INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

**Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600**

UMI[®]

**Using data mining to describe the
outcome of web-based negotiations**

by

Anantha Mahadevan, B. Comm.

**A Thesis submitted to the Faculty of Graduate Studies
and Research in partial fulfillment of the requirements
for the degree of**

Masters of Management Studies

School of Business

Carleton University

Ottawa, Canada

May 10, 2000

© Copyright

2000, Anantha Mahadevan

**National Library
of Canada**

**Acquisitions and
Bibliographic Services**

**395 Wellington Street
Ottawa ON K1A 0N4
Canada**

**Bibliothèque nationale
du Canada**

**Acquisitions et
services bibliographiques**

**395, rue Wellington
Ottawa ON K1A 0N4
Canada**

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-52400-0

Canada

The undersigned recommend to the Faculty of Graduate
Studies and Research acceptance of the thesis

**“Using data mining to describe the
outcome of web-based negotiations”**

submitted by Anantha Mahadevan, B. Comm. in partial
fulfillment of the requirements for the degree of Master
of Management Studies

Thesis co-supervisor

Thesis co-supervisor

Director, School of Business

Abstract

This study models the outcome of web-based negotiations. Negotiations are a common aspect of electronic businesses. The items negotiated in these situations are analogous to face-to-face negotiations, with technology and remote access being the key differentiators. Web-based negotiations are being studied by the InterNeg group; one of the projects that has been developed is the INSPIRE negotiation support system.

The objective of this study is to model the outcome of web-based negotiations (that is, whether an agreement was reached or not reached) using data mining. Four data mining techniques are chosen for modeling.

The evidence from this study indicates that the outcome of web-based negotiations may be analogous to face-to-face negotiations. The results indicate issues that should be considered in future research of web-based negotiations. Further, this study also presents the results of model comparison of four data mining techniques.

Acknowledgement

I would like to extend my sincerest gratitude to Professor Kersten for his continuous encouragement, helpful advice and considerable patience through the development of this thesis. His constant input and support was instrumental in the success of this thesis.

I am extremely grateful to Professor Thomas, who through his deep understanding and valuable knowledge of statistics, helped me develop a high respect for the field. His advice helped me develop high quality research.

I am also indebted to Professor Matwin for introducing me to the exciting possibilities in the area of machine learning, and for his ongoing advice. I also thank him for granting me continued access to the resources in his research facility.

Sunil Noronha introduced me to the possibility of research in the field of data mining and data analysis more than three years ago. His ideas, insights, and support of this thesis were invaluable.

A special thanks to my colleagues at the InterNeg research group for companionship and intellectual discussions.

This thesis would not have been possible without the unequivocal support from my parents, Meenakshi Mahadevan and SriVidhya Mahadevan, and my brother, Sundaram Mahadevan. It is to them that I dedicate this work as a token of my affection and appreciation.

Table of Contents

1. Introduction	6
2. Web-based negotiations	9
2.1 Overview	9
2.2 Negotiation process.....	11
2.3 Data collection	13
2.4 Data structure	14
2.4.1 Exogenous variables.....	14
2.4.2 Response variables	15
2.4.3 Intermediate variables	15
2.5 Modeling web-based negotiations	16
3. Data mining	19
3.1 Data mining and knowledge discovery in databases	21
3.2 Implications.....	22
4. Data mining techniques	23
4.1 Association rules analysis	23
4.1.1 Steps in association rules analysis.....	24
4.2 Cluster analysis using simple and iterative k-means methods.....	27
4.2.1 Simple <i>k</i> -means method	27
4.2.2 Iterative <i>k</i> -means method	31
4.3 Naive Bayes method	33
4.3.1 Assessment of naive Bayes method	37
4.4 Entropy-based decision tree method.....	38
4.4.1 Assessment of entropy-based decision tree.....	41
4.5 CHAID decision tree method.....	41
4.5.1 Assessment of CHAID.....	44
4.6 Recommended techniques for data mining web-based negotiations	45
5. Research Objective	48
6. Methodology	50
6.1 Select dependent variable.....	50
6.2 Select independent variables and preprocess	50
6.3 Data mining.....	51
6.4 Model selection and presentation.....	52
7. Describing the outcome of web-based negotiations	53
7.1 Answering the primary objective.....	53
7.1.1 Association rules analysis	55
7.1.2 Naive Bayes.....	59

7.1.3 Entropy-based decision tree	61
7.1.4 CHAID	66
7.2 Integrating models of outcome of web-based negotiations	69
7.2.1 Model from association rules and naive Bayes	70
7.2.2 Models from CHAID and entropy-based decision tree.....	74
7.3 Answering the secondary objectives.....	76
8. Data mining to support statistics: Case of loglinear modeling	90
8.1 Introduction to Loglinear modeling	90
8.2 Results from loglinear analysis	95
9. Conclusions	98
9.1 Key observations from this study.....	99
9.2 Suggestions for future research.....	102
10. References	104
Appendix A: Data Dictionary	110
Appendix B: List of possible independent variables	115
Appendix C: Categorization of variables for loglinear modeling	117
Appendix D: Sample INSPIRE negotiation	118

1. Introduction

Significant research has been conducted on the use of computing tools to conduct and support negotiations (Chatterjee and Lilien 1984; Bui 1994). In such negotiations the parties are aware of each other's identities, and the use of the tool supports face-to-face bargaining. Recent advancements in network technology (such as the Internet and World Wide Web) have accelerated development of net-centric tools (Carroll 1996). Due to the nature of network platforms, users of net-centric tools can communicate with one another remotely. For example, users of on-line auctions (for example, www.ebay.com and www.buy.com) are able to meet on-line, present products, finalize the terms of the trade on-line. As of yet, users cannot yet exchange money for services or products on-line.

