An investigation into the relative abilities of three alternative data mining methods to derive information of business value from retail store-based transaction data

Shane Butler

BInfoTech (Deakin University)

8th of November, 2002

Submitted for the degree of Bachelor of Information Technology (Honours) in the School of Computing and Mathematics, Faculty of Science and Technology, Deakin University, Victoria 3217, Australia

Abstract

For many retail organisations, data collection is a routine activity and there is widespread recognition of the potential for analysis of past transaction data to improve the quality of business decisions.

The use of several data mining tools will be explored and their outputs compared and contrasted. Experiments have been designed and carried out, including a detailed analysis of results. Decision trees will be found to produce overly complex rules with too much emphasis on negative items.

Contrary to Bay and Pazzani's assertion, association rules can be used to find differences between groups.

Keywords: data mining, transaction data, decision trees, association rules, contrast sets

Table of contents

List of tables

List of figures

Abbreviations

Abbreviation	Meaning
AR	Association Rules
CS	Contrast Sets
DT	Decision Trees

Acknowledgements

I would like to acknowledge the following people:

- My supervisors: Prof. Geoff Webb and Dr Doug Newlands, for all their support and guidance throughout the project;
- Industry partner: for their participation in the project, real world data and feedback;
- Stephen Bay: for making the STUCCO binary available to me;
- BIT program industry sponsors and supporters: Without their support such a rewarding program would not be possible;
- BIT colleagues: Rich, JP, Paul, Tim, Jeremy, Byron, Mark and our Program Director, Dr. David Mackay; and
- Deb, Brian, Ian and Kelly for all their support.

Chapter 1: Introduction

1.1 Background

For many retail organisations, data collection is a routine activity. The advent of electronic registers and barcoding of supermarket goods has not only increased efficiency in the retail industry, but also meant that every transaction can be easily captured and then stored in a database [1, 2].

More and more organisations are realising the hidden value of their large databases [3]. There is widespread recognition of the potential for analysis of past transaction data to improve the quality of future business decisions. Although several data mining tools exist that are well suited to this type of problem, it is not apparent which techniques are more valuable for deriving such interesting information.

1.2 Research question

The research that will be conducted asks, "What are the relative abilities of three alternative data mining methods to derive information of business value from retail store-based transaction data?"

In answering this research question, the following sub-questions will also be addressed:

- How well suited are each of the methods to extracting the required knowledge?
- To what extent are domain experts likely to understand information from these methods?
- How unexpected do the domain experts find information produced by these methods?
- How potentially valuable do the domain experts find information produced by these methods?
- What do these methods have in common, and what is different?

1.3 Significance

The Australian retail industry has a large turnover, exceeding AU$14 billion monthly [4] and AU$150 billion in 2000/2001 financial year [5]. The industry is

important to the Australian economy as it contributes approximately 8% of Australia's Gross Domestic Product (GDP).

It has been shown that derived knowledge can be applied to add value [6], for example, to improve efficiency, improve services delivered to community (such as helping supply the products people really want), provide competitive advantage, etc.

Applying this in the Australian retail industry has the potential for a big impact. For example, the retail industry spends approximately 2% of its turnover in advertising and marketing [5], so even adding a small amount of value here could have a big impact on the bottom line. Therefore this research is important to the retail industry and through that to the Australian economy.

1.4 Methodology

As a data mining project, a data mining methodology will be used. The methodology employed is called CRISP-DM and it is consists of several phases: business understanding; data understanding; data preparation; modelling; evaluation and deployment. CRISP-DM will be used to mine the organisations data using the three alternative data mining methods.

After the data mining has been completed, the results will be transformed into plain English information and put into a questionnaire. The domain experts will then be asked to rate the information based on their knowledge and experiences.

Initially a quantitative research approach was planed for the analysis of the questionnaire results. However after insufficient responses a quantitative research approach was adopted.

1.5 Structure of this thesis

In Chapter 2 literature surrounding these data mining methods is reviewed. Chapter 3 then outlines in detail the research methodology being employed and why this approach has been taken. Chapter 4 presents, analyses and discusses the results. Chapter 5 draws conclusions about the research and presents some further implications.

1.6 Conclusion

This chapter has laid the foundations for the thesis. It introduced the research questions after which the research was justified, definitions were presented, the methodology was briefly described and the thesis was outlined. On these foundations, the thesis can proceed with a detailed description of the research.

Chapter 2: Literature review

2.1 Introduction

This chapter discusses the literature in the area of data mining and machine learning. It is organised as follows: section 2.2 introduces data mining and its applications. In sections 2.3-2.5 several set mining techniques are examined. Section 2.3 looks into decision trees, section 2.4 investigates association rules and section 2.5 introduces contrast-sets.

2.2 Data mining

Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [7]. It involves the automated prediction of trends and behaviours to discover previously unknown patterns.

Data mining has recently become popular as the three technologies by which it is supported, have started to mature [8], namely:

- Massive data collection: Companies have realised the value of information and are collecting all sorts of data.
- Powerful multiprocessor computers: The advent of cost effective and powerful computers.
- Data mining algorithms: These techniques have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.

While there are many different data mining algorithms in use today, the majority of them can be categorised as one of the following types of algorithms:

- Decision trees: Tree shaped structures that represent decisions processes. These decisions generate rules for the classification of a dataset. Popular decision tree methods include ID3 [9], C4.5 [10] and CART [11]. Decision trees are discussed further in section 2.3.
- Artificial Neural Networks: Non-linear predictive models that learn though training and resemble biological neural networks in structure.

- Rule induction: The extraction of useful if-then rules from data based on statistical significance. This includes association rules [12] which are useful for market basket analysis and contrast sets [13, 14]. Association rules are discussed further in section 2.4 and contrast sets in section 2.5.
- Nearest neighbours: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset.
- Genetic algorithms: Optimisation techniques that use processes such as genetic combination, mutation, and natural selection in a design based on concepts of evolution.

2.2.1 Relationship to statistical analysis

Data mining can be seen as a successor to traditional statistical analysis. However data mining techniques have several notable differences when compared with statistical analysis.

Traditional statistical analysis attempts to investigate the amount of support a dataset affords a hypothesis, this is hypothesis testing. However, data mining is concerned with the automated information discovery of knowledge without prior generation of hypotheses. The data mining process is often of a computer automated exploratory nature, conducted on large and complex datasets [15].

Some authors believe that classical statistical models, dominated by linear models, are models for modest, not large, datasets [16]. Many statistical models can be too efficient at setting rigorous standards for modelling and statistical proof, this often coming at the expense of simplifying assumptions [16].

Data mining techniques also have advantages over using a domain expert. These techniques can find hidden patterns and predictive information that domain experts may miss because they become lost in the volume of data or uncover information that was not previously known.

2.2.2 Data mining applications

Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. They can be used to answer business questions

that traditionally were too resource intensive to resolve. It should be noted however, that data mining is not a business solution, it is just a technical one. That is, data mining can be used to make smarter business decisions by being used in conjunction with a solid understanding of the business.

Data mining has been used in a number of different and wide-ranging applications. Within the retail industry for example, data mining can be applied in many areas, such as:

- Sales
    - A comparison of sale day and non-sale day data
    - Identify seemingly unrelated products that are often purchased together
    - Identify buying patterns from customers
    - Help account for peak periods of consumption, merchandise throughput
- Quality Control
    - Identifying anomalous data – such as keying errors (pattern recognition)
    - Using refunds and sales data to identify problem products
- Fraud
    - Credit card transactions – using pattern recognition to detect fraudulent use
    - Irregular transactions
    - Resolve unusual patterns: refunds, discounts, price overrides, credit cards, store cards, debit cards, staff discounts, voids, reversals, overage and shortages

2.2.3 Ethics

Using data about people can have serious ethical and legal implications. Data miners need to be aware and responsible for the use of such information in their mining [17].

Data mining is often used to discriminate amongst people, for example, in an insurance risk assessment. However, discrimination on racial, sexual, or religious bases is potentially both unethical and illegal.

Data miners may also need to be aware of privacy and personal information handling laws. For example, in Australia laws enforcing ten National Privacy Principles (NPP) [18] are now in place. These laws cover the collection and use of personal information in following areas:

- Collection: Organisations cannot collect personal information unless it is necessary for them to do so.
- Use and disclosure: Personal information cannot be disclosed other than the purpose it was collected for.
- Data quality: Personal information must be kept accurate, complete and up-to-date.
- Data security: Protection of such personal information should be maintained.
- Openness: Organisations should document their privacy policies and make them available.
- Access and correction: Organisations must provide a means for personal information to be accessed and corrections made.
- Identifiers: Personal information cannot be labelled such that it becomes uniquely identifiable.
- Anonymity: Where possible, individuals must have the option of not identifying themselves when entering transactions with an organisation.
- Transborder data flows: Transborder data flow cannot occur unless the receiving organisation conforms to practices similar to the NPP.
- Sensitive information: Organisations must not collect sensitive information about individuals.

2.2.4 Selection of data mining methods

Although there are many data mining techniques available, only three have been selected for this study. The three that have been selected are:

- Decision trees
- Association rules
- Contrast sets

Selection of data mining methods and tools is discussed in detail in section 3.5.4.1.

2.3 Decision trees

Decision trees are a graph of choices or decisions. Each node, from the root node down, represents a decision. The final node of any branch (the leaf node) is used for classification. An example decision tree is shown in Figure 1.

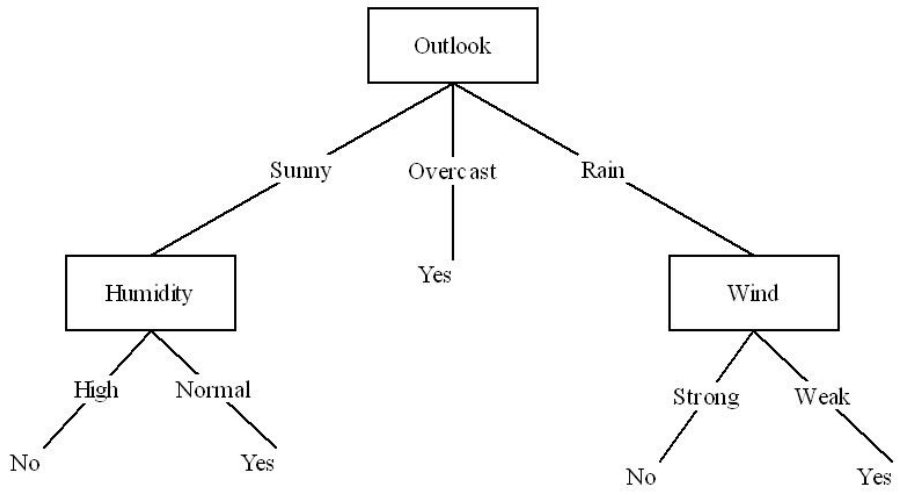

Figure 1: An example decision tree [19]

During training, the tree is *greedily* constructed from the root down using a basic divide and conquer algorithm. At each decision node the algorithm chooses the next 'best' attribute on which to branch. Many different strategies have been developed for attribute selection.

To classify a case, the learner follows a path down the decision nodes until reaching a leaf node. The case is then classified by the information in that leaf node.

Decision trees are an intuitive data mining method as the resultant model can be easily read and understood by non-data miners and domain experts alike. However, as the data becomes more complex, so too does the model [20].

2.3.1 Overfitting and pruning

A learned decision tree can be quite large, which presents the problem of overfitting the training data. A hypothesis is said to *overfit* the training data if another hypothesis has a higher error rate on the training data, but a lower error rate on the entire distribution of instances [21].

One method of avoiding overfitting is *pruning*. Pruning attempts to reduce the search space by rearranging or removing decision nodes. There are two types of pruning methods: *Pre-pruning* and *post-pruning* [19]. Pre-pruning approaches try to stop the tree growing earlier, during the growing process, while post-pruning methods allow the tree to first overfit the data before pruning. Two methods for post-pruning are [19]:

- *Subtree replacement*: Replacing subtrees with a leaf node.
- *Subtree raising*: Moving a subbranch of the tree up one or more levels.

