Weiss, 2003; Japkoweicz & Jo, 2004; Prati *et al.*, 2004). These topics are outside the scope of this literature review and will be discussed at a later stage. Instead the review will focus on the aforementioned approaches to class imbalance.

3.6.1 Balance the Data Set

Under-sampling and over-sampling are two methods for balancing a data set. Under-sampling consists of eliminating elements of the over-sized class until it matches the size of the other class (Japkowicz & Stephen, 2002). Under-sampling (randomly or selectively) the majority class while keeping the minority class is the simplest way (Kubat & Matwin, 1997). This method results in information loss for the majority class (Wang & Japkowicz, 2008). In over-sampling instances of the minority class are duplicated. Although over-sampling does not lose any information about the majority class, an unnatural bias is introduced in favour of the minority class (Wang & Japkowicz,

2008). Another disadvantage of this method is that it creates noise which could result in the loss of classifier performance. Both methods have been criticized for altering the original class distribution (Garcia *et al.*, 2007).

A lot of research conducted in the area of class imbalance attempts to improve classification performance through data sampling techniques (Drummond & Holte , 2003; Maloof, 2003; Barandela *et al.*, 2004; Han *et al.*, 2005). Some researchers have highlighted the inadequacies of under-sampling and over-sampling methods. Barandela *et al.* state, (2003):

> *"Replicating the minority class to eliminate imbalance in the TS* [**Training Set**] *does not add new information to the system. Moreover, working in that direction means to worsen the known computational burden of some learning algorithms, such as the NN rule and the Multi-Layer Perceptron... downsizing the majority class can result in throwing away some useful information".*

3.6.2 Modifying the Classifiers

As an alternative to the disadvantages presented by resampling techniques, the imbalance problem can be addressed from an algorithmic standpoint. Adapting existing algorithms and techniques can fall into the following categories (Garcia *et al.*, 2007):

- Cost-sensitive learning,
- Classifier ensembles,
- Classifier biasing
- One-class classifiers

Cost-sensitive Learning The main objective in cost sensitive learning is to minimise the cost of misclassification (Garcia et al., 2007). Given a specification of costs for correct and incorrect predictions, an object should be predicted to have the class that results in the lowest expected cost, where the expectation is computed using the conditional probability of each class given the object (Elkan, 2001). A confusion matrix C, as illustrated in Table 2, is used to measure the cost of predicting that an example belongs to class i when in fact it belongs to class j (Elkan, 2001). The cost matrix

rows correspond to alternative predicted classes, while columns correspond to actual classes, i.e. row/column = i/j= predicted/actual.

Table 2: Cost Matrix

-	actual negative	actual positive
predict negative	$C(0,0) = c_{00}$	$C(0,1) = c_{01}$
predict positive	$C(1,0) = c_{10}$	$C(1,1) = c_{11}$

Some studies assign specific costs to the classification errors for positive and negative examples (Gordon & Perlis, 1989; Domingos, 1989). Conceptually, the cost of labelling an example incorrectly should always be greater than the cost of labeling it correctly (Elkan, 2001). Japkowicz and Stephen (2002) propose the use of non-uniform error costs defined by means of the class imbalance ratio present in the data set. Refer to Elkan (2001) for further details on cost-sensitive learning.

Classifier Ensembles Ensembles consist of a set of individually trained classifiers whose decisions are combined when classifying new objects. Research has demonstrated that the predictive accuracy of a combination of independent classifiers out performs that of the single best classifier (Garcia *et al.*, 2007). For the class imbalance problem, ensembles have been used to combine several classifiers whose training sets have used under-sampling and over-sampling techniques (Garcia *et al.*, 2007).

Classifier Biasing This process involves biasing the discrimination based process so as to compensate for class imbalance (Garcia *et al.*, 2007). Garcia *et al.*, (2007) conducted a study of existing classifier biasing work, these include; Pazzani *et al.* (1994), who assigned different weights to the instances of the different classes, Ezawa *et al.*, (1996) bias the classifier in favour of certain attribute relationships, Brandela *et al.* (2003) propose a weighted distance function to be used in the k-nearest neighbour classification. Weights are assigned to the respective classes and not to the individual instances (Garcia *et al.,* 2007).

The final approach, OCC, is discussed in it entirety in a separate section.

3.7 Conclusion

This section considered the traditional form of the Classification problem, two-class classification. Definitions were provided and discussed in Section 3.1. Sections 3.2 and 3.3 examined the theoretical formulation of classification. The concepts involved in Model Selection and Model Estimation were reviewed. A concise description of classifier evaluation techniques was listed in Section 3.4. Of particular interest to one-class-classification is the ROC curve. In Section 3.5 existing models used in credit scoring techniques were listed and discussed. In Section 3.6 the notion of biased or unbalanced data sets was introduced. This final subsection marked the shift in emphasis from two-class classification to OCC, which is reviewed in Section 4.

4 One-Class Classification

The goal of one-class classification (OCC) (Juszczak *et al.*, 2008) is to take one of the classes, the target class, and distinguish it from all other possible objects, called non-targets (Juszczak *et al.*, 2008). OCC is often called outlier detection (Ritter & Gallegos, 1997), novelty detection (Bishop, 1994), concept learning (Japkowicz, 1999), single-class classification (El-Yaniv & Nisenson, 2007) and data description (Tax & Duin, 2000). OCC approaches can be applied to problems that cannot easily be addressed by more conventional approaches. One example of such problems is when non-target objects are very expensive or difficult to obtain (Japkowicz *et al.*, 1995). Applications that utilise OCC or novelty detection include:

- Loan application processing - Identify potential defaults and thus prevent loss (Hodge & Austin, 2004),
- Fraud detection - Detect fraudulent applications for credit cards (Chan & Stolfo, 2001),
- Intrusion detection - to detect unauthorised computer network access (Hodge & Austin, 2004),
- Activity Monitoring - Detect suspicious trades in the equity markets (Hodge & Austin, 2004),
- Network Performance - Detect network bottlenecks in the performance of computer networks (Hodge & Austin, 2004),
- Fault diagnosis - Detect fault in mechanical devices such as motors, aeronautical instruments (Markou & Singh, 2003a; Weiss & Hirsch, 2000),
- Satellite image analysis - Identify oil spills (Kubat *et al.*, 1997),
- Motion segmentation - Detect image features moving independently of the background (Hodge & Austin, 2004),
- Medical condition monitoring - such as heart-rate monitors (Hodge & Austin, 2004),
- Pharmaceutical research - identifying novel molecular structures (Tarassenko *et al.*, 1999; Hempstalk & Frank, 2008),

- Hand written digit recognition (Tax & Duin, 1998)
- Detecting mislabelled data in a training data set (Hodge & Austin, 2004).

In all of these examples little or no non-target data is available. Some financial institutions attempt to address the low-default portfolio problem by accepting all credit applicants over a short period of time. Financially this is costly as a higher proportion of loan-defaulters are accepted. Instead of this practice, financial institutions use reject inference which results in fewer loan defaults and helps gives rise to the low-default portfolio. In Figure 6 the feature space is composed of an applicant's income and the requested loan amount. It is kept at 2-features in order to represent the classification problem in 2-dimensions. In this example an OCC classifier places a boundary around the target data. If the distance of a specific instance to the target class is within a specified threshold it is accepted as part of the target class, otherwise it is rejected as non-target data. There are two distinct features of all OCC classifiers. The first is *"a measure of the distance d(z) or resemblance (or probability) p(z) of an object z to the target class"* (Tax p.57, 2001). The second consideration is a threshold, θ, on this distance or resemblance (Tax, 2001). New objects are accepted as part of the target class when the distance to the target class is smaller than the threshold or when the resemblance is larger than the threshold (Tax, 2001).