The InterNeg group (<http://interneg.org>) is researching negotiations by the development of several inter-related projects. Through these projects, researchers develop and host a comprehensive negotiation support system (NSS) and negotiation software agents on the World Wide Web, and collect and analyze associated data. The negotiation support system provides support for the negotiation process through electronic bargaining facilities, visualization suites, analytical tools, and quantitative and qualitative models (Kersten and Noronha 1999).

One of the current systems developed within the InterNeg project is the INSPIRE negotiation support system (<http://interneg.org/inspire>). The INSPIRE system allows for bilateral negotiations based on a mock business case (which presents a parts manufacturer positioned to supply components to a bicycle producer). As part of the negotiation process, each user is assigned the role of bargainer for either side, and is

not made aware of their opponent's identity. Negotiators then use the system and its facilities to bargain and reach an agreement, during which they may also elect to reveal their identity.

There are few hypotheses that explain the nature of web-based negotiations (Kersten and Noronha 1999), and studies of this social interaction are few. Tukey (1977) and Romesburg (1990) suggest that empirical modeling prior to confirmatory analysis aids the construction of significant hypotheses that may not have been apparent through confirmatory analysis alone. The preceding reasons suggest that there is possibility of extracting new and meaningful models to describe web-based negotiations, as an alternative to constructing hypotheses from literature and domain expertise alone.

In this study, models to explain the outcome of web-based negotiations will be constructed using data mining. The use of data mining provides advantages, such as exploratory analysis and detection of patterns (Chen, Han et al. 1996; Fayyad 1996). Specifically, four data mining techniques will be used in this study. They are:

- a) naive Bayes method
- b) entropy-based decision tree
- c) CHAID
- d) association rules analysis

These methods are typically used on categorical datasets, and have been widely researched and applied (Eisenbeis and Avery 1972; James 1985; Weiss and Indurkha 1998). Further, commercial software implementations exist that can be used to apply these techniques (SGI 1998; SPSS 1998; Cognos 1999; SAS 1999).

The primary objective of this research study is to explain the outcome of web-based negotiations using data mining. A secondary research objective is to compare differences and similarities present in models of web-based negotiations constructed with association rules, naive Bayes, entropy-based decision tree, and CHAID techniques. Six investigative questions are suggested to address the primary research objective. To the knowledge of the author, this is the first study in social research that is proposing to use these four techniques.

This proposal has five subsequent chapters. In the next chapter (Chapter 2), web-based negotiations and data gathered from INSPIRE are discussed. In Chapter 3, relevant issues in data mining and knowledge discovery are presented. In Chapter 4, key data mining techniques are described. The research question is formulated in Chapter 5. The methodology, described in Chapter 6, presents steps for preprocessing, constructing and comparing models. In Chapter 7 the results of modeling are presented. The conclusions, key observations, and next steps from this research study are presented in Chapter 8.

2. Web-based negotiations

2.1 Overview

A negotiation is the process by which at least two parties try to reach an agreement on matters of mutual interest (Hendon, Hendon et al. 1996). It is also defined as an operation where two or more parties with common (and conflicting) interests enter into a process of interaction with the goal of reaching an agreement (preferably of mutual benefit) (Lax and Sebenius 1986). The negotiation process, therefore, allows the parties to redress their needs through mutual communications.

Significant research has been conducted on the use of computing tools to conduct and support negotiations (Chatterjee and Lilien 1984; Noronha and Szpakowicz 1996). Bui (1994) presents situations where computers were used to support face-to-face negotiations. In these experiments, negotiators used a negotiation support system to construct mathematical representations of their subjective preferences. Balakrishnan and Eliashberg (1995) used computer simulations to process offers and counter-offers as negotiations were being conducted. Negotiation support systems (NSS) have also been used to develop efficient outcomes of negotiations, that is final agreements were Pareto-optimal to all negotiating parties. Rangaswamy and Shell (1994) presented results from experiments where NSS was used to improve the negotiated outcomes after formal negotiations were completed. In such negotiations, bargaining is conducted in person, and the use of computing tools, usually, supports communications.

Advancements in inter-network technology (such as the World Wide Web, Gopher, and FTP) have led to the development of net-centric tools (Carroll 1996).

Such tools are now facilitating new forms of communications, including the possibility of communicating remotely. Users of on-line applications and services (such as auction-house www.ebay.com and information portal www.yahoo.com) are able to conduct several tasks on-line. Specifically, commercial activities such as searching for buyers and sellers, searching for products and services, bartering for goods, and finalizing the terms of the trade can be conducted on-line using a myriad of software applications. As of yet, users cannot yet exchange money for services or products on-line. The nature of web-based negotiations, though new, has not received much theoretical assessments. This is a newly researched social phenomenon.

The InterNeg group (<http://interneg.org>) is researching negotiations by the development of several inter-related projects. Through these projects, researchers develop and host a comprehensive negotiation support system (NSS) and negotiation software agents on the World Wide Web, and collect and analyze associated data. The negotiation support system provides support for the negotiation process through electronic bargaining facilities, visualization suites, analytical tools, and quantitative and qualitative models (Kersten and Noronha 1999).