2.3.2 Regression trees

The types of decision trees mentioned so far have only been applicable to nominal data. This means that numeric or continuous data must first be discretized. Unfortunately, discretization can lead to loss of accuracy and meaning of the underlying information. One popular discretization method is to split the range by specifying bounds. An example of this would be a choice on the variable *age* with the decision based on $age < 34$ and $age \geq 34$.

Regression trees are a similar tree representation to classification trees, however they allow numeric values at the leaves [17].

2.3.3 Algorithms

Popular decision tree algorithms include ID3 [9], C4.5 [10], CART [11] and CHAID [22]. These are discussed below.

2.3.3.1 ID3

The basic technique used by the ID3 (*Itemized Dichotomizer 3*) [9] family of algorithms is called Top-down Induction of Decision Trees (TDIDT). The algorithm uses a divide-and-conquer strategy combined with logical dichotomization (splitting into segments) to produce impressive results in comparatively short processing times.

The ID3 algorithm has several advantages, when compared to other algorithms at that time, including its simplicity of algorithm design, conservative use of system

resources and the simplicity of generated models. It also works well on large and complex datasets, and has a linear computation time [23].

ID3's weaknesses lie in that the generated models can sometimes be confusing to humans when trained on large or noisy datasets. Another potential problem is that you cannot efficiently change a learned tree. Attempting to do so can lead to a tree that is no longer optimal [23].

2.3.3.2 C4.5

Quinlan's C4.5 [10] is the successor to his earlier ID3 [9] algorithm. The new C4.5 algorithm produces small and accurate trees resulting in fast, reliable classifiers.

C4.5 and its predecessor, ID3, use formulas based on information theory to evaluate the *goodness* of a test; in particular, they choose the test that extracts the maximum amount of information from a set of cases, given the constraint that only one attribute will be tested [21]. Although other authors have suggested alternative measures of information gain, the measure used by C4.5 is entropy based [11].

Many datasets contain cases for which some attribute values are unknown. C4.5 handles unknown values by adjusting the information gain for the attribute accordingly.

C4.5 uses a pruning approach to avoid overfitting. C4.5's pruning method is based on estimating the error rate of every subtree, and replacing the subtree with a leaf node if the estimated error of the leaf is lower [21].

Quinlan has since published methods for the improved use of continuous attributes in C4.5 [24]. An updated (commercial) version, C5.0 [25] is also available.

2.3.3.3 CART

Classification and Regression Trees (CART) [11] algorithm provides decision trees that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. It creates these trees by creating binary splits [22].

2.3.3.4 CHAID

Chi Square Automatic Interaction Detection (CHAID) [22] segments a dataset by using chi square tests to create multi-way splits.

2.3.4 Problem suitability

Problems suitable for decision tree learning are often:

- Comprising of attribute-value pairs – Decision tree learners work well with problems that consist of attribute-value pairs. An attribute-value pair is just an occurrence of an attribute, such that it consists of attribute and one of its possible values. For example $A_1 = V_1$ and $A_1 = V_2$.
- Expect discrete output values from the target function
- Good for disjunctive descriptions – the resulting node path can be read as a combination of independent rules.
- Have training data may contain errors (Often called 'noisy' training data).
- Have training data may contain missing attribute values – Decision tree learners can accommodate for data that has missing attribute values.

For a detailed comparison of the appropriateness of several types of tree classifiers see [26].

2.4 Association rules

Association rule discovery [12] methods learn relations between variables within a dataset. An association rule consists of two sets of items called the *antecedent* and *consequent*. It indicates that a relationship exists between the two sets, such that the occurrence of the first set (antecedent) implies the occurrence of the second set (consequent).

The initial application used to demonstrate association rules was supermarket transaction analysis. Association rules are well suited to this type of market basket problem, as it involves the analysis of products and customer patterns. Agrawal, *et al.* [12] give several supermarket examples of the types of tasks association rules are suited to:

- Find all rules that have "Diet Coke" as consequent. These rules help to plan what the store should do to boost sale of Diet Coke.
- Find all rules that have "bagels" in the antecedent. These rules may help determine what products may be impacted if the store discontinues selling bagels.
- Find all rules that have "sausage" in the antecedent and "mustard" in the consequent. This query can be phased alternatively as a request for the additional items that have sold together with sausage in order to make it highly likely that mustard will also be sold.

2.4.1 Representation

A *training set* is a finite set of *records* where each record is an element to which Boolean predicates called *conditions* are applied [27]. A *rule* can be expressed $A \Rightarrow C$, where antecedent *A* implies consequent *C*. The antecedent is often called the *left-hand-side* or *LHS* and the consequent the *right-hand-side* or *RHS* [28].

Initial association rule definitions [12] only allowed one variable in the consequent, however this variable might differ form rule to rule. More recent publications [27, 29] allow many target attributes to be discovered as the consequent in rules.

The association rule problem is interested in discovery of rules that meet the following constraints [12]:

- *Syntactic constraints:* Constraints that involve restrictions on the items that can appear in a rule.
- *Support constraints:* Constraints involve the number of transactions that support a rule. Support is discussed further in section 2.4.2.

It has also been suggested that by integrating item constraints into the mining algorithm can to dramatically reduce the execution time [30].

The basic rule discovery process can be decomposed into two parts [29]:

1. Find all large itemsets that satisfy the minimum support

2. Use the large itemsets to generate the desired rules

2.4.2 Rule selection metrics

The search for association rules is often restricted to association rules that have a reasonably large number of instances and have a reasonably high accuracy on the instances they apply to [17], although there are other rule selection methods.

The *support* (or *coverage*) of an association rule is the number of instances for which it predicts correctly [17]. Support corresponds to strength of statistical significance and is also useful for restricting interest to rules with support above some minimum threshold for business reasons [12].

Its *confidence* (or *accuracy*) is the number of instances that it predicts correctly, expressed as a proportion of all the instances to which it applies [17].

Other popular rule selection metrics include:

- *Lift*: The number of instances that contain the antecedent over the number of instances that contain the consequent [28].
- *Leverage*: Proposed by [31], leverage is the difference between the number of instances where the antecedent and consequent co-occur and the number of instances that would be expected if the two were independent [28].

2.4.3 Algorithms

Popular algorithms for mining association rules include AIS [12], Apriori [29] and OPUS_AR [27]. A new algorithm called OPUS_IR [32] is also discussed.

2.4.3.1 AIS

AIS [12] algorithm searches for rules that satisfy a minimum support and a minimum confidence. Rules only have one item in the consequent but may have any number of items in the antecedent. An estimation procedure is used to determine which itemsets to measure in a pass, pruning and buffer management are also features of this algorithm.

Candidate itemsets are generated on-the-fly during the database pass. After reading a transaction, it is determined which of the itemsets found to be large in the previous pass are present in the transaction. New candidate itemsets are generated by extending these large itemsets with other items in the transaction [12].

One shortcoming of the AIS [12] is that it unnecessarily generates and counts too many candidate itemsets that turn out to be small [29].

2.4.3.2 Apriori

Apriori [29, 33] is a popular and widely available [34] algorithm for mining association rules. Apriori [29, 33] like AIS [12] utilises the concept of an *itemset*, however Apriori allows both the consequent and antecedent to contain many items.

Apriori generates the candidate itemsets to be counted in a pass by using only the itemsets found to be large in the previous pass. This has the effect of pruning the search because any subset of a large itemset is also likely to be large [29].

2.4.3.3 OPUS_AR

OPUS_AR [27] is an algorithm for association rule analysis based on the efficient Optimised Pruning for Unordered Search (OPUS) [35] search algorithm. When compared with Apriori [29], OPUS_AR requires more passes though the dataset. However, if the data can be retained in memory, this is not such a problem [27].

A commercial implementation of the OPUS_AR algorithm, called Magnum Opus [36] was used in these experiments. Magnum Opus has the ability to search using leverage, lift, strength, coverage, or support.

2.4.3.4 OPUS_IR

Traditional association rules cannot have the consequent be a numeric variable. *Quantitative association rules* [37, 38] have therefore been proposed for the discovery of associations between numeric variables impact rules. Later it was suggested they be called *impact rules* [32] to avoid confusion with other research.

OPUS_IR [32] is an efficient algorithm for discovery of impact rules using the based on the efficient Optimised Pruning for Unordered Search (OPUS) [35] search algorithm.

2.5 Contrast sets

Contrast-sets [13, 14] can be used to identify differences between groups. Contrast-sets are defined as conjunctions of attributes-value pairs that differ meaningfully in their probabilities across several groups.

2.5.1 STUCCO algorithm

STUCCO (Search and Testing for Understandable Consistent Contrasts) [13, 14] is an algorithm for mining contrast-sets. STUCCO has several important features [13]:

- Admissible pruning rules
- Guaranteed control over false positives: STUCCO has mechanisms to avoid concluding there is a difference when none exists [13].
- Compact summarization of results

The STUCCO mining algorithm treats the problem of mining contrast sets as a tree search problem. The root node is an empty node, and children of a node are generated by specializing the set by adding one more term. Canonical ordering is used to avoid visiting the same node twice. This drastically reduces the search space, as shown in Figure 2.

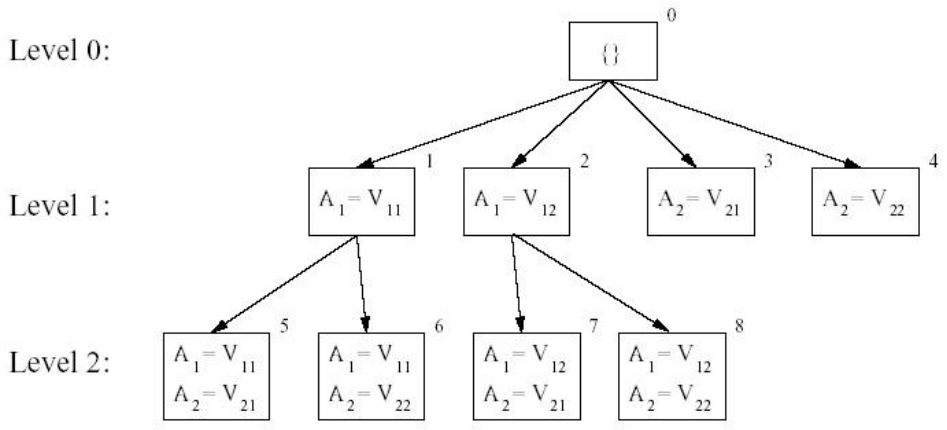

Figure 2: Example of the search tree for two attributes $A_1 = \{V_{11}, V_{12}\}$ and $A_2 = \{V_{21}, V_{22}\}$ [13]

2.5.2 Surprising contrast sets

Dealing with the large amount of data produced by data mining is a difficult problem. Therefore contrast-sets are post-processed to present only a subset that is surprising given what has already been shown. The user is shown the most general contrast sets first, those involving a single term, and then only shown more complicated conjunctions if they are surprising based on the previously shown sets [14].

2.5.2.1 Insightfulness

Current mining algorithms can produce rules which differentiate the groups with high accuracy, but often human domain experts find these results neither insightful nor useful.

Bay and Pazzani [39] compared two data mining algorithms to determine if a discriminative or characteristic approach was more useful for describing group differences. Discriminative approaches attempt to find differences that can be directly used to classify the instances of the groups, while characteristic approaches attempt to find differences in the class descriptions ([40] in [39]).

The first algorithm used in the experiment was STUCCO [13], a characteristic approach. The second was C5 [25], a discriminative approach which is a commercial derivative of C4.5 [10].

Using UC Irvine (UCI) admissions data, the researchers used the two different approaches to see why admitted students choose to enrol or not to enrol.