From a broad perspective, much of the academic research into classification focuses on theoretical statistical classifiers that assume well defined, well sampled, balanced and stationary data distributions (Tax, 2001). This is somewhat contrary to the data obtained by actual classifiers in a live-environment which may contain noise, missing attributes, outliers/unbalanced and non-stationary data distributions (Tax, 2001). One-class classifiers represent an attempt at addressing this gap. OCC has received much attention from the machine learning and pattern recognition communities (El-Yaniv & Nisenson, 2007). However, El-Yaniv and Nisenson (2007) state that *"The extensive body of work on SCC* [**Single Class Classification**]*, which encompasses mainly empirical studies of heuristic approaches, suffers from a lack of theoretical contributions and few principled (empirical) comparative studies of the proposed solutions"*. From this observation the author is of the opinion that there is the scope to conduct a benchmark experiment comparing OCC methods as well as popular two-class classification approaches.

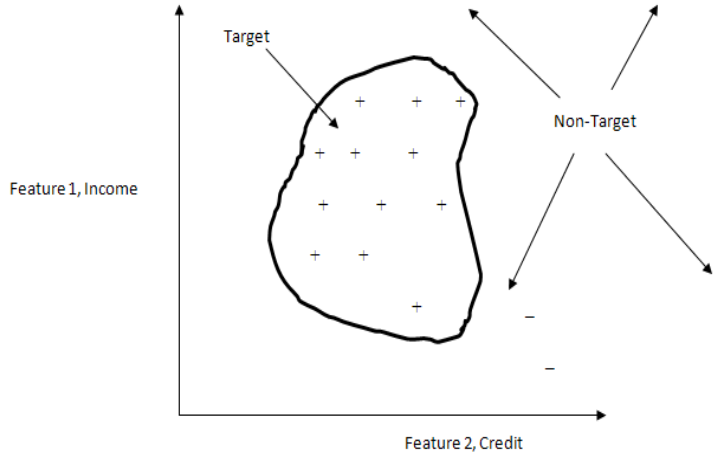

Figure 6: Low-Default Portfolio Problem. Loan defaulters are represented by -. Successfully repaid loans are represented by +.

4.1 Problem Formulation

According to Tax (2001, p14) *"the problem in one-class classification is to make a description of a target set of objects and to detect which (new) objects resemble this training set"*. OCC attempts to maximise the detection of true novel data while at the same time minimising the false positives (Markou & Singh, 2003b). This is echoed by Tax (p.58, 2001) who states:

> *"The most important feature of one-class classifiers is the trade off between the fraction of the target class that is accepted, fT+, and fraction of outliers that is rejected, fO-".*

Identifying an outlier is a subjective exercise. There is no universally accepted definition of what constitutes an outlier. Grubbs (1969) defined an outlier as:

> *"An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs."*

A further definition is provided by Hawkin (1980):

> *"an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism."*

In this literature review outliers are not considered as noise points lying outside a set of defined clusters. Separating noise and outliers is outside the scope of this review. Outliers arise because of error (human or instrument), natural deviations in populations, fraudulent behaviour or faults in systems (Hodge & Austin, 2004). In the case of loan defaults, outliers can arise because of changes in the behaviour of the system, in the sense that a person has missed a scheduled loan repayment. Different systems respond to outliers in different ways. If an outlier is due to a typographical error by an entry clerk then the clerk can be notified and simply correct the error so the outlier will be restored to a normal record (Hodge & Austin, 2004). In the case of credit scoring, a bank manager is alerted and steps to minimise exposure to loss are implemented.

A problem faced by most classification algorithms is that they fail at automatically detecting novel classes because they are discriminators rather than detectors (Markou & Singh, 2003b). This can be attributed to the fact that many classification algorithms use open decision boundaries, such as a hyperplane, to distinguish between classes and fail to decide when a feature set does not represent any known class (Markou & Singh, 2003b). Hodge and Austin (2004) offers three fundamental approaches to the problem of outlier detection.

Type 1 - *"Determine the outliers with no prior knowledge of the data"* (Hodge & Austin, 2004, p.88). This approach is essentially the same as unsupervised clustering. The data is processed as a static distribution and the most remote points are flagged as potential outliers. We have previously stated that unsupervised learning is outside the scope of this review. However future research (experiments) will be conducted to assess the applicability of unsupervised learning methods.

Type 2 - Model both normality and abnormality. This approach requires pre-labelled data and is akin to supervised learning. We have already dismissed this approach as *"Classification algorithms require a good spread of both normal and abnormal data"* (Hodge & Austin, 2004, p.89). As previously discussed, the root of the low default portfolio problem is down to class imbalance were one class is underrepresented. As a result of this it is difficult to obtain a good spread of abnormal data. Type 2 methods are unsuitable

for our research area.

Type 3 - *"Model only normality or in very few cases model abnormality"* (Hodge & Austin, 2004, p.90). This approach uses labelled data but only infers from target data. *"It aims to define a boundary of normality"* (Hodge & Austin, 2004, p.90). An object is classed as target data if it lies within the boundary and as non-target data otherwise. *"This approach requires no non-target data"* (Hodge & Austin, 2004, p.90). This approach appears to match the requirements of a OCC model.

The approach to addressing the low-default portfolio problem gains little traction by splitting OCC into the broad categories of Type 1 and Type 2. Supervised and Unsupervised learning are the broadest categories of machine learning approaches available. Type 3 offers some light. The notion of defining a boundary or density around the target data is expanded by Tax (2001). OCC approaches are discussed in detail a little later.

Before a further discussion on approaches can take place, it is necessary to discuss the relative considerations when selecting an appropriate methodology for OCC.

4.2 One-Class Classification Considerations

There exist several important considerations or issues related to one-class classification (OCC). Combining Tax (2001) and Markou & Singh (2003a) these issues can be summarised as:

1. Principle of robustness and trade-off - A one-class classifier should offer robust performance on test data that minimises the exclusion of target data while maximising the exclusion of non-target data (Tax, 2001; Markou & Singh, 2003a).

2. Principle of uniform data scaling - All test and training data should lie within the same range after normalisation (Roberts & Tarassenko, 1994).

3. Principle of parameter minimisation - An OCC classifier should aim to minimise the number of parameters that are set by the user (Markou & Singh, 2003a). Consider this as the ease of operation by the user. Tax (2001) also calls this the *Magic Parameters*.

4. Principle of generalisation - *"The one-class classification classifier should be able to generalise without confusing generalised information as novel"* (Tax & Duin, 1998). The generalisation performance of one-class classifiers can be measured using three criteria (Moya *et al.*, 1993). First, within-class generalisation measures the classifiers performance on non-trained known classes (Markou & Singh, 2003b). Between-class generalisation indicates the performance on near-known class objects from other classes (Markou & Singh, 2003b). Finally, out-of-class generalisation indicates the classifiers' performance on unknown classes (Markou & Singh, 2003b).

5. Principle of independence - The performance of the OCC classifier should be independent of the number objects available and the amount of features used. The classifier should display a reasonable performance in relation to a low number of samples and noise (Markou & Singh, 2003a).

6. Principle of adaptability - the information on instances labelled as non-target data during test should be used during retraining (Saunders & Gero, 2000).