One of the projects developed by the InterNeg group is the INSPIRE negotiation support system. INSPIRE is a negotiation support system developed to meet both training and research requirements (Kersten and Noronha 1999). The system captures data at various points in the negotiation process that is available to researchers to study aspects of web-based negotiations. (A sample negotiation is provided in Appendix D).

Section 2.2 is used to present the negotiation process of the INSPIRE NSS. The media used for data collection in INSPIRE are described in Section 2.3. Section 2.4 describes the structure of the collected data. Sections 2.5 explains the objectives that would be met through a modeling exercise of the INSPIRE dataset.

2.2 *Negotiation process*

Negotiations in INSPIRE are conducted using a business case involving a seller and buyer. This case is a mock scenario where a parts supplier (called Itex Manufacturing) is positioned to supply key components to a bicycle producer (called Cypress Cycles). As part of the negotiation process, participants are assigned the role of bargainer for either side and negotiate on behalf of the company. The participants then use the system to mediate and to reach an agreement that would maximize their party's position. Users are, typically, unaware of their opponent's identity prior to negotiations. They may elect to reveal their identity during negotiations.

INSPIRE negotiations follow three phases: the *analysis*, the *conduct of negotiation*, and the *post-settlement*. The stages are reflective of tasks involved in negotiation studies involving face-to-face negotiations (Kersten, Koszegi et al. 1999).

The *analysis phase* involves an analysis of the situation, the problem and the opponent, formulation of preferences, reservation levels, and strategy. As part of this phase, the user specifies their preferences over four distinct items. These are given in brief below (Kersten and Noronha 1999):

1. The unit price that Cypress would pay to buy the parts from Itex. The user can select one of five options. The set of options is {\$ 3.47, \$ 3.71, \$ 3.98, \$ 4.12, \$ 4.37}.

2. The delivery time of the parts from Itex to Cypress. The user can select one of four options. The set of options is {20 days, 30 days, 45 days, 60 days}.
3. The payment schedule for payment of the parts. The user can select one of three options. The set of options is {payment on delivery, payment 30 days after delivery, payment 60 days after delivery}.
4. The terms for the return of defective parts. The user can select one of three options. The set of options is {full price returns, 75% refund with 5% spoilage, 75% refund with 10% spoilage}

The negotiator provides ratings for each item and, subsequently, for each option available under an item. The option ratings are, typically, dependent on the side the user represents — for example, the buyer may put more emphasis on lower price, faster delivery, longer payback, and favourable terms for returning defective parts. The INSPIRE system uses these preference ratings to build a utility function for the user.

The *conduct of negotiations phase* in INSPIRE negotiations involves the exchange of offers, messages, and offers with messages by users. Offers comprise the negotiated items and their options, such that the participant provides one option per item. The INSPIRE system presents ratings for each offer and counter-offers based on the user's constructed utility function. The user may use these ratings as guides to construct offers and to evaluate counter-offers from their counterparts.

The *post-settlement phase* may be static and involve only the evaluation of the negotiation outcomes generated by, and after, the negotiation activity (Tung 1988). These outcomes include the information about the compromise and the negotiators'

satisfaction. The analysis thus focuses on the evaluation of variables describing outcomes. Furthermore, INSPIRE users have the possibility to improve inefficient compromises.

The negotiation concludes either when the two users reach an agreement, when a negotiator terminates the mediation, or when the deadline has been met.

2.3 Data collection

The INSPIRE system provides two sources of data, which together describe the entire negotiation and the users. They are:

1. two questionnaires that are filled on-line by each negotiator, and
2. the complete computer records of the negotiation.

The first questionnaire is presented in the analysis phase of the negotiation, and is referred to as the “pre-negotiation questionnaire”. The second questionnaire is presented in the post-settlement phase, and is referred to as the “post-negotiation questionnaire”. The pre-negotiation questionnaire contains sixteen questions about the user's background. The questions include the user's date of birth, their country of birth and residence, countries they have lived in, their mother tongue, their prior level of negotiation experience, and other such indicators.

The post-negotiation questionnaire contains questions about the system and the just-concluded bargaining. The questions asked of the user include the user's opinion of the system, the negotiation process, and their counterparts.

The history recording mechanism in the INSPIRE system logs each negotiator's activities in detail and provides complete computer records of the negotiation. This

includes the user's preference ratings on items and options, the interaction measures between the two sides (for example, number of offers and messages exchanged), the length of the negotiations, and the score of the compromise (obtained by applying the negotiator's utility function to the final package). For a detailed breakdown of variables, and their source, please refer to Appendix A.

2.4 Data structure

The variables that describe the INSPIRE negotiations can be classified into three categories: exogenous variables, response variables, and intermediate variables.

2.4.1 Exogenous variables

Three sub-categories of exogenous variables are assumed: *subject*, *task*, and *system variables*.

Subject variables describe information about the subjects in this research. This includes measurements of country of birth, age, gender, occupation, and prior negotiation experience of the user.

Task variables relate to the problems undertaken by the subjects; these center on the negotiation case. Cases from different problem domains, having rich vocabulary, and cases with different agendas have potential for different effects on outcomes. Users are given only one case in INSPIRE, and thus this variable is held constant throughout.

System variables represent the specific functionalities that are available from INSPIRE. This includes the ratings that appear when the user is constructing an

offer, the negotiation history graph that displays the flow of past offers, and the item and option ratings constructed by the user.