After generating rules using both techniques, the researchers showed the rules in plain English form to several human domain experts (the UCI admissions officers). The human domain experts where asked to rate the rules according to their *insightfulness*.

The discriminative and characteristic approaches resulted in two very different rule sets describing the differences between the students who enrol and do not enrol, and the researchers concluded that characteristic differences are more useful to domain experts than purely discriminative differences [39].

2.6 Conclusion

This chapter has provided a brief overview of the field of data mining and introduced three specific data mining algorithms: decision trees, association rules and contrast sets.

Several set mining techniques have been discussed. Decision trees, where the goal is to find attribute-value pairs with high predictive power; association rules where the goal is to find relations between itemsets; contrast-sets where the goal is to find all sets that represent large differences in the probability distributions of two or more groups [41, 42].

Each of these techniques are useful for specific applications within data mining. However further research is required to compare their relative abilities of these data mining methods for deriving interesting information.

Chapter 3: Methodology

3.1 Introduction

This chapter will examine the methodology to be employed in the project. This will include the constraints, the research approach options and selected research approach, the data mining methodology and questionnaire details.

3.2 Constraints

Several constraints had a limiting effect on this research study:

- Time constraints: Time constraints were very influential in limiting the scope of the research project. Relaxation of the project time frame could have allowed for a more comprehensive review of the literature, and the data mining activity could have involved more in-depth analysis.
- Transaction data availability constraints: It was initially believed that data would be available in the early days of the project. Unfortunately the industry partner's systems were not fully deployed until near the end of the project. As such, the transaction data extracted from these systems was not fully available until the end of the project, and extracts were only available from six stores/retail outlets. Had the data arrived earlier and been of larger quantities, more analysis could have been done.
- Feedback constraints: The major constraint to the feedback in the project was the availability of staff resources at the sponsor organisation to participate in the questionnaire. Also the questionnaire had to be completed in a short time frame and subsequently participants could only be asked to rate five pieces of information for each data mining method. Ideally more questions would have been asked of more people to obtain a higher quality of feedback. An alternative data gathering method could have been used had there been higher access to staff at the industry partner organisation.
- Software availability and licensing constraints: Initially it was envisaged that the availability and licensing of the software would be a constraint. However, while no source code for STUCCO and Magnum Opus was available, pre-compiled binaries suitable for the purpose of this research study were made available.

3.3 Research methodology options

There are several methodology options when conduction any research project. Most research methodologies fall into one of two types: quantitative or qualitative methods.

Qualitative methods attempt to capture and understand individual definitions, descriptions and meanings of events, while quantitative research methods, count and measure occurrences [43].

3.3.1 Quantitative research methods

3.3.1.1 Experimental and quasi-experimental research

Experimental research involves the discovery of causal relationships. At least one independent variable, called the experimental variable, is deliberately manipulated or varied by the researcher [44]. In quasi-experimental research, the research attempts to uncover a causal relationship even though all the variables that might affect the outcome cannot be controlled [44].

3.3.1.2 Survey research

In research it is often the case that variables cannot always be manipulated in an experimental fashion [44]. The major forms of survey are descriptive and explanatory surveys [43]. Descriptive surveys aim to estimate as precisely as possible the nature of existing conditions, while explanatory surveys aim to establish cause and effect relationships but without experimental manipulation.

Survey research can involve both closed items and open items [43]. Closed items usually allow the respondent to choose from two or more fixed alternatives. Open items simply allow the respondent to provide their own answers.

Unfortunately, with surveys the researcher has no scope to find out about the beliefs, feelings or perceptions of the respondent that do not fit into the response categories. Surveys are also fairly impersonal.

3.3.2 Qualitative research methods

3.3.2.1 Ethnographic research

Ethnography encompasses the study of a group of people for the purpose of describing the socio-cultural activities and patterns [43]. Ethnographic research involves a variety of data-collection methods, the primary procedure being observation [44]. Other data collection methods that may be appropriate include videotaping and interviewing.

3.3.2.2 Unstructured interviewing

Interviewing is the face-to-face data collection. It usually occurs via a meeting however can also be completed via telephone. Interviews can be conducted one-on-one or in a group.

Interviews can be structured in several different ways [44]:

- Unstructured (or open-ended) interviews: are useful for in-depth interviews or group interviews. The interviewer will only prepare a light outline, and provoke discussion. There is plenty of scope for the interviewer to vary the depth, breadth and pace of the interview, allowing questions to be revisited or expanded on if required [45].
- Semi-structured: useful for survey interviews and group interviews where fixed wording or fixed ordering of the questions need not be used. A direction is given to the interview so that the content focuses on the crucial issues of the study.
- Structured: structured interviews are rarely used as for quantitative analysis as they are better suited to quantitative analysis. See section 3.3.1.2 for further details).

3.3.2.3 Action-research

Action-research is the application of fact-finding to solving practical problems. It is situational, collaborative and participatory. Action-research involves teachers as generators of knowledge in a bottom-up approach to professional development with the professional researcher as the resource person for the teacher [43].

3.3.2.4 Case study

Case studies are quite extensively used in qualitative research [44]. A case study is a detail examination of something; a specific event, an organisation, a system, just to name a few examples [44]. Organisation case studies and observational case studies are the two most commonly use designs. Case studies may also use other techniques, such as interviews, questionnaires, data analysis, document reviews and observations over time [45].

3.3.2.5 Historical research

Historical research is intended to understand, explain or predict, through some systematic collection and objective evaluation of data relating to the past occurrences [43]. This is done through the use of qualitative tools such as documents, interviews, biographies and events. Historical research can also make use of quantitative data, such as the changing demographics.

3.3.3 Sampling methods

Whereas quantitative research uses probability sampling, qualitative research employs non-probability sampling, especially snowball sampling and theoretical sampling. In snowball sampling, a person, who is identified as a valid member of a specific group to be interviewed, is asked to provide the names of others who fit the requirements [43].

3.4 Selected research methodology

It was the intention of the researcher to conduct a quantitative study. A questionnaire was planned and the indication from the industry partner organisation was that this would be possible.

The planned quantitative analysis was to include a detailed analysis of the questionnaire results. However, based on insufficient responses to the questionnaire a qualitative approach was adopted.

3.5 Data mining methodology

The CRoss-Industry Standard Process for Data Mining (CRISP-DM) [46] is a reference model by the CRISP-DM Consortium (NCR, DaimlerChrysler, SPSS and OHRA). Since being published CRISP-DM has become a standard in the industry [47]. It aims to describe the individual phases of the data mining process [46], as illustrated in Figure 3.

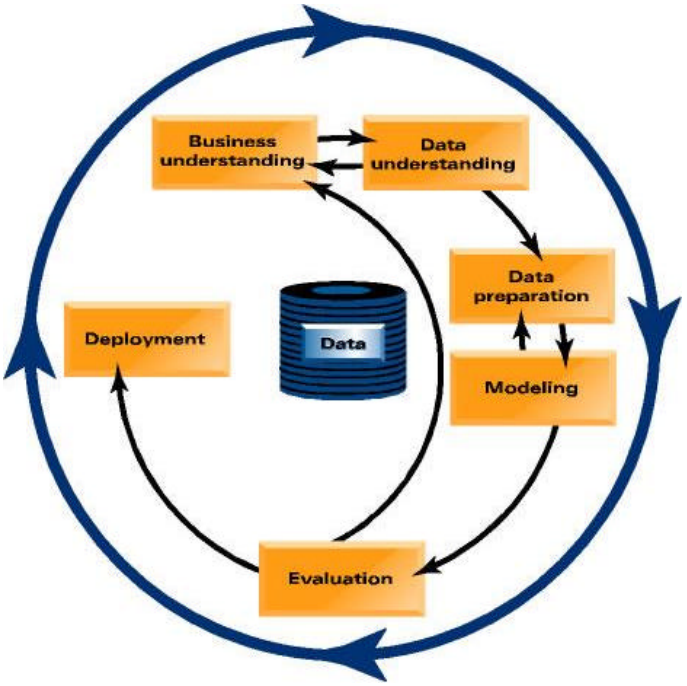

Figure 3: The CRISP-DM reference model consists of six phases [46]

The CRISP-DM model consists of six phases [48]:

- *Business understanding* – Addresses the understanding of the project objectives and requirements from a business perspective
- *Data understanding* – collecting, describing and exploring data for further examination
- *Data preparation* – This covers all the activities in preparation from the initial data format to the final data set to be used for modelling.
- *Modelling* – where modelling techniques are selected and applied
- *Evaluation* – models are evaluated to fulfil the quality requirements
- *Deployment* – including report generation and documentation

Due to the constraints and limitations on the project (see Section 3.2), the CRISP-DM reference model was only loosely followed. The way it was used in this project is discussed in more detail below.

3.5.1 Phase 1: Business understanding

This phase addresses the understanding of the project objectives and requirements from a business perspective. There are several steps involved: determining the business objectives, assessing the situation, determining the data mining goals and a project plan. In this project this was approached by attending several meetings with staff from the industry partner organisation.

3.5.1.1 Determine business objectives

This project was conducted with an industry partner who is a major Australian discount department store retailer with over 250 stores Australia wide. By the very nature of their business, the organisation has a very strong customer focus and they pride themselves on listening to their customers.

Therefore any data mining activity that can improve the customer experience is particularly of value to the partner organisation.

The industry partner was particularly interested in data mining techniques for accessing the success of catalogue advertising campaigns. That is, a comparison between items appearing in the catalogue and those that did not.

3.5.1.2 Assess situation

During this stage, the involvement of the various departments in the catalogue advertising campaigns was assessed in further detail.

The process of determining what will be in the catalogue advertising campaign is an 11 week cycle that involves the both buying office and marketing department. A buyer will create a proposal to justify a place in the catalogue. This proposal would include things such as sales projections.

The decision of which products make it into the catalogue is made by management. Proposals could be rejected if there is not enough markdown budget available, or if the bid is 'out done' by another buyer.

After the campaign has been run, marketing and buying will review the outcomes of the process which will also help determine future advertising allocations.

3.5.1.3 Determine data mining goals

Having determined that the industry partner was particularly interested comparing catalogue and non-catalogue sales, the next step was to determine the matching data mining goals.

One possible way of doing this would be to require that items on sale be flagged as such. However, this would be too presumptuous, and not account for other factors that could potentially influence purchasing patterns. Instead, it was decided that the domain experts would be the best judge of as to why a particular piece of information is interesting.

Thus the data mining goal is to determine comparative information between sales for different days. The business objective of comparison between items appearing in the catalogue and those that did not appear in the catalogue could be achieved by making one (or more) of these days a catalogue sale.

3.5.1.4 Produce project plan

As a research project the project plan was produced in the form of an preliminary research proposal. This included the steps to be taken to achieve the project goals and a research methodology.

3.5.2 Phase 2: Data understanding

The next phase, data understanding, deals with the collecting, describing and exploring data for further examination. Steps in this phase include collecting the initial data, describing the data, exploring the data and verifying the data quality.

3.5.2.1 Collect initial data

The first step of the data understanding phase is data collection. After obtaining a sample data extract from the industry partner, it was determined that only the one data source would be required.

3.5.2.2 Describe data

Once a sample of the store-based transaction data had been acquired, an investigation into the nature of the data was undertaken.

The input data format was a transaction file format. Each store may have many log files, each with many transactions. Each transaction may be stored as many lines. Each line represents a product purchased in a transaction, with each line consisting of a transaction identifier, the item code, department code, quantity sold, unit price and the amount (equal to quantity multiplied by unit price).

An analysis of the data revealed the following properties:

- Approximately 140,000 unique transactions a day;
- Approximately 65,000 unique item codes occurred in the data; and
- Approximately 100 unique department codes occurred in the data.

This information would be used in later phases and decision making.

3.5.2.3 Explore data

This step attempts to answer the data mining question which can be answered using querying, visualisation and reporting [46]. At this stage *Microsoft Excel* was used to graph the distribution of the department attributes. This graph is show in Figure 4, with the different departments across the x-axis and the number of occurrences along the y-axis.