7. Computational complexity and Storage - The computational complexity of a OCC classifier should be kept to a minimum (Markou & Singh, 2003a). Typically training is performed off-line, as a result training costs are not a major consideration. However, training costs become a factor when it is necessary to adapt to a changing environment (Tax, 2001). Changes in the credit scorecard environment include market conditions, population drift or some low probability high impact event (for example, a natural disaster).

8. Incorporation of known outliers - Outliers can be used to tighten the description of the target data. It should be possible to add a parameter to the one-class classifier to regulate the trade-off between a target and outlier error (Tax, 2001).

4.3 One-Class Classification Approaches

This subsection identifies the main approaches to OCC. Related studies include Barnett and Lewis (1994) and Rousseeuw and Leroy (1996) who ex-

amined statistical approaches to identifying outliers. Tax (2001) describes a range of statistical, neural network-based and machine learning approaches to OCC. Tax's list *"does not pretend to be an exhaustive enumeration of all possibilities"* (Tax p.64, 2001). Markou and Sing (2003a, 2003b) describe both a statistical approach and neural network approach to OCC. A similar approach is detailed by Hodge and Austin (2004). Within this literature review the method of describing OCC approaches is split into three main sections; Statistical, Neural Networks and Machine Learning.

4.3.1 Statistical Approach

Statistical approaches are based on modelling the statistical properties of the training data to estimate if a test sample comes from the same distribution. The techniques used vary in their complexity (Oddin & Addisson, 2000). The simplest method consists of constructing a density function for data of a known class. The two main methods used to estimate the probability of a density function are parametric and non-parametric methods (Desforges *et al.*, 1998).

Parametric: The parametric approach assumes that the data originates from a family of known distributions, such as the normal (Gaussian) distribution and certain parameters are calculated to fit this distribution (Markou & Singh, 2003a). Parametric models can be rapidly evaluated for new instances and are suitable for large data sets (Hodge & Austin, 2004). A parametric model grows only with data complexity and not the number of instances (Hodge & Austin, 2004). Their applicability is limited by enforcing a pre-selected distribution model to fit the data (Hodge & Austin, 2004). Estimating the density of the training data and setting a threshold on this density represents a straightforward method of obtaining a one-class classifier (Tarassenko *et al.*, 1995). Of particular importance is the trade-off between the recognition rate (error probability) and the proportion of data rejected (reject probability) (Hansen *et al.*, 1997). The reject rate is necessary to safeguard against overfitting caused by noise and uncertainty.

Tax describes the Guassian model as a method for rejecting outliers based on their density distribution. Gaussian mixture modelling (GMM) allows for a more flexible density method by extending the normal distribution to a mixture of Gaussians (Duda & Hart, 1973). An optimisation algorithm such as Expectation-Maximisation (EM) is used to select the parameters of the

model. GMM requires a large number of samples to train the model (Markou & Singh, 2003a).

Other methods include minimum volume ellipsoid estimation, or MVE (Rousseeuw & Leroy, 1999). MVE fits the smallest permissible ellipsoid volume around the majority of the data distribution model (Hodge & Austin, 2004). Convex peeling is also another parametric method that works by peeling away the instances on the boundaries of the data distribution's convex hull (Rousseeuw & Leroy, 1996). This results in peeling away the outliers. Both methods are only applicable for low dimensional data (Hodge & Austin, 2004).

An extensive study of parametric outlier detection methods is detailed in Markou & Singh (2003a).

Non-Parametric: In non-parametric approaches no assumptions on the statistical properties of the data are made (Markou & Singh, 2003a). The overall form of the model structure is not defined *a priori*, instead it is derived from the data as are the parameters of the model. The k-nearest neighbour algorithm is a non-parametric technique for estimating the density function of data (Oddin & Addisson, 2000). A width parameter is set as a result of the position of the instance in relation to other instances by considering the k-nearest patterns in the training data to the test pattern (Markou & Singh, 2003a). A drawback with this technique is that for large datasets a larger number of computations have to be performed. The nearest neighbour method, NN-d, can be derived from a local density estimation by the nearest neighbour classifier (Duda & Hart, 1973). Hellman (1970) used the nearest neighbour (NN) classifier for rejecting patterns with higher risk of being misclassified. The advantage of NN-d is that it avoids explicitly estimating the complete density of the data and uses the first nearest neighbour. This method can be termed as a *boundary method* and follows one of the main ideas in learning theory, that when only a limited amount of data is available one should avoid solving a too hard intermediate problem. Roth (2005) argues that while, theoretically, this line of reasoning seems appealing, in practical applications it leads to problems. The restriction of estimating only a boundary renders the task of deriving *"a formal characterisation of non-target data without prior assumptions on the expected fraction of outliers or even on their distribution"* (Roth p.1169, 2005) impossible. In practice, however, it is difficult to justify any such prior assumptions. Roth (2005)

states:

> *"The fundamental problem of the one-class approach lies in the fact that outlier detection is a (partially) unsupervised task which has been 'squeezed' into a classification framework. The missing part of information has been shifted to prior assumptions which can probably only be justified, if the solution of the original problem was known in advance".*

Another shortcoming of the *boundary method* is the fact that they rely heavily on the distances between objects and as a result tend to be sensitive to scaling of the features.

Further studies of non-parametric based novelty detection classification methods appear in Markou and Singh (2003a), Hodge and Austin (2004), Agyemang *et al.* (2006), Bakar *et al.* (2006), Patcha and Park (2007) and Chandola *et al.* (2008).

4.3.2 Neural Networks

Typically, neural networks make no a priori assumptions on the properties of data. They generalise well to unseen patterns and are capable of learning complex class boundaries (Hodge & Austin, 2004). Training can be a lengthly process that involves traversing the training set numerous times in order to allow the network to model the data correctly. Many neural networks are susceptible to the curse of dimensionality, though less so than statistical methods (Hodge & Austin, 2004). Despite the difficulties with neural network retraining and the vast amount of parameter settings, neural networks are important novelty detectors (Markou & Singh, 2003b). Not all neural network approaches are examined or mentioned in the following section. Markou and Singh (2003b) offer an in-depth topology of neural network based approaches to novelty detection.

Supervised Neural Networks: The learning process of supervised neural networks is driven by the class labels. To correctly classify input, the neural network uses the class labels to adjust its weights and thresholds. Multilayer perceptrons are the best known and most widely used class of neural networks. Devising methods of novelty detection is a challenging task as MLPs do not generate closed class boundaries. This can cause interference

between the generalisation property of the network and its ability to detect novelties (Moya *et al.*, 1993).

According to Bishop (Bishop, 1994) one of the most important sources of error in neural networks arises from novel input data. A network that is trained to discriminate between a number of classes coming from a set of distributions will be confused when it encounters data coming from an entirely new distribution (Hodge & Austin, 2004). The novelty detection is implemented by estimating the density of the training data, thus modelling its distribution and checking whether an input data point comes from this distribution (Markou & Singh, 2003b). The density estimation is done by using either a kernel-based estimator or by a semi-parametric estimator constructed from a Gaussian Mixture Model (Markou & Singh, 2003b). A threshold is then placed on the output of the network and low confidence indicates an outlier.

Other supervised neural networks used for novelty detection include an auto-associative neural network (Japokowicz *et al.*, 1995), Hopfield networks (Hopfield, 1982) and the supervised radial basis function (RBF) network.

Unsupervised Neural Networks: Unsupervised neural networks are used when pre-classified data is unavailable. Unsupervised neural networks contain nodes which compete with one and other to represent portions of the data set (Hodge & Austin, 2004). Training data is necessary in order for the network to learn. Neural networks work on the assumption that related vectors have common feature values that can be identified to topologically model the data distribution (Hodge & Austin, 2004).