2.4.2 Response variables

Three sub-categories of response variables are assumed: measures of *goodness of the negotiation's outcome*, measures of *goodness of the negotiation process*, and measures of the *system's effectiveness*.

Measures of goodness of the negotiation's outcome gauge the effectiveness of the negotiation. This can be assessed by the presence of an agreement, and whether this agreement was efficient.

Measures of goodness of the negotiation process are subjective to each negotiator. Therefore, the same negotiation process can have polar impacts on the two users.

Measures of the system's effectiveness are elicited through the post-questionnaire, where users are asked their perception of various system functionalities, including the negotiation history graph, the score display for offers, and the in-built messaging systems.

2.4.3 Intermediate variables

Intermediate variables are measures that have a contributory or intervening effect on the relationship between independent and dependent variables (Emory and Cooper 1991). Intermediate variables can be classified into three sub-categories: "*psychological baggage*", *behaviour of negotiators during negotiations*, and *perceptions of the negotiations*.

The “psychological baggage” that a user brings to the table is dependent on the user’s personal background and impacts the process and outcome of negotiations. For example, users from cultures where competitive bargaining is the norm may expect greater hostility from counterparts than users from cultures where relationship building is the basis for negotiation.

Users’ behaviours during negotiations have a direct impact on the negotiation process and outcome. For example, exchanging a high number of offers and messages may lead to agreement between the users.

Perceptions of the negotiations may be effected by both the person’s background and by the process and outcome of negotiations. Perceptions of actual control during negotiations, satisfaction with the final agreement, and perceptions of negotiation performance fall within this group of factors.

2.5 Modeling web-based negotiations

The conduct of computer-supported negotiations through the World Wide Web (WWW) is a newly studied social phenomenon. There are few hypotheses that explain the nature of negotiations in this environment and explain the interactions between underlying concepts and constructs. Further, studies of this form of social interaction are few (Kersten and Noronha 1999). Data collected from INSPIRE negotiations present potential for constructing new models of web-based negotiations.

Modeling of newly studied phenomena involves inducting relations from the given dataset(s), and subsequently, validating these relations for the population. Tukey

(1977) and Romesburg (1990) suggest that empirical modeling prior to confirmatory analysis aids in the construction of hypotheses that may be validated on the population. Further, empirical modeling may be used to construct models that explain previously unknown relationships or provide information for future theory construction purposes (typically, predictive and confirmatory modeling) (Tukey 1977). For example, Adler, Brahm et al. (1992) measured the predictive impact of exogenous factors (such as cultural background and personal attributes) on negotiated agreements. Graham et al. (1994) measured bargaining strategies in ten cultures, in experiments of face-to-face negotiations, and concluded that opponent characteristics such as personality affect reaching negotiated agreements in certain cultures. Since web-based negotiations are a newly studied social phenomenon, there are few studies that explain the outcome of web-based negotiations and their quality (for example, efficiency).

Empirical modeling can be conducted to describe the behaviour of other response variables. Bargainers conducting web-based negotiations through the INSPIRE system can improve negotiated outcomes (or negotiated agreements) through the use of a post-settlement mechanism. The post-settlement mechanism is presented to negotiators if there exists possibility for improving either one negotiator's position, without declining the position of the other party. Experiments indicate that negotiators seldom use such facilities in real-life negotiations or experiments (Alemi and Fos 1990; Prasnikar and Roth 1992). Researchers indicate one of the reasons for negotiators' refusal to consider optimal packages is in possible dislike or distrust of a computer generated outcome. In another example, empirical models may be

constructed to describe bargainers that reach an agreement through negotiations, and compare results with bargainers that do not reach agreements.

This thesis proposes to use data mining to explain the outcome of web-based negotiations by considering data describing INSPIRE negotiators. It is to be noted that in this study the term “outcome of web-based negotiations” relates to whether users reached an agreement, improved on their negotiated agreement, or failed to reach an agreement.

This thesis is the first, known to the author, that proposes to use data mining techniques to analyze data from web-based negotiations.

3. Data mining

Data mining is a colloquial term used to address the empirical construction and verification of models from datasets containing a large number of elements (variables and observations) (Glymour, Madigan et al. 1997). The term “mining” is an artifact used to denote the aspect of analyzing data for information that may not have been extracted through the use of confirmatory analysis alone.

Data mining is used to construct summaries of large amounts of data, to identify significant structures and relationships, and to construct predictors of future observations (Harris-Jones 1997). Specifically, data mining is used for various tasks, including:

- predicting the class an observation belongs to;
- predicting the dependent variable value given independent variable values;
- formulating and describing clusters of similar observations;
- describing a group of observations;
- finding and describing relationships and associations among variables;
- identifying deviations and changes;
- identifying variables that control values of other variables.

These tasks are performed through the use of data mining techniques.

Techniques used in data mining, which are typically non-parametric in nature, have been developed by incorporating statistical methods into machine learning (ML) algorithms (Limb and Meggs 1994). Machine learning research has also developed

unique techniques for the purposes of data mining. Examples of data mining techniques include association rules analysis, naive Bayes classification, and Cubist regression (Berry and Linoff 1997).