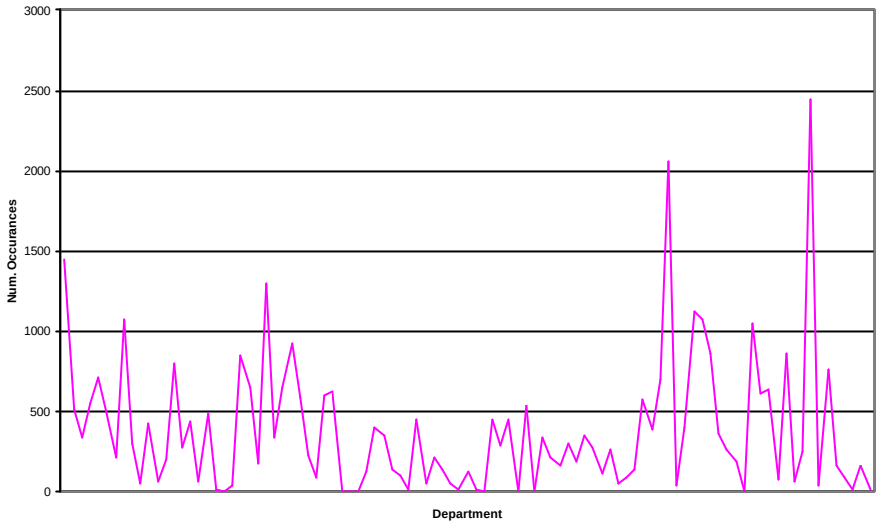

Figure 4: Distribution of departments

However, this brief examination of data did not reveal any initial findings.

3.5.2.4 Verify data quality

The data was obtained from raw transaction logs, such that the information was pretty much straight from the shop register. This meant that the data was of reasonably high quality – the data was fairly complete, was likely to feature few errors and unlikely to contain missing values.

It was impractical to account for the small number of voided or refunded transactions, as this would have involved a large amount of post processing and required considerably more input data than what was available.

The alternative would be to obtain data after it had been though the standard post-processing systems already implemented by the retailer, such as a merchandising system. However, such as system was not available in this case.

3.5.3 Phase 3: Data preparation

Data preparation covers all the activities in preparation from the initial data format to the final data set to be used for modelling. Steps undertaken in this phase include selection of data, cleansing of data, construct data, integration data, and formatting data.

3.5.3.1 Select data

This step involves the selection deciding on the data to be used for analysis. As with any data mining activity, there is the potential that there may be too much data. The dimensions of the data can exceed the capacity of the prediction program, or it may take too long to process and produce a solution [16]. The data selection process, therefore, involves the selection of both attributes (columns) as well as selection of records (rows) with the intention of reducing the size of the data.

Out of the attributes described in section 3.5.2.2, three were chosen to be used in the data mining exercise. The three attributes were the transaction identifier, the item code, and department code. Later, department code was selected over item code, as the granularity was too small at the item code level. Thus the data was aggregated by selecting only the transaction identifier and the department code for further analysis.

Although not planned, one data selection method used included only using two days worth of data. Another was reducing the number of stores involved to six. This reduced the number of transactions and consequently also the number of records. Both of these data reductions were actually due to the data availability constraints described in section 3.2.

3.5.3.2 Clean data

In this step data is cleansed to the required level of quality. However, since the data quality was established to be fairly high in the data understanding phase (See Section 3.5.2.4), no cleansing was performed in this project.

3.5.3.3 Construct data

This task includes constructive data preparation operations such as the production of derived attributes, entire new records or transformed values for existing attributes [46]. Only one operation was required for this project and that was deriving the date from the transaction identifier. This was fairly straight forward as it was a feature of the transaction identifier.

3.5.3.4 Integrate data

These are methods whereby information is combined from multiple tables or records to create new records or values. Since only one data source was used in this research project, there was no need to integrate data sources.

3.5.3.5 Format data

The next task was to transform the data from its original format, which was a raw transaction log format, into the various formats required by the different algorithms. For C4.5 [49] and STUCCO this was a comma separated data (CSV) format and corresponding '.names' file. A names file simply consists of a list of each attribute and the subsequent attribute values, although the names file format used by C4.5 is slightly different to that used by STUCCO. This is because C4.5 requires a class to predict while STUCCO does not.

STUCCO has the extra requirement that the data be separated into separate data files according to the groups. For the purposes of this data mining exercise, the data was split into two files according to the date. This was done using the UNIX *grep* tool.

Magnum Opus [36] has the added advantage that it can also take input from a transaction data format. The transaction data format consists of a transaction id, followed by an item. For the purposes of this analysis exercise, the date was added as the last item so as to determine associations that include the date.

Compared with the transaction data format, the CSV format is highly inefficient for storing this type of data, because each transaction has to indicate if each product was purchased or not. As the number of products increases, the size of the resulting data file explodes. Table 1 shows how the file size can quickly expand beyond the available hardware and software resources.

Number of days	*Number of transactions*	*Number of attributes*	*Estimated size*
1	138,246	65,101	17,167 Mb
2	276,492	65,101	34,335 Mb
3	414,738	65,101	51,502 Mb

Table 1: Estimated CSV filesize at the itemcode level

One way of reducing the size of the data is to reduce the number of attributes, as seen in the selection step (See section 3.5.3.1). In the context of the retail data, this means to group products in some way.

Using information stored in the product hierarchy, the data was aggregated to a higher level, in this case the department level. This allowed the data file size for one day to be reduced to around 27 Mb. The estimated file size for several days is shown in Table 2 below.

Number of days	*Number of transactions*	*Number of attributes*	*Estimated size*
1	138,246	99	27 Mb
2	276,492	99	55 Mb
3	414,738	99	82 Mb

Table 2: Estimated CSV filesize at the department level

This highlights the importance of using a suitable file format for the domain. The CSV format is very inefficient for storing transaction data.

3.5.4 Phase 4: Modelling

During this phase modelling techniques are selected and applied. Steps in this phase include selecting an appropriate modelling technique, generating a test design, building the models and then assessing.

3.5.4.1 Select modelling technique

The selection of a model involves selecting appropriate techniques for the problem, then from those selected, refining to those that meet any political requirements and other constraints. This is illustrated in Figure 5.

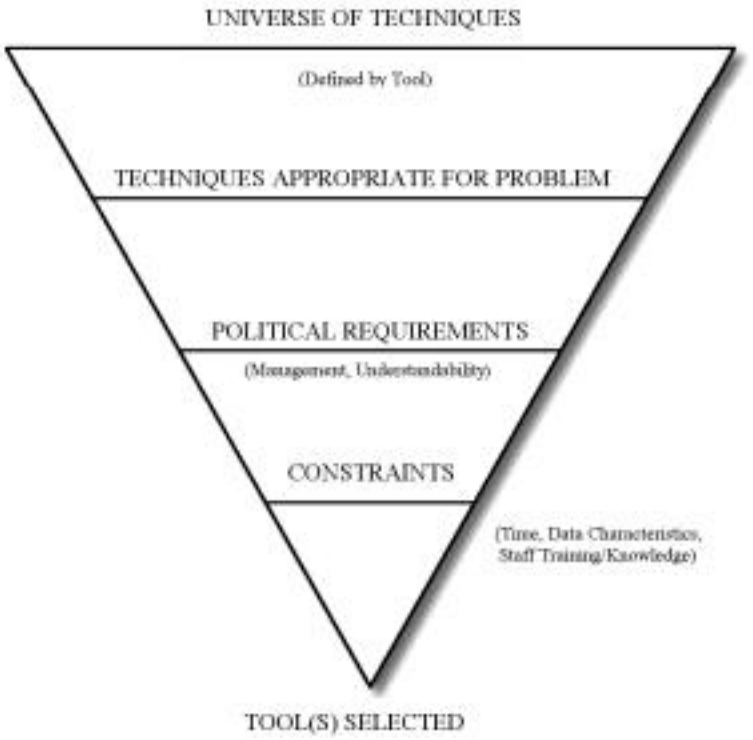

Figure 5: Universe of Techniques [46]

After a broad review of data mining techniques, decision trees, association rules and contrast sets were identified as the data mining techniques that would be used. One of the aims of the research will be to determine whether these methods are well suited to extracting the require knowledge. These methods were introduced in sections 2.3-2.5.

Software was obtained for all three data mining methods. C4.5 [49] was used for decision tree analysis, Magnum Opus [36] for association rule based analysis and STUCCO [50] for contrast sets.

3.5.4.2 Generate test design

This step refers to designing a test to confirm the model's quality and validity. In this project each of the tools used to produce the model handled the dividing of training data, and cross-validation that was taken place was automatically done.

3.5.4.3 Build model

This step involves running the modelling tools to produce the models. In this project it was the application of C4.5, Magnum Opus and STUCCO on the prepared datasets.

3.5.4.3.1 System specifications

The following computer systems were used in the production of these models:

- Dual Sun UltraSPARC-II 450MHz CPU machine with 4Gb RAM and running Sun Solaris 8. This machine was used to obtain the C4.5 and STUCCO results.
- Dual Intel Xeon 1.7Ghz CPU machine with 2Gb RAM and running RedHat Linux 7.3. This machine was used to obtain the Magnum Opus results.

The C4.5 and C4.5 rule generation was performed on the Sun machine. STUCCO was also run on this machine because the binaries obtained were for this architecture. Magnum Opus was run on the Xeon machine because the software license was for this machine.

3.5.4.3.2 C4.5

C4.5 [49] was used to generate a decision tree based on the transaction data, with the date being the class attribute. Then the C4.5 rule generator was then used to convert the tree to a more suitable rule based format.

The profiles of the rules produced by Magnum Opus, STUCCO and C4.5 were very different. Compared to Magnum Opus and STUCCO, C4.5 produced very large rules, the largest rule featured some 51 negative conditions and no positive conditions. A *positive condition* is a when an item is purchased as a part of a transaction and is a part of the market basket. A *negative condition* represents an item that was not purchased and is subsequently missing from the basket.

The large rules produced by C4.5 often had one positive condition and many negative conditions. It is unlikely that so many negative associations would be useful and possibly even confusing to the user. An objective of the questionnaire is to confirm this.

3.5.4.3.3 Magnum Opus

Magnum Opus was selected as the association rule learner because it allows constrained association rule discovery. This is in contrast with a more traditional association rule learner, which are not as flexible.

Magnum Opus [36] was used to generate rules from the data set. The default search method of leverage was used. The first run produced many rules, some featuring the date and some which did not. Since the data mining goal was to compare differences between two dates, interest was restricted to rules which featured the date in the consequent. Magnum Opus was configured to filter out rules that did not feature the date in the rule's consequent. This was done using the *rhs-available* command line option.

3.5.4.3.4 STUCCO

The STUCCO program binary was obtained from Contrast Sets author Stephen Bay and used to produce models of the data.

STUCCO requires that the data is split into separate files according to the groups you wish to compare. This was done using the UNIX tool *grep*, which searched based on the date.

Initially when running STUCCO on the dataset, the program appeared to suffer from a memory blowout. After running for more than 40 hours and using around 2.7Gb of memory the program was stopped.

Correspondence with the program author revealed that the problem may be the feature space being too big to handle. It was suggested that the feature space could be reduced by reducing the number of valid attribute-values from binary (purchased or not purchased) to only use a single attribute value (purchased). This was done by editing the names file such that there was only one possible value for each binary attribute. As an added advantage, this meant that negative contrasts sets, that is products that were not purchased, were not considered.

3.5.4.4 Assess model

At this stage the data miner attempts to assess each of the models, based on knowledge gained about the domain and thus the expected response of the domain experts. A good example is the output produced from C4.5. The resulting rules consisted of many negative conditions and it was not apparent how this would be useful.