Most self-organising maps (SOM) (Kohonen, 1997) based approaches are similar to statistical clustering. A threshold is set on some form of cluster membership value to determine whether a sample belongs to a cluster or not. SOMs are similar to k-means clustering with k equivalent to the number of nodes (Hodge & Austin, 2004). However, SOMs employ a user-specified global threshold distance whereas k-means autonomously determines the boundary during training allowing local setting of the radius of normality as defined by each cluster (Hodge & Austin, 2004).

4.3.3 Machine Learning

The following subsection examines a machine learning technique, support vector machines, applied to the OCC problem.

Support Vector Machines: As previously discussed in Section 3.6.5 support vector machines (SVMs) are based on the concept of determining optimal hyperplanes for separating different classes (Vapnik, 1998).

Tax and Duin create a Support Vector Data Description (SVDD) (Tax & Duin, 1999, Tax & Duin, 1999b) to address the problem of OCC by distinguishing between the class of objects that are represented by the training set and all other possible objects in the object space (Markou & Singh, 2003b). In the SVDD a model f(x;w) is defined in such a way that instead of searching for a optimal hyperplane it searches for a closed boiundary around the data, a hypersphere. The hypersphere is characterised by a centre a and radius R (Figure 7). Using a hypersphere and based on its minimum radius, almost all of the objects in the data set are encompassed (Markou & Singh, 2003b). An object is rejected if its distance from the centre of the sphere is larger than the radius of the sphere (Markou & Singh, 2003b). Objects lying outside of the sphere are called slack variables, i.e. outliers. The optimal hypersphere implements a trade off between two conflicting goals: (i) the volume of the hypersphere and the number of target objects included; and (ii) the size of the radius and the number of slack variables (Cherkassky & Mulier, 2007).

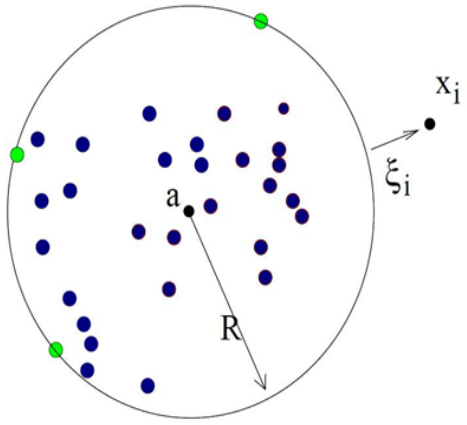

Figure 7: Support Vector Data Description (Tax & Duin, 1999, Tax & Duin, 1999b)

To find the optimal hypersphere an error function F is formulated so that

$$F(R, a) = R^2 \tag{28}$$

With the constraints that all of the training data x_i are within that R^2

$$\|x_i - a\|^2 \leq R^2, \quad \forall_i \tag{29}$$

In order to allow for the possibility of outliers in the training set, the distance from x_i to the centre a should not be less than R^2 but larger distances should be penalised. Therefore, slack variables $\xi_i \geq 0$ may be introduced and the error minimisation problem can be rewritten as

$$F(R, a, \xi) = R^2 + C \sum_i \xi_i \tag{30}$$

where the variable C gives the trade-off between the volume of the sphere (or simplicity) and the number of target objects rejected (misclassification errors). This must be minimised under the constraints

$$\|x_i - a\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0, \quad \forall_i \tag{31}$$

Constraints (31) can be incorporated into (30) by using Lagrange mulitpliers. The resultant Lagrangian formula is

$$L(R, a, \alpha, \gamma, \xi) = R^2 + C \sum_i \xi_i - \sum_i \alpha_i \times \left\{ R^2 + \xi_i - \left(\|x_i\|^2 - 2a \cdot x_i + \|a\|^2 \right) \right\} - \sum_i \gamma_i \xi_i \tag{32}$$

with the Lagrange multipliers $\alpha_i \geq 0$ and $\gamma_i \geq 0$. L should be minimised with respect to R, ξ, a and maximised with respect to α and γ with the constraints

$$\sum_i \alpha_i = 1 \tag{33}$$

$$a = \frac{\sum_i \alpha_i x_i}{\sum_i \alpha_i} = \sum_i \alpha_i x_i \tag{34}$$

$$C - \alpha_i - \gamma i = 0 \tag{35}$$

Since $\alpha_i \geq 0$ and $\gamma_i \geq 0$, the variables can be removed from Equation (35) and the constraint $0 \leq \alpha_i \leq C$ can be used.

Rewriting Equation (32) and resubstituting Equations (33, 34, 35) give to maximise with respect to α_i

$$L = \sum_i \alpha_i \left(x_i \cdot x_i\right) - \sum_{ij} \alpha_i \alpha_j \left(x_i \cdot x_j\right) \tag{36}$$

with constraints $0 \leq \alpha_i \leq C$, $\sum_i \alpha_i = 1$

Equation (34) states that the centre of the sphere is a linear combination of data objects, with weight factors α_i which are obtained by optimising Equation (36). A data object is located on the boundary of the sphere when equality in Equation (31) is achieved. For these data objects the coefficients α_i will be non-zero and are called support objects. Only these objects are needed in the description of the sphere. The radius R of the sphere can be calculated using the distance from the centre of the sphere to a support vector with a weight less than C. Objects for which $\alpha_i = C$ have hit the upper bound in Equation (35) and are outside the sphere. These support vectors are considered outliers.

To determine whether a test point z is within the sphere, the distance to the centre of the has to be measured. A test object z is accepted when this distance is less than the radius. Expressing the centre of the sphere in terms of the support vectors, z is accepted when

$$\left(z \cdot z\right) - 2 \sum_i \alpha_i \left(z \cdot x_i\right) + \sum_{ij} \alpha_i \alpha_j \left(x_i \cdot x_j\right) \leq R^2 \tag{37}$$

SVDD with kernels: The previous method only computes a sphere around the data in the input space. Rarely is the data so spherically distributed that one can expect a very tight description. As the problem is stated completely in terms of inner products between vectors (Equations (37) and (36)), the method can be more flexible, analogous to Vapnik (1995). Inner products of the objects $(x_i \cdot x_j)$ can be replaced by the kernel function $K(x_i, xj)$, when this kernel $K(x_i, xj)$ satisfies Mercer's theorem. This implicitly maps the objects x_i into some feature space and when a suitable feature space is chosen, a better, more tighter description can be obtained. No explicit mapping is required, the problem is expressed completely in terms of $K(x_i, x_j)$. All inner products $(x_i \cdot x_j)$ are replace by a proper $K(x_i, x_j)$ and the problem of finding a data domain description is given by

$$L = \sum_i \alpha_i K(x_i, x_i) - \sum_{ij} \alpha_i \alpha_j K(x_i, x_j) \tag{38}$$

with the constraints $0 \leq \alpha_i \leq C,\ \sum_i \alpha_i = 1$. A test object z is accepted when

$$K(z,z) - 2\sum_i \alpha_i K(z,x_i) + \sum_{ij} \alpha_i \alpha_j K(x_i,x_j) \leq R^2. \quad (39)$$

Different kernel functions K result in different description boundaries in the original input space. The problem is to find a suitable kernel function $K(x_i, x_j)$. Not all kernels perform equally well for the SVDD. Polynomial kernels result in less compact boundary representations (Cherkassky & Mulier, 2007). Additionally data points with the highest norms have a higher chance of becoming support vectors (Cherkassky & Mulier, 2007). With a radial basis function, the width kernel controls the flexibility of the boundary (Cherkassky & Mulier, 2007).