In the literature, data mining techniques have been contrasted with confirmatory statistical methods (Harris-Jones 1997; Pregibon 1998; Weiss and Indurkha 1998). According to these authors, statistical methods have been devised to be used with pre-specified set of variables, usually of small to moderate size. However, variants of statistical methods (for example, step-wise regression) are used with datasets where large numbers of variables exist, not all of which may be relevant. Since “searching” methods represent an evolution of classical confirmatory methods in the direction of what is now referred to as data mining.

Data mining techniques are, typically, used in the following situations:

1. Data mining is used when the underlying distribution of data is unknown or complex (Glymour, Madigan et al. 1997). This allows for the derivation of significant information without knowledge of exact distributions. It should be noted, though, that several statistical methods, such as General Additive models and non-parametric discriminant analysis, and non-parametric analyses, are also used when the underlying distribution is unknown.
2. Model verification in data mining is conducted through sampling methods such as cross-validation and holdout samples (Kohavi 1995). These verification methods guard against “capitalization on chance” (Stevens 1996). The cross-validation method may provide increased guard against “capitalization on chance” since it is not exposed to the possible bias inherent in hold-out sampling when the training

and test datasets are initially selected.

3. Data mining is used to present data summary. Even if the relationship of a dataset to a population is ill defined, it can still be useful to have a parsimonious model that describes the dataset. This is particularly useful if the dataset is large.

3.1 Data mining and knowledge discovery in databases

Data mining is one stage in the process known as knowledge discovery in databases (KDD) (Fayyad, Piatetsky-Shapiro et al. 1996). The KDD process includes steps to define initial problem and scope, select relevant elements to the chosen problem, standardize data elements, construct model and other summary information, and evaluate and validate information. In general, KDD is used to identify significant patterns in data or to map low-level data into compact, abstract models.

Researchers have studied the use of and provided definitions on the issues that are considered through the use of knowledge discovery. Wirth, Shearer et al. (1997) propose that knowledge discovery is used to extract new rules and patterns from a database. Further, KDD process should include various stages to collect raw data, process data for use by applications, select and use data mining techniques, and evaluate constructed models. They remark that a well-defined methodology eliminates the possibility of bias introduced by the researcher and by selected techniques.

John (1997) states that KDD stages are iterative and interactive. Although there is some canonical order to the process (e.g. collection of raw data precedes other stages), information gathered in one stage may be used to alter assumptions in prior

stages or to define further prior scopes.

Fayyad (1996) observes that knowledge discovery in databases is a nontrivial process of identifying models that are valid, useful, and understandable by domain experts and users. He views research of KDD as a multidisciplinary activity that encompasses techniques and issues beyond the scope of any one particular discipline.

3.2 Implications

Data mining provides the means to conduct exploratory modeling on a dataset. Data collected of INSPIRE users provides a resource that needs to be modeled to understand relationships present in the given dataset, and consequently to understand the underlying social interaction of web-based negotiations.

Chapter 4 presents an introduction to data mining techniques that are available in two software applications. The techniques are presented along with a critical assessment of their benefits and disadvantages.

4. Data mining techniques

Data mining techniques are based on theories of statistics, machine learning, neural networks, fuzzy logic, and genetic algorithms. Techniques used for data mining include association rules analysis, Bayesian classification, neural network classification and clustering, and decision trees (Limb and Meggs 1994; Harris-Jones 1997). This chapter discusses five data mining techniques:

- association rules analysis
- *k*-means cluster analysis
- naive Bayes classification
- entropy-based decision trees
- CHAID (Chi-square Automated Interactive Detection)

4.1 Association rules analysis

Association analysis is used to measure the strength of associations between variables. An association rule represents implication between two variables, that is, the presence of one variable in the database implies the presence of another variable (note: presence of a variable here is measured with observations that have non-zero and non-missing values of that variable) (Brin, Motwani et al. 1997; Megiddo and Srikant 1998).

An association rule is, typically, of format:

$$A \Rightarrow B \quad (1)$$

where A and B denote two variables, and the presence of A (precedent) implies the presence of B (consequent). The method of association analysis is defined below.

Let $I = \{i_1, i_2, \dots, i_M\}$ be a set of M binary variables and let S be a set of N observations, where each observation, s_n ($n = 1, \dots, N$) s_n is measured over each of the M dimensions. With *a priori* null hypothesis of independence between all pairs of i_m , ($m = 1, \dots, M$), and given *a priori* a value of minimum prevalence $\underline{p}$ and minimum predictability $\underline{c}$, association rule method is used to find all association rules that have prevalence and predictability greater than or equal to $\underline{p}$ and $\underline{c}$, respectively. The measure of prevalence for an association rule is the number of observations that take on a value of 1 for variable i_a and variable i_b , as a proportion of all observations in S (where i_a and $i_b \in I$). The measure of predictability for an association rule is defined as the proportion of the consequent to the precedent (more detail is presented in Section 4.1.1). Each association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \neq B$ (that is, A and B can represent either variables or a sub-set of variables in I). The values of $\underline{p}$ and $\underline{c}$ are typically in range $[0,1]$ (Agrawal, Imielinski et al. 1993).

4.1.1 Steps in association rules analysis

Association rules analysis follows three steps:

Step 1: Measure prior probabilities

The prior probability of each variable i_m ($m = 1, \dots, M$) is calculated as the number

of observations that take value of 1 for i_m as a proportion in N . Variables whose prior probability values are less than the *a priori* minimum prevalence, $\underline{p}$, are removed from further consideration.

Step 2: Consider two variables

Consider two variables, i_a and i_b ($i_a \subseteq I$ and $i_b \subseteq I$) that have prior probabilities greater than $\underline{p}$.