3.5.5 Phase 5: Evaluation

In this phase models are evaluated to assess the degree to which the model meets the business objectives and quality requirements. The steps involved are evaluating the results, reviewing the processes and determining the next steps.

The method used to evaluate the results in the project was a questionnaire which featured a sample of information output from each of the models and asked the domain expert to rate it. The questionnaire is discussed in detail in section 3.6.

3.5.6 Phase 6: Deployment

The deployment phase consists of report generation and documentation. The sub-steps involved are planning deployment, planning monitoring and maintenance, producing a final report and reviewing the project.

Since this is a research project, the deployment phase is not relevant. There are currently no plans for deployment, and the final reporting and reviewing is undertaken in the form of this research dissertation.

3.6 Questionnaire

Using the output of the different data mining methods, a questionnaire was prepared to gain feedback from the domain experts and to help partially validate the research with the industry partner.

This post-processing was completed using a number of tools, the main one being Microsoft Excel. STUCCO and MO results were transformed in the UNIX program *vim* before being imported into Microsoft Excel.

However, in order to present the C4.5 rules in the same format as the other methods in the questionnaire, some extra statistics were required and thus had to be manually gathered. This was done using a custom built application (shown in Figure 6). This graphical application was used to construct queries to a database based on the C4.5 rules. The application then produces some summary statistics about each rule, which were then exported as comma separated text and used in Microsoft Excel.

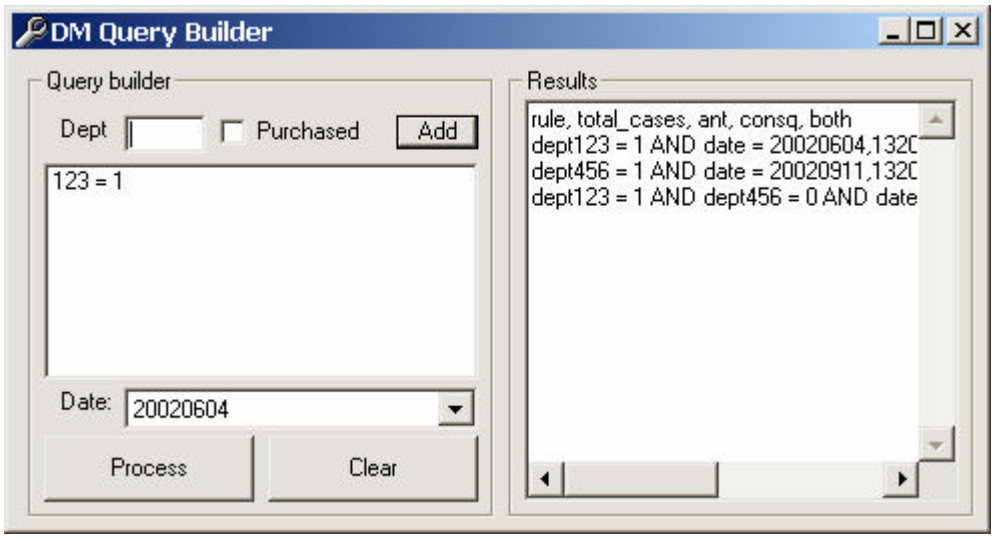

Figure 6: An application was built to extract further statistics from the database.

The questionnaire consisted of 3 sections, one for each analysis method. Each section contained 5 questions in the form of information that had been output from the learning algorithm. The decision to use only 5 questions was based on time and resource limitations. However, many more rules than the five required for the questionnaire were generated, so the selection of which rules made it to the questionnaire was dependent on how the rules were sorted.

Rules generated from the C4.5 rule generator were sorted on their estimated accuracy, with the 5 most accurate rules being used in the questionnaire. Magnum Opus rules were sorted according to their leverage, which is the default search method, and the 5 rules with the highest leverage were selected. STUCCO did not sort its output so they were sorted manually. The 5 rules with the highest relative measure of how many more customers are purchasing those products on a given day were chosen. This measure was manually calculated using the proportion of times the rule is true for one day over the proportion of times the rule is true for the other day.

The outputs from all 3 data mining methods were transformed into the same plain English format. The aim of using a common format is to determine the quality of the information being assessed, rather than the format it is presented in. Figure 7 shows a fictitious example of this format.

```
On July 5th customers were 10.0 times more likely to
purchase items from department 123 (INFANTS; Nappies)
than they were on October 22nd.  It was bought in 1.0%
of transactions on July 5th and 0.1% of transactions
on October 22nd.
```

Figure 7: Questionnaire information format

Many of the rules produced using C4.5 featured items that were not purchased. These rules were presented in a similar format to those produced using the other methods. A fictitious example of this is shown in Figure 8.

```
On October 22nd customers were 5.0 times more likely
to purchase items from department 123 (INFANTS;
Nappies) and nothing from department 345 (BEVERAGES;
Beer) than they were on July 5th.  This occurred in
2.5% of transactions on October 22nd and 0.5% of
transactions on July 5th.
```

Figure 8: C4.5 Questions were slightly different

For each piece of information, respondents were asked to use their knowledge and experience to rate the quality of the information in its understandability, degree of unexpectedness and its potential value to the organisation.

After participants had completed each of the 3 sections, they were asked to rate each of the 3 analysis methods, according to how useful the type of information that the analysis produces is, and how understandable the information is.

3.6.1 Selection of participants

Selection of participants was done using the snow ball sampling method (described in section 3.3.3). The questionnaire was distributed to contacts at the industry partner organisation. Unfortunately this meant it heavily influenced by the time and staff availability resource constraints described in section 3.2. For this reason, there were only three participants.

3.6.2 Design of the data to be captured

The questionnaire was designed with the intended audience of the marketing and buying office departments in mind. Unfortunately the respondents did fully not reflect the intended audience. Out of the three participants, two are classified as *buyers* (someone from the buying office) and the other is an *IT support manager* (from IT department).

3.7 Ethical considerations

Several ethical issues were considered during this research project. The data collected from the industry organisation was obtained under a confidentiality agreement and remained commercial in confidence. It was not of a sensitive nature in that it did not contain customer personal information or identifiers that could be used reference personal information.

The involvement of the staff in the feedback process did not require university human ethics approval since the study involved their professional opinion only.

3.8 Conclusion

This chapter introduced the methodology to be employed in this research project. Constraints including those relating to time, data availability, feedback, software availability and licensing were discussed.

For the research that is going to be undertaken, there is several research method options. These have been discussed and the selected methodology justified. Originally a quantitative study was planned but due to insufficient number of responses a qualitative approach was taken.

CRISP-DM, a data mining methodology used in this research project was introduced and processes described.

Chapter 4: Analysis of data

4.1 Introduction

This chapter will focus on addressing the research questions by presenting, analysing and discussing the associated results. Assessing the quality of the information is first addressed, then the rules generated using each method are examined in detail. Finally, a comparison between contrast sets and association rules is discussed.

4.2 Information quality analysis

A questionnaire was collated to gain feedback and help to answer the research question. In this section the results from this questionnaire are used to analyse the quality of the information produced by these data mining methods from a business perspective. This is in line with the evaluation phase of the CRISP-DM cycle (discussed in section 3.5.5).

The questionnaire was produced by taking the results from the three data mining techniques and putting them into a plain English questionnaire format. This was then given to domain experts at the industry partner organisation.

The questionnaire consisted of 4 sections: information produced using contrast sets, association rules and decision trees, as well as some general feedback about each of the methods. The results have been collected are analysed below.

4.2.1 Interpretation considerations

The output information to which the questions relate is open to interpretation based on the individuals experiences, perceptions and understanding of the domain. It is possible that a participant may give feedback that reflects the actual situation, basing it on their knowledge and experience. It is however, also possible that the information be provided based on their beliefs and thus may not be a true reflection of the actual situation.

This was taken into account when analysing the questionnaire results, as it was not apparent what involvement the IT Support Manager had in the marketing and buying

process. Ideally the questionnaire could have been designed to capture more details about the respondents involvement in the different organisational processes to try to gauge this.

4.2.2 Contrast Sets

The first section of the questionnaire introduced contrast sets to the domain expert. It included a brief description of the contrast set analysis method and asked the user to rate each of the pieces of information. This was on a scale of 1 – 5 according to their assessment of:

- Its understandibility: A value of 1 being the participant cannot understand what this means; to 5 being the meaning is very clear.
- Its degree of unexpectedness: A value of 1 means the information is already well known; to 5 meaning that the information is extremely surprising.
- Its potential value to the organisation: A value 1 meaning no value; to 5 meaning exceptional value.

The responses are given in Table 3.

		CS 1	CS 2	CS 3	CS 4	CS 5	Average
Buyer 1	Understandibility	5	5	5	5	5	5.0
	Unexpectedness	2	2	4	1	3	2.4
	Potential value	4	4	4	4	4	4.0
Buyer 2	Understandibility	5	5	5	5	5	5.0
	Unexpectedness	1	3	3	2	3	2.4
	Potential value	2	2	2	3	2	2.2
IT Support Manager 1	Understandibility	5	5	5	5	5	5.0
	Unexpectedness	2	2	2	2	2	2.0
	Potential value	2	2	2	2	2	2.0
Average	Understandibility	5.0	5.0	5.0	5.0	5.0	5.0
	Unexpectedness	1.7	2.3	3.0	1.7	2.7	2.3
	Potential value	2.7	2.7	2.7	3.0	2.7	2.7

Table 3: Contrast sets results

It is interesting to see that all of the respondents rated understandibility very highly with a rating of 5 out of 5.

The buyers found some of the rules to be more unexpected than others, overall rating an average of 2.4 out of 5, while the IT Support Manager rated all of the rules as 2 out of 5. Out of all the respondents, the average unexpectedness was 2.3.

Potential value was rated high by the first respondent, Buyer 1, while both Buyer 2 and the IT Support Manager rated it as low.

4.2.3 Association Rules

A selection of association rules output from Magnum Opus were presented to the domain experts in a similar format to before. The results of the association rules section are shown in Table 4.

		AR 1	AR 2	AR 3	AR 4	AR 5	Average
Buyer 1	Understandibility	5	5	5	5	5	5.0
	Unexpectedness	2	2	2	2	2	2.0
	Potential value	4	4	4	4	4	4.0
Buyer 2	Understandibility	4	4	4	4	4	4.0
	Unexpectedness	2	2	2	2	2	2.0
	Potential value	2	2	2	2	2	2.0
IT Support Manager 1	Understandibility	5	5	5	5	5	5.0
	Unexpectedness	2	2	2	2	2	2.0
	Potential value	2	2	2	2	2	2.0
Average	Understandibility	4.7	4.7	4.7	4.7	4.7	4.7
	Unexpectedness	2.0	2.0	2.0	2.0	2.0	2.0
	Potential value	2.7	2.7	2.7	2.7	2.7	2.7

Table 4: Association rules results

Similar to the contrast sets results, the domain experts rated understandibility very highly for association rules as well. Buyer 1 and the IT Support Manager rated understandibility 5 out of a possible 5 and the 2nd Buyer rated it 4 out of 5.

All of the participants were consistent in rating the unexpectedness of all of the information items as 2 out of 5.

When it came to the potential value to the organisation Buyer 2 and the IT Support Manager again rated it low, while Buyer 1 again found some high potential.

4.2.4 Decision Trees

The rules produced using C4.5 were post-processed and transformed into the same plain English format as used for the other questionnaire sections.

Unlike the contrast set and association rule methods, the decision tree technique produced rules that also featured negative conditions and were varying in size. The

sizes of the rules in the questionnaire were: DT1 had 1 condition; DT2 had 1 condition; DT3 had 2 conditions; DT4 had 2 conditions; and DT5 had 9 conditions.

The participant's responses have been summarised into Table 5.