A drawback with SVDD is that it ignores unlabelled data that may be readily available (Peng *et al.*, 2003). Partially supervised classification, also referred to as semi-supervised learning, provides a solution to OCC problems that utilise unlabelled data. Initially it is assumed that all unlabelled examples belong to a single class, the Expectation-Minimisation (EM) algorithm is applied to refine the assumption. Peng *et al.*, (2003) also offer a solution to this problem by constructing a contrast classifier that discriminates between labelled and unlabelled data. The output of the contrast classifier is a measure of difference, or contrast in density of a given data point between labelled and unlabelled data (Peng *et al.*, 2003).

Scholkopf and Smola (2002) provide an additional method of geometrically enclosing a fraction of the target data. This is achieved by defining a relationship between a hyperplane and the origin. Single-class SVM uses a hyperplane to separate the training data from the origin with a maximal margin (Cherkassky & Mulier, 2007). The hyperplane separates the surface region containing objects from the region containing no data.

The hyperplane is maximally distant from the origin with all data points lying on opposite sides of the origin (Markou & Singh, 2003b). Criticism of the method is available in Campbell and Bennett (2001) and Manevitz and Yousef (2007).

Additional studies of OCC methods are available in Campbell and Bennet (2001), Manevitz and Yousef (2007), Diehl and Hampshire (2002), Ratsch *et al.* (2002) and Davy and Godsill (2002).

Tax and Duin (2001) suggest creating artificial outliers uniformly in and around the target class. The authors used a dimensional Gaussian distri-

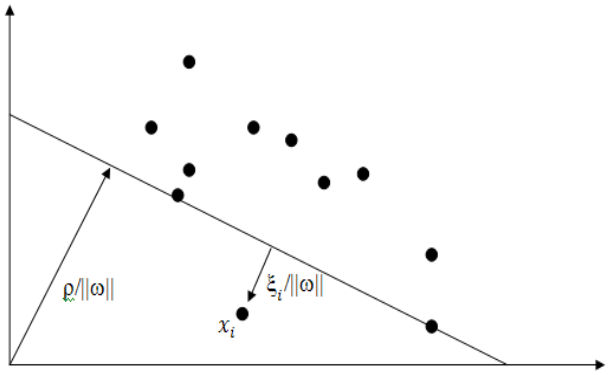

Figure 8: Single class SVM (Scholkopf & Smola, 2002)

bution for creating the outlier data and indicate that the method becomes infeasible in very high dimensional data (Markou & Singh, 2003b).

Chandola *et al.* (2009) report that the basic SVM technique described in Section 3.6.5 has been extended for anomaly detection in audio signal data (Davy and Godsill 2002), novelty detection in power generation plants (King *et al.* 2002), system call intrusion detection (Eskin *et al.*, 2002; Heller *et al.* 2003; Lazarevic *et al.* 2003), and detection of anomalies in temporal sequences (Ma & Perkins 2003a; 2003b).

4.4 Non-Target Data Inclusion

Should a one-class classifier include non-target data in its training set? Traditionally OCC techniques were reserved for classification techniques where only a single class of instances is exhibited at training time. At prediction time, new instances with unknown class labels either belong to this target class or a new class that was not available during training. Approaches have since evolved, but in the literature there is still no definitive answer to this question. Tax and Duin (2002) state:

> *"In one-class classification we assume that we have examples from just one of the classes ... one of the classes is character-*

ized well, while for the other class (almost) no measurements are available".

Similarly, Japkowicz *et al.*, (1995) states:

> *"Novelty detection approaches consequentially require very few, if any, negative training instances".*

From both these statements we can infer that only a small, or rare, amount of non-target data can be employed during training. This review will look at extending the capabilities of one-class classifiers and introducing a small number non-target data. This then raises the question; is rarity an absolute or relative property?

Raskutti & Kowalcyzk (2003) contend that in most case studies of class imbalance, the imbalance ratio of minority to majority class is around 10:90. In contrast to Tax and Japkowicz, Raskutti and Kowalcyzk (2003) assert that where the minority class consists of around 13% of the data, then *"one of the classes is ignored completely and learning is accomplished using examples from a single class".*

Recent research has addressed the issue of class imbalance, but few if any studies address the issue of rare events (Khoshgoftaar *et al.*, 2007). This rarity dilemma warrants further research, the author is of the opinion that one-class classifiers should make use of whatever data (target or non-target) is available.

Of interest to this review is the problem related to OCC when negative examples exist. The small number of negative examples prevents the possibility of making accurate assumptions about the characteristics of their distribution. In this sense, OCC can characterise the target class (positive examples) to distinguish it from all other non-target classes.

4.5 Conclusion

This Section examined the OCC problem. Section 4.1 considered the definition of an outlier and the general approaches to outlier detection. The characteristics of a one-class classifier were listed in Section 4.2. A brief description of the approaches to OCC were described in Section 4.3. Of particular interest to this review are the SVM approaches used by Tax and Duin(1999; 1999b) and Scholkopf and Smola (2002). The question about the inclusion of nontarget data in the training set of a one-class classifier was discussed in Section 4.4.

5 Conclusion

This section concludes the literature review. The goal of this review was to provide the reader with a snapshot of the proposed thesis subject area. The topics covered included; the role of the literature review, credit risk and scoring, classification and finally OCC. The review did not attempt an in-depth study of any of the topics, instead a broad approach was employed. This afforded an opportunity to establish a link between the topics, i.e. the tie-in between traditional classification and OCC all within the context of credit risk scorecards. The first Section of this literature review attempts to establish the framework of the review. The three main concepts of this section include Identification, Evaluation and Interpretation. In terms of future work this provides an approach for identifying current works, evaluating their intellectual context, and establishing the credibility and applicability of works. The author believes that the research methodology used in Section 1.3 is fundamentally sound. The review of how sources are organised and searched and can be employed for studies by the author.

Section 2 examined credit risk scorecards. The importance of credit scoring was outlined through an examination of the history of credit and the distinction between corporate and consumer credit. Any future publications will, at least, require an overview of the credit domain. The legislative importance of Basel 2 on credit scoring was also discussed. The idea of reject inference was also examined. This involves inferring the performance of rejected applicants, i.e. Their credit score was below the credit cut off point. The author is of the opinion that this would make an interesting sub-area as classification techniques are directly applicable to estimating the reject inference rate.

Section 3 provided an overview of classification. The theoretical framework offered an insight into how a classification model functions. Future work on this area should extenuate the role of the loss function, the author believes that classification examples in literature tend to overlook the importance of the loss function. Mainstream, but understandably vital, classification topics such as bias and variance were also discussed. Future work must also expand on the classification evaluation techniques listed in Section 3.4. A review of classification techniques used for estimating the probability of default were described in Section 3.6. Statistical methods are still the most popular, numerous studies (Baesen, 2002) have been completed in order to survey the effectiveness of such methods. Future thesis work would also benefit from

such an approach as it establishes the basis or starting point of credit scoring classification methods. Greater detail on the work of alternative methods (SVMs etc.) needs to be compiled by the author. Class imbalance was used as an introduction to OCC. Alternative techniques to OCC were discussed in Section 3.6. For future work the strengths of one-class classifiers could be enhanced by finding studies that highlight the limitations of methods that are used to balance the data set or modifying the classifier (e.g. undersampling, oversampling, cost-sensitive learning, classifier ensembles and classifier biasing).