Table 4.1: Cross-classification of i_a and i_b

	1	0	Total
i_b	n_{11}	n_{21}	n_{+1}
	n_{12}	n_{22}	n_{+2}
Total	n_{+1}	n_{+2}	N

n_{ij} — Number of observations (counts) corresponding to i th row and j th column
 n_{+j} and n_{i+} — Summation of counts across rows and columns, respectively
 N — Number of observations in S (i.e. $\sum n_{ij}$)

Table 4.1 represents the cross-classification of the two variables. The cell count n_{11} represents the number of observations that are true (that is, observations with non-zero values) for both variables. Association between the two variables could be bi-directional, that is the rules $i_a \Rightarrow i_b$ OR $i_b \Rightarrow i_a$ could be true. For either rule, the prevalence measure would have the same value. The measure of prevalence for an association rule is the number of observations that take on a value of 1 for variable i_a and variable i_b , as a proportion of all observations in S (Megiddo and Srikant 1998), that is:

$$\text{Prevalance} = \frac{n_{11}}{N} \quad (2)$$

In this step, the direction of the rule is not yet determined, that is the cell count n_{11} in Table 4.1 may represent either:

$$\text{Rule 1: } i_a \Rightarrow i_b, \text{ or} \quad (3)$$

$$\text{Rule 2: } i_b \Rightarrow i_a \quad (4)$$

Step 3: Calculate predictability of association rule

The prevalence measure of an association rule indicates the possibility of association between i_a and i_b . However, direction of implication cannot be determined by prevalence measure alone, and thus the predictability of each association rule has to be calculated. The predictability value of an association rule is the conditional probability of observing the consequent variable given the precedent variable. For example, from Table 4.1 two probabilities are $P(i_b|i_a)$ and $P(i_a|i_b)$.

The values of predictability are computed for (3) and (4). The value:

$$r_1 = \frac{n_{11}}{n_{1+}} \quad (5)$$

denotes predictability for (3), and similarly:

$$r_2 = \frac{n_{11}}{n_{+1}} \quad (6)$$

denotes predictability for (4). Valid rules are those where:

$$r_i \geq \text{minimum specified predictability, } \underline{c}. \quad (7)$$

If both r_1 and r_2 exceed $\underline{c}$, the respective rules will be considered “valid”. Thus, associations can be uni- or bi-directional.

4.2 Cluster analysis using simple and iterative k-means methods

A cluster is a set of similar data elements (observations or variables). The mean of observations present in a cluster is referred to as that cluster's center. Cluster analysis is used to elicit schemas on groups (clusters) present in data, such that, elements in each group are similar according to some metric (Sun, Xu et al. 1994). Typically, cluster analysis is used to determine similar groups of data observations.

There are several approaches for cluster analysis; including single link, Ward's, k th neighbour, and self-organizing methods (Romesburg 1990). Approaches to clustering differ in initial cluster assignment and differ in the distance calculations between clusters.

4.2.1 Simple k-means method

Let S be set of M dimensional N observations, such that $S = \{s_i : i = 1, \dots, N\}$, where s_i is denoted as vector:

$$s_i = \langle s_{i1}, s_{i2}, \dots, s_{iM} \rangle \quad (1)$$

composed of M -elements, each element denoted as s_{im} ($i = 1, \dots, N$ and $m = 1, \dots, M$).

Given *a priori* number of clusters k , and set number of iterations L , simple k -means method is used to partition S into k clusters, such that, each cluster, denoted by C_j , is non-empty, that is:

$$C_j \neq \emptyset \quad (j = 1, \dots, k) \quad (2)$$

In addition, all observations in S are decomposed into an exhaustive and mutually exclusive set of clusters, that is:

$$S = \bigcup C_j \quad (j = 1, \dots, k) \quad (3)$$

and:

$$C_g \cap C_j = \emptyset \quad (g \in k \text{ and } j \in k) \quad (4)$$

The simple k -means approach uses the following steps (Sun, Xu et al. 1994).

Initial iteration $\ell = 1$

Step 1: Assign cluster centers

K observations from S are arbitrarily assigned as cluster centers and denoted as c_j .

Therefore, cluster center:

$$c_j = \langle c_{j1}, c_{j2}, \dots, c_{jM} \rangle \quad (j = 1, \dots, k) \quad (5)$$

is composed of M -elements, and each element denoted as c_{jm} ($j = 1, \dots, k$ and $m = 1, \dots, M$).

Step 2: Measure distance between s_i and c_j

The distance between observation s_i and cluster center c_j ($i \in N$ and $j \in k$) is calculated. This distance is represented by vector d_{ij} (i denotes observation s_i ; j denotes cluster center c_j). Each element in d_{ij} is denoted as:

$$d(s_{im}, c_{jm}) \quad (i \in N, j \in k, \text{ and } m = 1, \dots, M) \quad (6)$$

where $d(\cdot, \cdot)$ denotes distance between corresponding elements in two different vectors.

Thus:

$$d_{ij} = \langle d(s_{i1}, c_{j1}), d(s_{i2}, c_{j2}), \dots, d(s_{iM}, c_{jM}) \rangle \quad (i \in N \text{ and } j \in k) \quad (7)$$

Typically in k -means, $d(\cdot, \cdot)$ is a Euclidean measure.

Step 3: Repeat calculation for other cluster centers and for remaining observations

Distance measurement is repeated between s_i and all other cluster centers, and then distance measurement is repeated for every observation.