		DT 1	DT 2	DT 3	DT 4	DT 5	Average
Buyer 1	Understandibility	4	4	4	3	3	3.6
	Unexpectedness	2	2	3	3	3	2.6
	Potential value	4	4	3	4	1	3.2
Buyer 2	Understandibility	5	5	5	5	5	5.0
	Unexpectedness	1	1	2	2	2	1.6
	Potential value	1	1	2	2	2	1.6
IT Support Manager 1	Understandibility	5	5	5	5	5	5.0
	Unexpectedness	2	2	2	2	2	2.0
	Potential value	2	2	2	2	2	2.0
Average	Understandibility	4.7	4.7	4.7	4.3	4.3	4.5
	Unexpectedness	1.7	1.7	2.3	2.3	2.3	2.1
	Potential value	2.3	2.3	2.3	2.7	1.7	2.3

Table 5: Decision Trees results

Understandibility was again rated very highly by the domain experts, with an overall average of 4.5 overall. It was interesting to see that Buyer 1 rated understandibility between high and medium but rated every other analysis method very high. Further examination of the questionnaire information DT4 and DT5 revealed that DT4 consisted of one positive condition and one negative condition; and that DT5 consisted of one positive condition and eight negative conditions. The reason that Buyer 1 user rated understandibility lower for DT4 and DT5 could have been related to the complexity of these rules.

Both of the buyers rated the last three pieces of information, DT3, DT4 and DT5, of higher unexpectedness than the other information provided. This could have again been related to the complexity, as they may have not expected these products to have a relationship, or a negative one. These rules had one positive condition and at least one negative condition, so they also may not have expected this. The IT Support Manager again rated the unexpectedness of all of the information low.

Buyer 1 found some of the information of high potential value with the exception of DT5 which was rated very low. DT5 was the rule containing 1 positive condition and 8 negative conditions which could have been interpreted by the domain expert as being too specific, or the items unrelated. Consequently, this could have lead to the buyer

being confused as to how this more complex information could be of value to the organisation. Another explanation is that they cannot conceive of how to use the information that something is not purchased.

Unlike Buyer 1, Buyer 2 found DT1 & DT2, which were single item rules, to be of very low potential value to the organisation. DT3, DT4, and DT5 rated slightly higher, at 2 out of 5, which is interesting considering these are the more complex rules. The IT Support Manager again rated the information in DT1-5 as having a low potential value.

4.2.5 Method comparison

The final section of the questionnaire asked the participants to rate each of analysis the methods with respect to:

- How useful is the type of information that this analysis produces? With 1 being not at all; to 5 being extremely.
- How understandable is the type of information that this analysis produces? With 1 being not at all; to 5 being extremely.

The responses are displayed in Table 6. Unfortunately the IT Support Manager did not fill in this section of the questionnaire.

		CS	AR	DT	Average
Buyer 1	Usefulness	4	4	4	4.0
	Understandibility	2	2	2	2.0
Buyer 2	Usefulness	1	2	2	1.7
	Understandibility	4	4	4	4.0
IT Support Manager 1	Usefulness				
	Understandibility				
Average	Usefulness	2.5	3.0	3.0	2.8
	Understandibility	3.0	3.0	3.0	3.0

Table 6: Comparison of analysis methods results

Both the respondents thought the methods were similar as they rated all methods the same (with the exception of one answer).

For usefulness, Buyer 1 said that all of the techniques were highly useful. Buyer 2 scored differently with both Association Rules and Decision Trees rated as low usefulness and the Contrast Set method as very low.

The next question related to the understandibility of the methods themselves. Buyer 1 rated all the methods as low understandibility while Buyer 2 rated all methods as high understandibility.

4.2.6 Summary

The questionnaire has resulted in some interesting results, and there were certainly unexpected information and some potential for business value. At some stages there were mixed responses, even amongst the two buyers.

Unfortunately it is unclear how involved the IT Support Manager is in the relevant company processes. This participant rated understandibility, unexpectedness and potential value the same across every analysis method, and consequently it is hard to know if this respondent actually read the questionnaire in detail.

However, some interesting results were obtained and a summary has been collected into Table 7 below.

	CS	AR	DT	Average
Understandibility	5.0	4.7	4.5	4.7
Unexpectedness	2.3	2.0	2.1	2.1
Potential value	2.7	2.7	2.3	2.6
Average	3.3	3.1	3.0	

Table 7: Summary of results

We can see that all methods rated very highly with understandibility.

The decision trees analysis rated lower on potential value than both the contrast set and decision tree analysis methods. The calculated average for decision trees was 2.3 and contrast sets and association rules were on 2.7.

All round, the contrast sets method rated the highest on an average of 3.3. It was followed by association rule son 3.1 and then decision trees on 3.0.

4.3 Analysis of the rules generated

To answer the research sub-question, "Is one analysis technique any different from another?" a detailed analysis of each of the methods and their results was undertaken.

4.3.1 Quantity

When comparing the methods it is important to look at the quantity of information that is produced by each method. The ideal method would find the 'right' information without leading to an 'information overload'.

The figures for the number of rules produced can be viewed in Table 8 below.

Method	Num. Rules
STUCCO	19
MO	83
C4.5	24

Table 8: Quantity of rules produced using each method

STUCCO and C4.5 produced a similar number of rules. Magnum Opus, however, produced many more. In the coming sections method similarity and other aspects such as complexity will be further examined.

4.3.2 Similarity

In order to determine whether the rules output by the different algorithms were fundamentally different the researcher compared the frequency that each of the rules occurred in the output of each of the data mining programs.

Table 9 shows a summary of frequency of the rules from each method, only the first five of which were used in the questionnaire. The full data is available in Appendix B.

Dept. Label	CS Rule Num.	AR Rule Num.	DT Rule Num.	Total Num. Occurrences
1	CS 18	AR 18, AR 62*, AR 67*	DT 22	3
2	CS 1	AR 5	DT 1	3
3	CS 14	AR 3, AR 63*, AR 65*, AR 69*, AR 82*	DT 15	3
4	CS 4	AR 6	DT 8	3
5	CS 3	AR 8	DT 5	3
6	CS 17	AR 14, AR 52*, AR 60*, AR 63*, AR 66*	DT 17	3
7	CS 2	AR 9	DT 3	3
8	CS 9	AR 10	DT 11	3
9	CS 8	AR 12	DT 14	3
10	CS 7	AR 4, AR 62*	DT 12	3
11	CS 5	AR 1, AR 19*	DT 7	3
12	CS 6	AR 2, AR 19*, AR 51*	DT 9	3
13	CS 11	AR 17, AR 51*	DT 13	3
14		AR 29	DT 19	2
15		AR 36	DT 2	2
16	CS 19	AR 15, AR 39*, AR 60*		2
17		AR 31	DT 20	2
18	CS 16	AR 7		2
19	CS 12	AR 16		2
20	CS 13	AR 21, AR 69*		2
21		AR 47	DT 4	2
22		AR 57	DT 6	2
23	CS 15	AR 22, AR 78*		2
24		AR 24, AR 59*	DT 10	2
25	CS 10	AR 20, AR 61*		2
26		AR 37	DT 24	2
27		AR 67*, AR 71*, AR 73*, AR 75*	DT 23	2

Table 9: Frequency of similar rules

The first column in this table is just a label for the rule. Values in the next 3 columns correspond to information items from contrast sets, association rules and decision tree software. The final column gives a count of the number of times this rule occurred and the results are sorted on this count for presentation purposes.

Association rule items that are followed by a star (*) indicate that there is more than one condition in the antecedent. These rules with more than one condition are often further specializations of the one condition rules.

To illustrate an example, look at the first row in Table 9. Department 1 was the only item in the contrast set CS 18. It was also the only item in the antecedent in AR 18. Department 1 along with department 10 made up the antecedent for AR 62 and department 1 and 27 were used in AR 67's antecedent. The decision tree analysis method produced DT 22 which was made up of one branch on department 1. Although not shown in Table 9, DT 22 also featured 6 negative conditions.

By treating the C4.5 rules as if they did not include negative conditions, similarity of the techniques can be to be compared easier. Using the above example, DT 22 would include items from department 1 and no items from departments 12, 11, 4, 5, 39, and 15. The rules with negative conditions are addressed in Section 4.3.3.

From Table 9 and the more detailed data in Appendix B, it seems that there were 13 rules that were similar across the 3 methods, 14 similar rules across 2 methods, 41 that were unique to that method. Of the 99 attributes, 29 attributes that were not featured in any rules.

Approximately one third of the rules produced bear some similarity. To investigate this further some summary statistics were compiled into Table 10. These were gathered based on the data in Table 9 and Appendix B.

	CS	AR	DT
# unique to this method	0	40	1
# common with another method	6	14	8
# shared with all methods	13	13	13

Table 10: Summary statistics for each of the methods

This table shows the number the number of rules that were unique to each method, common with one other method and shared with all of the other methods.

It is interesting to note that STUCCO is producing rules that are very similar to the other methods, with the majority of rules produced using also being produced by both the other methods. This was 13 of the 19 contrast sets. STUCCO also produced 6 rules that were common to another method and no were unique to just STUCCO.

Magnum Opus achieved all of the rules produced by both of the other methods, with the exception of one rule. The rules produced with Magnum Opus were similar to both methods, however Magnum Opus also output more information about the data – producing 40 rules unique to this method!

Out of all the rules produced, there was only one rule that Magnum Opus did not produce. This rule was DT 16 from the C4.5 results (see Appendix B). Further investigation into why Magnum Opus did not discover this found that it was a fairly weak rule, with a leverage of just 3. Further investigation found that department 68 is in fact company sundries area. This should not theoretically be used, however

departments move items to this code that are not for re-order, or are for deletion from sale, etc. Thus DT 16 did not contain any business value, and it was of no loss that Magnum Opus did not pick this rule up.

The decision trees method, C4.5, was only slightly different to the contrast sets method. They shared 13 rules in common with the association rules method. The decision trees method also produced 8 rules that were also produced by Magnum Opus but not STUCCO and one rule that was unique to just the decision trees method. However, as has already been discussed, this rule did not contain any business value. One possible reason for C4.5 including this fairly weak rule is that it is forced to cover all training examples.

4.3.3 Rule complexity

Rules that are extremely complex have the potential to overwhelm the domain expert, rending the information nearly useless. Data was collected about the sizes of the rules produced using the three different data mining methods and is presented below in Table 11. More detailed data can be found in Appendix C.

	Number of occurrences		
Complexity	**CS**	**AR**	**DT**
1	19	56	5
2	0	23	2
3	0	4	3
>3	0	0	14
Total	19	83	24

Table 11: Complexity of rules produced

These figures show that STUCCO tended to produce contrast sets with a single condition. In fact all 19 of the contrast sets output from STUCCO consisted of only one item, which is very different to the profile of the other methods.

Magnum Opus produced a total of 83 rules, 56 of which had only one condition in the antecedent, 23 with two conditions and 4 with three conditions in the antecedent. There were no rules any larger this produced.

C4.5 tended to produce more complex rules than the other methods, with the majority made up of more than three items and a class. This covers 17 of the 24 rules

produced, and out of the remaining 7 rules, 5 rules were made up of two items and a class and 2 rules of 3 items and a class.

For C4.5, there was usually at least one item as a positive condition, however unlike the other methods, there were often a large number of negative conditions as well. A negative condition means that for a given transaction customers did not purchase a product. This is shown in Table 12 below.

Purchased	Was not purchased	Size	Estimated Accuracy
1	0	2	53.6%
1	0	2	66.2%
1	0	2	82.2%
1	0	2	86.8%
1	0	2	96.2%
1	1	3	84.2%
1	1	3	86.2%
1	2	4	67.2%
1	2	4	71.2%
1	2	4	73.3%
1	3	5	69.6%
1	3	5	77.1%
1	4	6	61.7%
1	5	7	77.7%
1	6	8	56.9%
1	6	8	57.4%
1	6	8	66.6%
1	6	8	71.0%
1	6	8	76.4%
1	8	10	84.2%
1	12	14	55.6%
1	13	15	65.5%
1	44	46	57.1%
1	50	52	62.0%

Table 12: The C4.5 rule generator produced rules varying in size

The first column is the number of items that were positive conditions, that is products were purchased by the customer. Next is the negative conditions, or products that were not purchased or a part of the transaction. These are tallied in column three. The last column is the estimated accuracy produced by C4.5 rule generator.