OCC was reviewed in Section 4. Perhaps greater emphasis should have been placed on this section, in particular OCC approaches, as it is central to the proposed thesis. Can a one-class classifier still be considered such if non-target data is employed during its training? This question was explored in Section 4.1 and the author is of the opinion that the one-class classifier should employ whatever data is available during training. The characteristics of a one-class classifier were also examined. Of these the most important of which is identified by Tax (2001) *"a measure of the distance d(z) or resemblance (or probability) p(z) of an object z to the target class"* (Tax p.57, 2001) and a consideration or threshold σ on this distance or resemblance (Tax, 2001).

The approaches to categorising OCC methods vary within literature. Tax (2001) offers a clear categorisation by splitting the methods into three areas; density estimation, boundary estimation and reconstruction methods. However this approach is not exhaustive and the previously employed approach of statistical, neural and machine learning was used. Future work should consider a proper taxonomy of OCC approaches based on their characteristics. Furthermore greater detail on machine learning approaches needs to be furnished. Perhaps a comparative study of SVDD, one-class SVM and other SVM approaches would yield worthwhile observations and results.

Overall this literature review provided the author with an opportunity to consolidate the literature accumulated thus far. Undoubtedly blind spots exist in the literature review. The implementation of credit scoring classifiers and one-class classifiers requires further research, as the description of such methods lacked a satisfactory depth.

6 Bibliography

Agyemang, M., Barker, K., & Alhajj, R., 2006, A comprehensive survey of numeric and symbolic outlier mining techniques. Intelligent Data Analysis 10, 6, 521-538.

Allen, L., DeLong, G. & Saunders, A., 2004. Issues in the credit risk modelling of retail markets. Journal of Banking and Finance, 28(4), 727-752.

Alpaydin, E., 2004. Introduction To Machine Learning, MIT Press.

Altman, E., I., & Saunders, A., 1997, Credit risk measurement: Development over the last 20 years. Journal of Banking and Finance 21 11/12 (1997), pp. 1721-1742

Anderson, R.A., 2007. The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management, Oxford University Press: UK.

Baesens B., 2003, Developing Intelligent Systems for Credit scoring using Machine Learning Techniques, Ph.D.thesis, Katholieke Universiteit Leuven.

Bakar, Z., Mohemad, R., Ahmad, A., & Deris, M., 2006, A comparative study for outlier detection techniques in data mining. Cybernetics and Intelligent Systems, 2006 IEEE Conference on, 1-6.

Barandela, R. et al., 2003, Strategies for learning in class imbalance problems, Pattern Recognition 36, no. 3 (March 2003): 849-851, doi:10.1016/S0031-3203(02)00257-1

Barandela R, Sanchez JS., Garcia, V. & Rangel, E. (2003) Strategies for learning in class imbalance problems. Pattern Recognition 36:849-851

Barandela, R., Valdovinos, R. M., Sanchez, J. S., & Ferri, F. J, 2004, The imbalanced training sample problem: Under or over sampling? In Joint IAPR International Workshops on Structural, Syntactic, and Statistical Pattern Recognition (SSPR/SPR'04), Lecture Notes in Computer Science 3138, (806-814), 2004

Barnett, V. & Lewis, T., 1994. Outliers in Statistical Data, John Wiley & Sons New York.

Batista, G.E., Carvalho, A.C. & Monard, M.C., 2000. Applying One-Sided Selection to Unbalanced Datasets. LECTURE NOTES IN COMPUTER SCIENCE, 315-325.

Bellman,R., 1961, Adaptive Control Processes. Princeton University Press, Princeton, NJ.

Beranek, W. & Taylor, W., 1976. Credit-Scoring Models and the Cut-Off Point-A Simplification. Decision Sciences, 7(3), 394-404.

BIS, 1999, Bank for International Settlements, Consultative paper issued by the Basel Committee on Banking Supervision Issued for comment by 30 November 1999 `http://www.bis.org/publ/bcbs54.pdf?noframes=1`, accessed 8th January 2009.

BIS, 2000, Bank for International Settlements, `http://www.bis.org/publ/bcbs75.htmaccessed`, 7th January 2009.

BIS, 2005, Bank for International Settlements Newsletter No. 6, Validation of low-default portfolios in the Basel II Framework, `http://www.bis.org/publ/bcbs_nl6.pdf`

BIS, 2006, Basel Committee on Banking Supervision, International Convergence of Capital Measurement and Capital Standards, A Revised Framework. Technical Report, Bank for International Settlements, Basel. `http://www.bis.org/publ/bcbs128.pdf`

Bishop, C., 1994, Novelty detection and neural network validation, Proceedings of IEE Conference on Vision and Image Signal Processing, 1994, pp. 217-222.

Bourner, T., 1996, The research process: four steps to success, in Research methods: guidance for postgraduates, edited by T. Greenfield. Arnold: London, pp. 7-11.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984, Classification and regression trees.

Bruce, C. S. ,1994, 'Research student's early experiences of the dissertation literature review' Studies in Higher Education, vol. 19, no. 2, pp. 217-229.

Campbell, C. & Bennett, K.P., 2001, A linear programming approach to novelty detection, Advances in NIPS, Vol. 14, MIT Press, Cambridge, MA.

Chan, P.K. & Stolfo, S.J., 2001, Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, pages 164-168.

Chandola, V., Banerjee, A. & Kumar, V., 2008. Anomaly detection-a survey. ACM Computing Surveys (To Appear).

Chawla, N. V., 2003, "C4. 5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure," in Workshop on Learning from Imbalanced Data Sets II.

Cherkassky, V. & Mulier, F., 1999. Guest Editorial Vapnik-Chervonenkis (VC) Learning Theory and Its Applications. IEEE TRANSACTIONS ON NEURAL NETWORKS, 10(5), 985.

Cherkassky, V. & Mulier, F., 2007, Learning from Data: Concepts, Theory and Methods, Wiley Interscience

Crook, J. & Banasik, J., 2004, Does reject inference really improve the performance of application scoring models? Journal of Banking and Finance, 28(4), 857-874.

Crook, J.N., Edelman, D.B. & Thomas, L.C., 2007, Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183(3), 1447-1465.

Davy, M. & Godsill, S., 2002, Detection of abrupt spectral changes using support vector machines: an application to audio signal segmentation, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, Orlando, FL, 2002, pp. II-1313-II-1316.

Desai, V.S., Crook, J.N., & Overstreet, G.A., 1996, A comparison of neural networks and linear scoring models in the credit union environment, European Journal of Operational Research 95 (1), pp. 24-37.

Desforges, M.J., Jacob, P.J. & Cooper, J.E., 1998, Applications of probability density estimation to the detection of abnormal conditions in engineering, Proceedings of the Institute of Mechanical Engineers, Vol. 212, 1998, pp. 687-703.

Diehl, C.P. & Hampshire II, J.B., 2002, Real-time object classification and novelty detection for collaborative video surveillance, Proceedings of IEEE IJCNN Conference, Honolulu, HI, 2002.

Dietterich, T.G., 2000. Ensemble Methods in Machine Learning. LECTURE NOTES IN COMPUTER SCIENCE, 1-15.

Dinh, T.H.T. & Kleimeier, S., 2007. A credit scoring model for Vietnam's retail banking market. International Review of Financial Analysis, 16(5), 471-495.