Step 4: Assign observations to clusters

From (4), it follows that observation s_i can belong to one and only one cluster. The criterion for assigning an observation s_i to cluster C_j rather than to cluster C_g (that is, $s_i \in C_j \wedge s_i \notin C_g$) (Sun, Xu et al. 1994) is:

$$d_{ij} < d_{ig} \quad (8)$$

where C_j denotes cluster with center c_j , and C_g denotes cluster with center c_g ($j \in k, g \in k; g \neq j$). Typically, s_i is assigned to cluster C_j when distance d_{ij} is less than all other distances in S , that is:

$$d_{ij} < d_{iz} \quad (i \in N, j \in k, z = 1, \dots, j-1, \text{ and } j+1, \dots, k) \quad (9)$$

at a given iteration ℓ ($\ell = 1, \dots, L$). (Note: each observation in S is similarly assigned to a cluster.)

Step 5: Re-calculate cluster centers

All cluster centers are re-calculated, such that one center c_j is a measurement of mean of all the observations that belong to cluster C_j . Let s_j denote observation that is assigned to cluster C_j , then each element in c_j , denoted as c_{jm} , is recalculated as:

$$c_{jm} = \left(\frac{\sum_{j=1}^{n_j} s_{jm}}{n_j} \right) \quad (10)$$

where n_j is the number of observations in C_j , and s_{am} denotes element in s_a , such that:

$$s_j = [s_{jm}] \quad (s_j \in C_j \text{ and } m = 1, \dots, M) \quad (11)$$

Step 6: Measure dispersion in C_j

Dispersion in a given cluster C_j is given as:

$$D_j = \sum_{i \in C_j} \frac{d(s_i, c_j)}{n_j} \quad (12)$$

where D_j denotes the overall dispersion in cluster C_j .

Step 7: Reassess and reassign cluster assignments

The membership of an observation to its initial cluster is reassessed. An observation s_h ($s_h \in C_j$) is reassigned to cluster C_b , if:

$$d_{hb} < d_{hj} \quad (13)$$

where $j \in k$, $b \in k$, and $b \neq j$.

Subsequent iterations ($\ell = 2, \dots, L$)

In all subsequent iterations, Steps 5, 6 and 7 from the initial iteration ($\ell = 1$) are repeated. The simple k -means method is terminated when:

$$\sum_{j=1}^k D_j^{\ell+1} \geq \sum_{j=1}^k D_j^{\ell} \quad (14)$$

where D_j^{ℓ} is the dispersion in cluster C_j at iteration ℓ , $D_j^{\ell+1}$ is the dispersion in cluster C_j at iteration $\ell + 1$. The term:

$$\sum_{j=1}^k D_j^{\ell} \quad (15)$$

referred to as global dispersion, denotes summation of dispersions of all clusters in S at iteration ℓ ($\ell = 1, \dots, L$). In the case where the value of global dispersion value decreases with every iteration, the method terminates when $\ell = L$.

From (2) it follows that every observation in S is assigned to one of k clusters.

Thus, outliers in S can increase dispersion in a cluster (SGI 1998). The iterative k -means method allows for the adjustment of the number of clusters k taking into account the global dispersion in S .

4.2.2 Iterative k -means method

Given set $S = \{s_i : i = 1, \dots, N, s_i = \langle s_{i1}, s_{i2}, \dots, s_{iM} \rangle\}$ an interval $[k, \bar{k}]$, and a choice point a , the iterative k -means method is used to find k clusters in S such that $C_j \neq \emptyset$ ($j = 1, \dots, k$). In addition, conditions from equations (2), (3) and (4) are also met.

The iterative k -means method follows four steps.

Step 1: Assign cluster centers

K observations from S are arbitrarily assigned as k initial cluster centers. Each cluster center, denoted by c_j , is an M -dimensional vector of type:

$$c_j = \langle c_{j1}, c_{j2}, \dots, c_{jM} \rangle \quad (j = 1, \dots, k) \quad (16)$$

composed of elements c_{jm} , where $m = 1, \dots, M$, and $j = 1, \dots, k$.

Step 2: Measure distance between s_i and c_j

The distance between observation s_i and cluster center c_j ($i \in N$ and $j \in k$) is calculated. This distance is a vector, denoted by the term d_{ij} , and contains elements of type:

$$d(s_{im}, c_{jm}) \quad (i \in N, m = 1, \dots, M, \text{ and } j \in k) \quad (17)$$

where $d(\cdot, \cdot)$ denotes the distance between corresponding elements in two different vectors.

Step 3: Assign observations to clusters

The criterion for assigning an observation to cluster C_j rather than to cluster C_r

(that is $s_i \in C_j \wedge s_i \notin C_g$) is:

$$d(s_i, c_j) < d(s_i, c_g) \quad (18)$$

where C_j denotes cluster with center c_j and C_g denotes cluster with center c_g , and where $j \in \underline{k}$, $g \in \underline{k}$, and $g \neq j$. Typically, s_i is assigned to the corresponding cluster with lowest value d_{ik} .

Step 4: Measure dispersion in all clusters

Dispersion in a given cluster C_j is given by (12). The dispersion for all clusters in S , C_j ($j = 1, \dots, \underline{k}$), is measured.

Step 5: Find cluster with maximum dispersion

Find a cluster, denoted by C_p , such that $D_p = \text{Max}(D_j)$, where $p \in \underline{k}$, and $j = 1, \dots, \underline{k}$.