The table shows that while C4.5 produced some very large rules, this was mainly due to the negative conditions.

Further analysis of the C4.5 results showed that an interesting relationship between the size of the C4.5 generated rules and the estimated accuracy of that rule. The graph

below (Figure 9) shows an interesting trend in the data that appeared when the data was sorted based on the size of the rule. Trendlines have also been added.

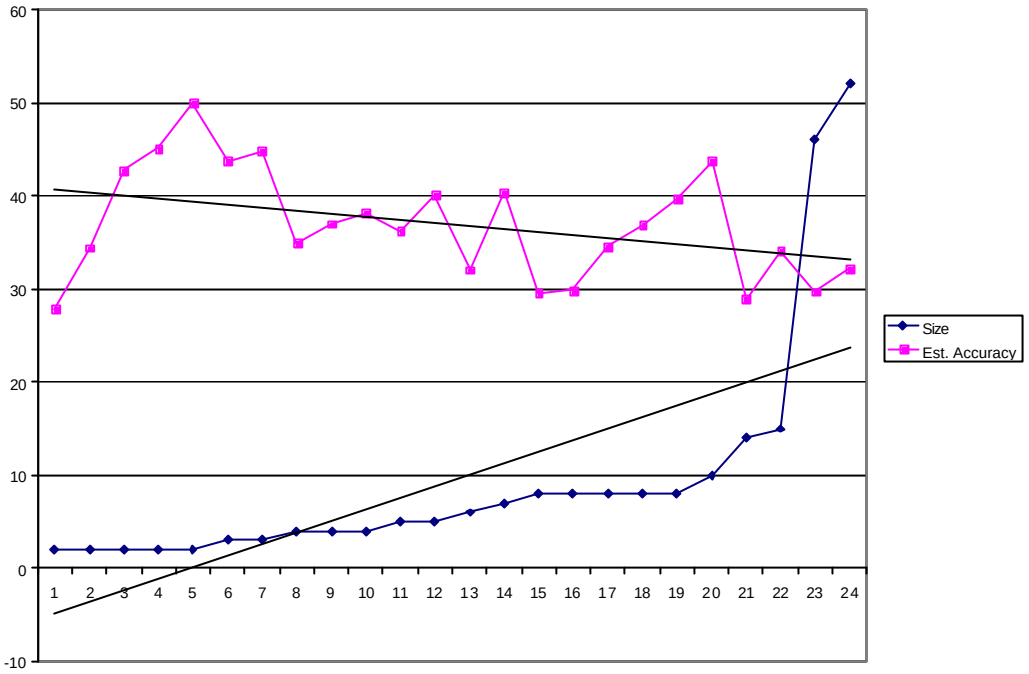

Figure 9: Trend in C4.5 the rules - as the size increases the estimated accuracy decreases

The graph demonstrates a trend developing in the C4.5 results. The estimated accuracy has been plotted on the same plane as the size. This shows the estimated accuracy for a rule is decreasing as the complexity of the rule increases, although not at the same rate.

4.3.4 Summary

In this section, the question of whether one analysis technique is different from the other was looked at. This included analysis of rule quantities, similarity of techniques and rule complexity.

The three data mining techniques produced different quantities of rules. Magnum Opus produced many more rules than the other methods. It output 83 rules while STUCCO and C4.5 output 19 and 24 rules respectively.

Some level of similarity was found between the three data mining methods were, with approximately one third of the rules were similar to those produce by another method in some way. The majority of the rules produced by STUCCO and C4.5 were

also by Magnum Opus. On top of this Magnum Opus also generated 40 rules that were not produced by either of the other methods. Out of all of the rules produced, there was only one which Magnum Opus did not discover.

Rule complexity was also analysed. STUCCO produced smaller sized, simple contrast sets, consisting of only one item. Magnum Opus rules were slightly more complex. Most Magnum Opus rules were of size 1, but there were also rules of size 2 and 3. Majority of the decision tree rules were of a complex nature and size, characterised by usually having one positive item and many negative items. The C4.5 results also revealed an interesting trend: as size of the rule increases the accuracy is decreasing.

4.4 Comparison of contrast sets and association rules

Bay and Pazzani [14] claim that contrast sets differ substantially from association rules, with the difference being that contrast sets are concerned with multiple groups and use different search criteria.

The majority of Bay and Pazzani's comments apply to traditional association rule discovery, however it appears that contrast set discovery can be handled by an extension to association rule discovery that is supported by Magnum Opus.

Constrained association rule discovery, unlike the unconstrained methods described by Bay and Pazzani, operates in a similar fashion to contrast set analysis. The method relies on running normal association rule discovery with the consequent restricted to a variable that distinguishes the groups. From an association rule point of view, this could be represented as $S \Rightarrow G$ where contrast set S implies a difference in group G.

The analysis performed during this research project involved running both contrast set and association rule analysis on the same data. This allowed the results to be compared. It was interesting to see that not only did constrained association rule analysis produce all of the rules that were produced using the contrast set method but it was also simple to do and produced many more rules.

Bay and Pazzani state that the search criterion for contrast sets also differs from that of association rules. However, when an analysis of the similarity of the rules generated

using each method was conducted, this was not apparent (see section 4.3.2). In fact using the default Magnum Opus search criteria of leverage, all of the rules from contrast sets also existed in the association rules results.

Bay and Pazzani suggest two association rule approaches that may be comparable with contrast set learners. The first is to run association rule discovery for each of the groups separately and then compare them [14]. They point out, however, that this would lead to a loss of pruning opportunities and thus efficiency.

Taking the fixed consequent approach there is no need to run the learners separately for each group and pruning is based on the groups the consequent has been restricted to. As a result there is no substantial efficiency loss. The pruning strategy used in this case was based on OPUS search technique (as employed by Magnum Opus). For further discussion of the efficient mining of in association rules using constraints see Srikant *et al.* [30], Ng *et al.* [51], and Bayardo *et al.* [52] and Webb [27, 35].

The second approach Bay and Pazzani suggest would be comparable to contrast sets is to encode the group explicitly as a variable and let an association learner run on this representation. Bay and Pazzani claim that such a method:

- Would not return group differences;
- Would have results that would be difficult to interpret, for example, having too many rules to compare;
- Would not enforce using the same attributes to separate groups, or *consistent contrast* ([53] in [14]); and
- Would require matching of rules to find differences and need a proper statistical comparison to see if differences in support and confidence are significant.

However, by forcing the consequent to be a group, comparing or matching rules would not be required in order to find group differences and consistent contrast is enforced. Also statistical comparisons could be handled by the learning algorithm.

It is also interesting to note that Bay and Pazzani's claim that results from association rule analysis would be difficult to interpret was not reflected in the questionnaire results. Actually, the respondents rated all methods fairly similarly (see Section 4.2).

In summary, Bay and Pazzani's comments about the differences between contrast sets and association rules only apply when running more traditional association rules discovery methods unconstrainedly. Association rules can be efficiently used to mine group differences, with similar results to Bay and Pazzani's contrast set method.

4.5 Conclusion

In this chapter, we have presented a qualitative analysis of the business value of information gained using the three data mining methods. A questionnaire targeted domain experts to get their views about the information produced using each of the techniques.

We have also looked at the abilities and behaviors of the three data mining methods. Included in this discussion was analysis of rule quantities, rule similarity between the different methods and rule complexity.

Finally, we also looked at the relationship of contrast sets to association rules as presented by Bay and Pazzani [14] and suggest constrained association rule discovery as an alternative method for mining group differences.

Chapter 5: Conclusions and future research

5.1 Conclusions

Three alternative data mining methods have been examined: contrast sets, association rules and decision trees. All three of these methods are potentially useful for deriving information from retail store-based transaction data. In this thesis the methods were compared on several different levels. First the quantity of rules produced by these methods was compared and it was found that the contrast set and decision tree methods found a similar number of rules. Association rule analysis found substantially more. Another aspect examined was rule similarity. The contrast sets and decision tree methods did not find many, if any, unique rules, where as association rules matched all of the rules from the other techniques and introduced some unique rules. Rule complexity was also a part of the study. The analysis revealed that contrast sets produced rules with a small number of conditions, association rules produced rules with small-medium number of conditions and decision trees produced rules with mostly large number of conditions. The decision tree results were largely made up of negative conditions of products customers did not purchase.

Another major component of the research study was the feedback gained from the domain experts. A questionnaire was prepared by transforming the results of the various analysis techniques into plain English statements. Although the number of questionnaire results was inadequate for any quantitative analysis, a qualitative approach was adopted. These results suggest the rules produced are easily understandable, may be surprising and have potential for business value.

Contrast set authors Bay and Pazzani [14] claim that contrast sets differ substantially from association rules and that association rule methods are not suitable for mining group differences. Investigating this, the constrained association rule mining method was compared with the contrast set performance. By restricting the consequent to be one of the group variables then it is in fact possible to mine group differences. Bay and Pazzani's comments about the differences between contrast sets and association rules only apply when running more traditional association rules discovery methods unconstrainedly. When suitably constrained association rules can be

efficiently used to mine group differences, with similar results to Bay and Pazzani's contrast set method

5.2 Limitations

The research encountered the following limitations:

- Respondents: Ideally there would have been more respondents, as the questionnaire was planned for participants 30 but only 3 responses were received. This would have allowed for a quantitative analysis or a mixed analysis technique.
- Partner organisation: Only one organisation was available to the researchers. It would have been good to have involved several retail organisations to get a better cross-section of the retail industry. However, this could potentially introduce data privacy issues, or at least multiply the work involved in obtaining the data mining results.
- Only data gathering technique used was a questionnaire: Another technique such as a semi-structured interview could have been used instead of or in conjunction with the questionnaire.

5.3 Future research

There are several areas where future research could be conducted, including:

- Adding a stage to the processes followed in this research project where the data mining expert (in this case the researcher) attempts to use acquired knowledge of the organisations processes to make some qualitative judgements about the quality of the information produced. This would have been a useful analysis for the 40 rules that were produced by the association rule discovery technique but not other methods.
- Conducting a quantitative study, as was originally intended for this research project. Due to the insufficient number of questionnaire responses received a qualitative approach was adopted. Conducting a quantitative study would allow for the validation of some of the key findings in this thesis.
- Research into different ways of data mining for catalogue advertising campaigns. This could make it possible to use more than just the date as a group, for example week number and catalogue number could be interesting.

- Research into other areas with the retail industry where data mining could be applied to bring about some benefit.

5.4 Summary of contributions

There were three main contributions of the research:

- Decision tree learner (C4.5) was found to be unsuitable for this sort of task. The rules produced were often very large, and featuring many negative conditions.
- Of the rules produced by the three methods examined, it was found that roughly one third of the rules being produced were similar to those produced by another method.
- Contrary to Bay and Pazzani's claims, constrained association rule discovery was found to be an efficient and effective method of mining group differences, with similar results. Thus constrained association rule discovery should be considered as an alternative method for mining group differences

References

1. Wilson, R., *Discerning habits*. Marketing Week, 1999. **22**(22): p. 45-48.
2. Hayes, F., *The story so far*. Computerworld, 2002. **36**(16): p. 28.
3. Abend, J., *Keep the customer coming back*. Bobbin, 1999. **41**(1): p. 98-100.
4. Australian Bureau of Statistics, *Retail Trade Australia*. February 2002: Canberra.
5. Australian Retailers Association, *Retail Industry Profile*. 2001.
6. Brijs, T., et al. *Using Association Rules for Product Assortment Decisions in Automated Convenience Stores*. in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1999. San Diego, Calif.
7. Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, *From Data Mining to Knowledge Discovery in Databases*. AI Magazine, 1996. **17**: p. 37--54.
8. Thearling, K., *An Introduction to Data Mining: Discovering hidden value in your data warehouse*. Accessed 17/06/2002 http://www.thearling.com/.
9. Quinlan, J.R., *Induction of Decision Trees*. Machine Learning, 1986. **1**: p. 81-106.
10. Quinlan, J.R., *C4.5: Programs for Machine Learning*. 1993: Morgan Kaufmann.
11. Breiman, L., et al., *Classification and Regression Trees*. 1984: Wadsworth International Group.
12. Agrawal, R., T. Imielinski, and A. Swami. *Mining Association Rules between Sets of Items in Large Databases*. in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*. 1993. Washington, D.C.
13. Bay, S.D. and M.J. Pazzani. *Detecting Change in Categorical Data: Mining Contrast Sets*. in *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1999.