Domingos P., 1999, Metacost: a general method for making classifiers cost-sensitive. In: Proc. 5th Intl. Conf. on Knowledge Discovery and Data Mining, pp. 155-164.

Drummond, C. & Holte, R.C., 2003, C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In Workshop on Learning from Imbalanced Data Sets II, International Conference on Machine Learning, 2003.

ECB, 2008, European Central Bank, http://www.ecb.int/stats/money/aggregates/bsheets/html/outstanding_amounts_2008-11.en.html

El-Yaniv, R. & Nisenson, M., 2007. Optimal Single-Class Classification Strategies. In Advances in Neural Information Processing Systems: Proceedings of the 2006 Conference. MIT Press.

Ezawa K.J., Singh M. & Norton, S.W., 1996, Learning goal oriented Bayesian networks for telecommunications management. In: Proc. 13th Intl. Conf. on Machine Learning, pp. 139-147

Fan, A. & Palaniswami, M., 2000, Selecting bankruptcy predictors using a support vector machine approach. In Proceedings of the international joint conference on neural networks.

Fawcett T., 2006, ROC graphs with instance-varying costs. Pattern Recognition Letters 27:882-891

Federal Reserve, 2008, Assets and Liabilities of Commercial Banks in the United States 1, Seasonally adjusted, http://www.federalreserve.gov/releases/h8/current/

Fisher, R.A., 1936, The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7: 179-188.

FTC, Federal Trade Commission, 1998, http://www.ftc.gov/bcp/edu/pubs/consumer/credit/cre15.pdf

Garca, V., Garreta, J.S., Cardenas, R.A.M., Alejo,R. & Sotoca, J.M., 2007, The class imbalance problem in pattern classification and learning. CEDI 2007 II Congreso Espaol de Informatica. Zaragoza: 11-09-2007. Nacional. 2007 Thomson. ISBN: 978-84-9732-602-5.

Gestel, T.V., Suykens, J.A.K., Baestaens, D.-E., Lambrechts, A., Lanckriet, G., & Vandaele, B., 2001, Financial time series prediction using least squares support vector machines within the evidence framework, IEEE Transactions Neural Networks 12 (2001) (4), pp. 809-821.

Gordon D.F. & Perlis, D. 1989, Explicitly biased generalization. Computational Intelligence 5:67-81

Han, H., Wang, W. Y., & Mao, B. H., 2005, Borderline-smote: A new over-sampling method in imbalanced data sets learning. In In International Conference on Intelligent Computing (ICIC'05). Lecture Notes in Computer Science 3644, pages 878-887. Springer-Verlag, 2005.

Hand, D.J. & Henley, W.E., 1997. Statistical Classification Methods in Consumer Credit Scoring: a Review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523-541.

Hand, D. & Henley, W., 1993. Can reject inference ever work? IMA Journal of Management Mathematics, 5(1), 45-55.

Hanley, J.A. & McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.

Hansen, L.K. , Liisberg C. & Salamon, P. (1997) The error-reject trade-off. Open Systems Inform. Dynamics 4 (1997), pp. 159-184.

Hart, E., 1998, Doing a literature review: releasing the social science research imagination, by E. Hart and M. Bond. London: Sage., p.1.

Hastie, T., Tibshirani, R., & Friedman, J., 2001. The elements of statistical learning, Springer New York.

Hawkins, D., 1980, Identification of Outliers, Chapman and Hall, London.

Haykin, S., Neural networks: a comprehensive foundation. 1999. Printice Hall, New Jersey, USA, 2.

Hellman, M.E., 1970, The nearest neighbour classification with a reject option. IEEE Trans. Systems Sci. Cybernet. 6 3 (July 1970), pp. 179-185

Hempstalk, K. & Frank, E., 2008, Discriminating Against New Classes: One-Class versus Multi-Class Classification, Springer Link Lecture Notes in Computer Science, Volume 5360/2008, November 27, 2008, ISBN 978-3-540-89377-6

Hempstalk, K., Frank, E. & Witten, I.H., 2008. One-Class Classification by Combining Density and Class Probability Estimation. In Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases-Part I. Springer, pp. 505-519.

Hodge, V. & Austin, J., 2004, A survey of outlier detection methodologies. Artificial Intelligence Review 22, 2, 85-126.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H. & Wu, S., 2004, Credit rating analysis with support vector machine and neural networks: A market comparative study, Decision Support Systems 37 (2004), pp. 543-558.

Huang, C.L., Chen, M.C. & Wang, C.J., 2007, Credit scoring with a data mining approach based on support vector machines. Expert Systems With Applications, 33(4), 847-856.

Japkowicz, N., Myers, C. & Gluck, M.A., 1995, A novelty detection approach to classification. In Proceedings of IJCAI.

Japkowicz, N., 2000, The class imbalance problem: Significance and strategies. In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI'2000). pp. 111-117.

Japkowicz N., 2001, Concept-learning in the presence of between-class and within-class imbalances. In: Proc. 14th Conf. of the Canadian Society for Computational Studies of Intelligence, pp.67-77

Japkowicz N., Stephen S., 2002, The class imbalance problem: a systematic study. Intelligent Data Analysis 6:429-450.

Jo T., Japkowicz N., 2004, Class imbalances versus small disjuncts. SIGKDD Explorations 6:40-49

Kiefer, N.M., 2008. Default estimation for low-default portfolios. Journal of Empirical Finance.

Juszczak, P., David M.J. Tax, D. M. J., Pekalska, E. & Duin, R.P.W., 2008, Minimum spanning tree based one-class classifier, Neurocomputing, In Press, Corrected Proof, Available online 16 July 2008, ISSN 0925-2312, DOI: 10.1016/j.neucom.2008.05.003.

Khoshgoftaar, T.M. et al., 2007. Learning with Limited Minority Class Data. In Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on. pp. 348-353.

Kim, K.J., 2003, Financial time series forecasting using support vector machines, Neurocomputing 55 (2003), pp. 307-319

Kulkarni, V.G., 1995, Modeling and Analysis of Stochastic Systems. Chapman Hall.

Lawrence, E.L., Smith, S., Rhoades, M., 1992, An analysis of default risk in mobile home credit, Journal of Banking and Finance 299 - 312.

Li, X. & Liu, B., 2005. Learning from Positive and Unlabeled Examples with Different Data Distributions. LECTURE NOTES IN COMPUTER SCIENCE, 3720, 218.

Liano, K., 1996. Robust error measure for supervised neural network learning withoutliers. Neural Networks, IEEE Transactions on, 7(1), 246-250.

Maloof, M., 2003, Learning when data sets are imbalanced and when costs are unequal and unknown. In Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets, 2003.

Markou, M. & Singh, S. 2003a. Novelty detection: a review-part 1: statistical approaches. Signal Processing 83, 12, 2481-2497.

Markou, M. & Singh, S. 2003b. Novelty detection: a review-part 2: neural network based approaches. Signal Processing 83, 12, 2499-2521.

Mitchell, T. et al., 1990, Machine Learning. Annual Reviews in Computer Science, 4(1), 417-433.

Min, J.H. & Lee, Y.C., 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems With Applications, 28(4), 603-614.

Moya, M.R., Koch, M.W. & Hostetler, L.D., 1993, One-class classifier networks for target recognition applications, in: Proceedings on World Congress on Neural Networks, International Neural Network Society (INNS), Portland, OR, 1993, pp. 797-801

Neuman, W. L., 2003, Social research methods: qualitative and quantitative approaches, 5th ed, Allyn and Bacon, Boston.

Odin, T. & Addison, D., 2000, Novelty detection using neural network technology, Proceedings of the COMADEN Conference, Houston, TX, 2000.

OECD, `http://stats.oecd.org/wbos/Index.aspx?DataSetCode=SNA_TABLE710`

Orriols A, Bernard E (2005) The class imbalance problem in learning classifier systems: a preliminary study. In: Proc.Intl. Conf. on Genetic and Evolutionary Computation, pp. 74-78

Patcha, A. & Park, J.-M., 2007, An overview of anomaly detection techniques: Existing solutions and latest technological trends. Comput. Networks 51, 12, 3448-3470.

Pazzani M., Merz C., Murphy P., Ali K., Hume T. & Brunk, C., 1994, Reducing misclassification costs. In: Proc. 11th Intl. Conf. on Machine Learning, pp. 217-225

Peng, K. et al., 2003. Exploiting unlabeled data for improving accuracy of predictive data mining. In Proceedings of the Third IEEE International Conference on Data Mining (ICDM). pp. 267-275.

Phipps, J., Seal, K. & Duncan, E., 2004, Introductory Paper on Low Default Portfolios, Joint Industry Working Paper, British Bankers Association, London Investment Banking Association, international Swaps and Derivatives Association

Poggio, T., Girosi, F. & MIT, C., 1990. Networks for approximation and learning. Proceedings of the IEEE, 78(9), 1481-1497.

Prati RC, Batista GE, Monard MC, 2004, Class imbalance versus class overlapping: an analysis of a learning system behaviour. In: Proc. 3rd Mexican Intl. Conf. on Artificial Intelligence, pp. 312-321

Provost F., Fawcett T., 1997, Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In: Proc. 3rd Intl. Conf. on Knowledge Discovery and Data Mining, pp. 43-48

Rasmussen C., 1996, Evaluation of Gaussian Processes and Other Methods for Non-Linear Regression. PhD thesis, Department of Computer Science, University of Toronto

Raskutti, B. & Kowalczyk, A., 2004. Extreme re-balancing for SVMs: a case study. ACM SIGKDD Explorations Newsletter, 6(1), 60-69.

Ratsch, G., Mika, S., Scholkopf, B. & Muller, K., 2002, Constructing boosting algorithms for SVMs: an application for one-class classification. IEEE Trans. Pattern Anal. Machine Intell. 24 9 (2002), pp. 1184-1199

Roberts, S. & Tarassenko, L., 1994, A probabilistic resource allocating network for novelty detection. Neural Comput. 6 (1994), pp. 270-284.

Roth, V.,2005, Outlier Detection with One-class Kernel Fisher Discriminants. Advances in Neural Information Processing Systems, 17, 1169-1176, MIT Press.

Sabato, G., 2006, Managing Credit Risk for Retail Low-Default Portfolios.

Saunders, R. & Gero, J.S., 2000, The importance of being emergent, Proceedings of the Artificial Intelligence in Design, 2000

Scholkopf, B. & Smola, A.J., 2002, Learning with Kernels, MIT Press.

Senf, A., Chen, X. & Zhang, A., 2006. Comparison of One-Class SVM and Two-Class SVM for Fold Recognition. LECTURE NOTES IN COMPUTER SCIENCE, 4233, 140.

Siddiqi, N., 2005, Credit Risk Scorecards: Developing And Implementing Intelligent Credit Scoring, SAS Publishing.

Stephan, F., 2001, On one-sided versus two-sided classification. Archive for Mathematical Logic, 40(7), 489-513.

Tarassenko, L., Nairac, A., Townsend, N., & Cowley, P., 1999, Novelty detection in jet engines, IEE Colloquium on Condition Monitoring, Imagery, External Structures and Health, Birmingham, UK, 1999, pp. 41-45

Tax, D.M.J. & Duin, R.P.W., 1999, Support vector domain description. Pattern Recognition Lett. 20 (1999), pp. 1191-1199

Tax, D.M.J. & Duin, R.P.W., 1999b, Data domain description using support vectors, Proceedings of ESAN99, Brussels, 1999b, pp. 251-256.

Tax, D. & Duin, R., 2000. Data Description in Subspaces. In INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION. pp. 672-675

Tax, D.M.J. & Duin, R.P.W., 2000b, Outlier detection using classifier instability, in: Advances in Pattern Recognition, the Joint IAPR International Workshops, Sydney, Australia, 1998, pp. 593-601.

Tax, D., 2001. One-class Classification. Ph.D. thesis, Delft University of Technology, The Netherlands

Tax, D.M.J. & Duin, R.P.W., 2002. Uniform object generation for optimizing one-class classifiers. The Journal of Machine Learning Research, 2, 155-173.

Tay, E.H. & Cao, L., Application of support vector machines in financial time series forecasting, Omega 29 (2001), pp. 309-317

Taylor, D. & Procter, M., 1998, The literature review: a few tips on conducting it. Writing at the University of Toronto Website, `http://utl1.library.utoronto.ca/disk1/www/documents/writing/litrev.html`, Site accessed 30th January 2009.

Thomas, L.C., 2000. A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers. International Journal of Forecasting, 16(2), 149-172.

Thomas, L.C., Ho, J. & Scherer, W.T., 2001. Time will tell: behavioural scoring and the dynamics of consumer credit assessment. IMA Journal of Management Mathematics, 12(1), 89-103.

Thomas, L.C., Edelman, D.B. & Crook, J.N., 2002. Credit scoring and its applications, Society for Industrial Mathematics.

Titterington, D.M., 1992. Discriminant analysis and related topics. In: Thomas, L.C., Crook, J.N. and Edelman, D.B., Editors, 1992. Credit scoring and credit control, Oxford University Press, Oxford, pp. 53-73.

Tsai, C.F. & Wu, J.W., 2008. Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Systems with Applications, 34(4), 2639-2649.

Vapnik, V.N. & Chervonenkis, A., 1968, "On the uniform convergence of relative frequencies of events to their probabilities," Doklady Academii Nauk USSR, vol. 181, no. 4, 1968.

Vapnik, V. N., 1995, The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Vapnik, V.N., 1998, Statistical learning theory, Wiley/InterScience, New York (1998).

Viaene, S., Derrig, R.A., Baesens, B., & Dedene, G., 2002 A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection, The Journal of Risk and Insurance 69 (2002) (3), pp. 373-421

Wang, Y., Wang, S. & Lai, K.K., 2005. A new fuzzy support vector machine to evaluate credit risk. IEEE Transactions on Fuzzy Systems, 13(6).

Wang, B.X. & Japkowicz, N., 2008. Boosting Support Vector Machines for Imbalanced Data Sets. LECTURE NOTES IN COMPUTER SCIENCE, 4994, 38.

Weiss, G. M. & Hirsh, H,2000, A quantitative study of small disjuncts. In Proceedings of the Seventeenth National Conference on Artificial Intelligence, pages 665-670, AAAI Press, 2000.

Weiss GM (2003) The Effect of small disjuncts and class distribution on decision tree learning, PhD thesis, Rutgers University

Wiginton, J.C., 1980. A note on the comparison of logit and discriminant models of consumer credit behavior. Journal of Financial and Quantitative Analysis, 757-770.

Wu, G. & Chang, E.Y., 2003, Class-boundary alignment for imbalanced dataset learning. In Proceedings of the ICML.

Ye, N., 2004. Handbook of Data Mining,

Zhang, D. et al., 2007. A Comparison Study of Credit Scoring Models. In Natural Computation, 2007. ICNC 2007. Third International Conference on Natural Computation.