14. Bay, S.D. and M.J. Pazzani, *Detecting Group Differences: Mining Contrast Sets*. Data Mining and Knowledge Discovery, 2001.
15. Friedman, J.H., *Data Mining and Statistics: What's the Connection?* 1997.
16. Weiss, S.M. and N. Indurkhya, *Predictive Data Mining: A Practical Guide*. 1998, San Francisco, CA: Morgan Kaufmann.
17. Witten, I.H. and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. 2000, Sydney: Morgan Kaufmann.
18. Office of the Fedral Privacy Commissioner, *The National Privacy Principles in the Privacy Amendment (Private Sector) Act 2000*. 10/01/2001.
19. Mitchell, T.M., *Machine Learning*. 1997, Sydney: McGraw-Hill.
20. Berry, M.J.A. and G.S. Linoff, *Mastering Data Mining: The Art and Science of Customer Relationship Management*. 2000, Brisbane: John Wiley & Sons.
21. Salzberg, S.L., *Book Review: C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993*. Machine Learning, 1994. **16**(3): p. 235-240.
22. Kass, G.V., *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. Applied Statistics, 1980. **29**: p. 119-127.
23. Gestwicki, P., *ID3: History, Implementation, and Applications*. 1997.
24. Quinlan, J.R., *Improved use of Continuous Attributes in C4.5*. Journal of Artificial Intelligence Research, 1996. **4**: p. 77-90.
25. Rulequest Research, *Information on See5/C5.0: Data Mining Tools See5 and C5.0*. http://www.rulequest.com/see5-info.html Accessed via the Internet 17/06/2002.
26. Buntine, W. *Learning Classification Trees*. in *Artificial Intelligence Frontiers in Statistics*. 1993. London: Chapman & Hall.
27. Webb, G.I. *Efficient Search for Association Rules*. in *The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2000.

28. Webb, G.I., *Association Rules*, in *Handbook of Data Mining*, N. Ye, Editor, Lawrence Erlbaum: To appear.
29. Agrawal, R., et al., *Fast Discovery of Association Rules*, in *Advances in knowledge Discovery and Data Mining*, U.M. Fayyad, et al., Editors. 1996, AAAI Press: Menlo Park, CA. p. 307-328.
30. Srikant, R., Q. Vu, and R. Agrawal. *Mining Association Rules with Item Constraints*. in *Proceedings of the 3rd International Conference on Knowledge Discovery in Databases and Data Mining*. 1997. Newport Beach, California.
31. Piatetsky-Shapiro, G., *Discovery, analysis, and presentation of strong rules*. Knowledge Discovery in Databases, 1991: p. 229-248.
32. Webb, G.I. *Discovering Associations with Numeric Variables*. in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*. 2001: ACM Press.
33. Agrawal, R. and R. Srikant. *Fast Algorithms for Mining Association Rules*. in *Proceedings for the 20th Int. Conf. Very Large Data Bases*. 1994.
34. Borgelt, C., *Apriori (Computer Software).* http://fuzzy.cs.uni-magdeburg.de/~borgelt/ Accessed via the Internet 17/06/2002.
35. Webb, G.I., *OPUS: An Efficient Admissible Algorithm for Unordered Search*. Journal of Artificial Intelligence Research, 1995. **3**: p. 431-465.
36. Rulequest Research, *Magnum Opus (Computer Software).* http://www.rulequest.com/ Accessed via the Internet 17/06/2002.
37. Aumann, Y. and Y. Lindell. *A Statistical Theory for Quantitative Association Rules*. in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*. 1999.
38. Srikant, R. and R. Agrawal. *Mining Quantitative Association Rules in Large Relational Tables*. in *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*. 1996. Montreal, Quebec, Canada.

39. Bay, S.D. and M.J. Pazzani. *Discovering and Describing Category Differences: What makes a discovered difference insightful?* in *Proceedings of the Twenty Second Annual Meeting of the Cognitive Science Society*. 2000.
40. Rubinstein, Y.D. and T. Hastie. *Discriminative vs informative learning*. in *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*. 1997.
41. Bay, S.D. *Multivariate Discretization of Continuous Variables for Set Mining*. in *Proceedings for the International Conference on Knowledge Discovery and Data Mining*. 2000.
42. Bay, S.D., *Multivariate Discretization for Set Mining.* Knowledge and Information Systems, 2001. **3**(4): p. 491-512.
43. Burns, R., *Introduction to research methods*. 4th ed. 1990, Malaysia: Logman.
44. Wiersma, W., *Research Methods in Education*. 7th ed. 2000, Sydney: Allyn and Bacon.
45. Anonymous, *Methodology Manual*. 1995, Texas State Auditor's Office. Available http://www.sao.state.tx.us/Resources/Manuals/Method/.
46. Chapman, P., et al., *CRISP 1.0 Process and User Guide*. 2000, CRISP-DM Consortium (Available http://www.crisp-dm.org/).
47. King, J. and O. Linden, *Data mining isn't a 'cookbook' activity*. National Underwriter, 2002. **106**(39): p. 11-12.
48. Sattler, K. and E. Schallehn. *A Data Preparation Framework based on a Multidatabase Language*. in *Proceedings of the International Conference on Database Engineering and Applications Symposium.* 2001. Grenoble, France: IEEE Computer Society.
49. Quinlan, J.R., *C4.5 (Computer Software).* http://www.cse.unsw.edu.au/~quinlan/ Accessed via the Internet 01/06/2002.
50. Bay, S., *STUCCO 1.0 (Computer Software).* Obtained via email correspondence <sbay@apres.stanford.edu>. 2001.

51. Ng, R.T., et al. *Exploratory mining and pruning optimizations of constrained associations rules*. in *Proceedings of the 1998 ACMSIGMOD International Conference on Management of Data*. 1998. Seattle, Washington.

52. Bayardo, R.J., R. Agrawal, and D. Gunopulos. *Constraint-based rule mining in large, dense databases*. in *Proceedings of the 15th International Conference on Data Engineering*. 1999.

53. Davies, J. and D. Billman. *Hierarchical categorization and the effects of contrast inconsistency in an unsupervised learning task*. in *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. 1996.

Appendix A: Glossary of terms

Association rule discovery	Association rule discovery methods learn relations between variables within a dataset.
Attribute value pair	An occurrence of an attribute, such that it consists of attribute and one of its possible values. For example $A_1 = V_1$ and $A_1 = V_2$.
Contrast Set	Conjunctions of attributes-value pairs that differ meaningfully in their probabilities across several distributions across groups.
Data Cleansing	Process of ensuring that all values in a dataset are consistent and correctly recorded.
Data Mining	Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.
Decision Tree	A decision tree is a graph of choices or decisions. The leaf nodes represent a class.
Itemset	A conjunction of attributes-value pairs.
Noise	Non-systematic errors in either values of attributes or class information.
Rule Confidence	Confidence (or accuracy) is the number of instances that it predicts correctly, expressed as a proportion of all the instances it applies to [8].
Rule Support	Support (or coverage) of an association rule is the number of instances for which it predicts correctly.

Appendix B: Rule similarity

Dept. Label	CS Rule Num.	AR Rule Num.	DT Rule Num.	Total Num. Occurrences
1	CS 18	AR 18, AR 62*, AR 67*	DT 22	3
2	CS 1	AR 5	DT 1	3
3	CS 14	AR 3, AR 63*, AR 65*, AR 69*, AR 82*	DT 15	3
4	CS 4	AR 6	DT 8	3
5	CS 3	AR 8	DT 5	3
6	CS 17	AR 14, AR 52*, AR 60*, AR 63*, AR 66*	DT 17	3
7	CS 2	AR 9	DT 3	3
8	CS 9	AR 10	DT 11	3
9	CS 8	AR 12	DT 14	3
10	CS 7	AR 4, AR 62*	DT 12	3
11	CS 5	AR 1, AR 19*	DT 7	3
12	CS 6	AR 2, AR 19*, AR 51*	DT 9	3
13	CS 11	AR 17, AR 51*	DT 13	3
14		AR 29	DT 19	2
15		AR 36	DT 2	2
16	CS 19	AR 15, AR 39*, AR 60*		2
17		AR 31	DT 20	2
18	CS 16	AR 7		2
19	CS 12	AR 16		2
20	CS 13	AR 21, AR 69*		2
21		AR 47	DT 4	2
22		AR 57	DT 6	2
23	CS 15	AR 22, AR 78*		2
24		AR 24, AR 59*	DT 10	2
25	CS 10	AR 20, AR 61*		2
26		AR 37	DT 24	2
27		AR 67*, AR 71*, AR 73*, AR 75*	DT 23	2
28		AR 79*, AR 82*		1
29		AR 26		1
30		AR 83*		1
31		AR 41		1
32		AR 54		1
33		AR 35		1
34		AR 38, AR 39*		1
35		AR 66*		1
36		AR 53		1
37		AR 58*, AR 74*, AR 82*		1
38		AR 34		1
39		AR 27		1
40		AR 32, AR 52*		1
41		AR 23		1
42		AR 42, AR 72*		1
43		AR 56, AR 81*		1
44		AR 48		1
45		AR 43		1
46		AR 68		1
47		AR 30		1
48		AR 28		1
49		AR 70*		1
50		AR 44		1
51		AR 45		1
52		AR 33, AR 80*		1
53		AR 46, AR 75*		1
54		AR 50		1
55		AR 49		1
56		AR 83*		1
57		AR 11, AR 61*, AR 71*, AR 72*, AR 78*, A		1
58		AR 64		1
59		AR 25		1
60		AR 65*, AR 79*		1
61		AR 74*		1
62		AR 73*		1
63		AR 59*, AR 80*		1
64		AR 13, AR 58*, AR 70*, AR 71*, AR 79*, A		1
65		AR 76		1
66		AR 40		1
67		AR 77		1
68			DT 16	1

Appendix C: Complexity of rules produced

Rule Num.	CS	AR	DT
1	1	1	1
2	1	1	1
3	1	1	2
4	1	1	2
5	1	1	9
6	1	1	1
7	1	1	6
8	1	1	4
9	1	1	7
10	1	1	3
11	1	1	3
12	1	1	7
13	1	1	4
14	1	1	3
15	1	1	7
16	1	1	1
17	1	1	14
18	1	1	51
19	1	2	5
20		1	7
21		1	45
22		1	7
23		1	13
24		1	1
25		1	
26		1	
27		1	
28		1	
29		1	
30		1	
31		1	
32		1	
33		1	
34		1	
35		1	
36		1	
37		1	
38		1	
39		2	
40		1	
41		1	
42		1	
43		1	
44		1	
45		1	
46		1	
47		1	
48		1	
49		1	
50		1	
51		2	
52		2	
53		1	
54		1	
55		2	
56		1	
57		1	
58		2	
59		2	
60		2	
61		2	
62		2	
63		2	
64		1	
65		2	
66		2	
67		2	
68		1	
69		2	
70		2	
71		3	
72		2	
73		2	
74		2	