Step 6: Arbitrarily split C_p

The cluster C_p is arbitrarily split into 2 halves (C_{p1} and C_{p2}), and each new half is considered a new cluster. The centers for the two new clusters C_{p1} and C_{p2} are calculated as in (7). Thus, the number of clusters in S increases to $\underline{k} + 1$.

Step 7: Reassess and reassign cluster assignments

The membership of an observation to its initial cluster is reassessed. Let s_h be an observation in S such that $s_h \in C_j$ ($j \in \underline{k}$), then s_h is reassigned to cluster C_b , if:

$$d_{hb} < d_{hj} \quad (19)$$

where $j \in \underline{k} + 1$, $b \in \underline{k} + 1$, and $b \neq j$.

Step 8: Iterate until termination

Steps 4, 5, 6, and 7 are repeated until the number of clusters in S equals $\bar{k}$.

Step 9: Calculating k

The final number of clusters, k , is determined by comparing average global

dispersion for a given number of clusters k' ($k' = \underline{k}, \dots, \bar{k}$), which is given by:

$$\frac{\sum_{j=1}^{k'} D_j}{k'} \quad (20)$$

to the choice point a . The final number of clusters, k , is where the value of (20) for a given number of clusters is closest to a . Typically, the choice point is standardized such that it lies in interval $[0,1]$ and a value of 0 typically leads to a choice of $\underline{k}$ final clusters and a value of 1 leads to a choice of $\bar{k}$ final clusters (SGI 1998).

4.3 Naive Bayes method

The naive Bayes method is used to construct classification models that measure observed association between the dependent variable and each independent variable (Mitchell 1997). The constructed model is used to classify future observations into one of known groups. The naive Bayes method is used under the following assumptions:

Assumption 1: Association between the dependent variable and an independent variable is not impacted by the relationship between the dependent variable and another independent variable

and,

Assumption 2: There is no association between independent variables

The naive Bayes method is used to build classification models given the following.

Let S be set of N observations with k ($k \geq 2$) known groups. Let each observation in S be represented by the sets X and Y , where $X \subseteq S$ and $Y \subseteq S$. The set X denotes set of M independent variables:

$$X = \{X_1, \dots, X_M\} \quad (1)$$

and Y denotes the categorical dependent variable, with each value in Y denoting a group in S , such that:

$$Y = \{y_j\} \quad (2)$$

where $j = 1, \dots, k$. An observation in S takes exactly one dependent value:

$$y_g \cap y_j = \emptyset \quad (3)$$

where $y_g, y_j \in Y$, and $g \neq j$.

Let the pair of a independent attribute and its corresponding value be of format:

$$(X_m, x) \quad (4)$$

where $X_m \in X$, and x is a possible value in X_m . For an observation s_n ($s_n \in S$) let the rule:

$$(X_m, x) \cap (X_m, \neg x) = \emptyset \quad (5)$$

be true — that is, an observation takes only one value for every independent variable.

By extension from (1), (2), and (4), s_n is represented as:

$$s_n = \langle a_1, a_2, \dots, a_M, y_j \rangle \quad (6.1)$$

$$= [\bar{a}_n, y_j] \quad (6.2)$$

where a_m ($m = 1, \dots, M$) denotes one pair of a independent variable and its corresponding value (6.1). From (6.2), the elements in s_n that denote pairs of independent attributes and corresponding values be denoted by $\bar{a}_n$ ($n \in N$).

The naive Bayesian task is decomposed into two steps:

- a) the step of model construction, and
- b) the step of classifying new observations to one of k groups – from equation (2).

Step 1: Construct model

The Bayesian model has two components (James 1985). The first is the probability distribution of the dependent variable – denoted by Y . The second component is the conditional probability of an observed independent attribute-value pairing given a dependent variable value, which is calculated as:

$$P(a_i | y_j) \quad (7)$$

Table 4.2: Example of dependent and independent variables

			Y
			y_1 y_2
X_m	1	n_{11} n_{12}	n_{1+}
	2	n_{21} n_{22}	n_{2+}
Total		n_{+1} n_{+2}	n_{++}

n_{ij} — Number of records at the i th row and j th column
 n_{+j} and n_{i+} — Summation of counts across rows and columns, respectively
 n_{++} — Summation along n_{i+} and n_{+j}

Table 4.2 presents an example of a cross-classification between a binary dependent variable (denoted by Y), with possible values of y_1 and y_2 , and an independent variable X_m (where $X_m \in X$) with two possible values. Though this example presents a dichotomous dependent variable, naive Bayes classification can be used on data where the dependent variable contains more than two values.

The naive Bayes method is used to calculate the probability distribution of Y , and the conditional probability distributions. Table 4.3 is used to represent the classification model. The probability distribution of Y is given in the first row. The conditional probability distributions are represented in the subsequent rows.

Table 4.3: Representation of naive Bayes model

Dependent distribution of y	$P(y_1) = \frac{n_{+1}}{n_{++}}$	$P(y_2) = \frac{n_{+2}}{n_{++}}$
Conditional distributions	$P(a_1 y_1) = \frac{n_{11}}{n_{+1}}$ $P(a_2 y_1) = \frac{n_{21}}{n_{+1}}$	$P(a_1 y_2) = \frac{n_{12}}{n_{+2}}$ $P(a_2 y_2) = \frac{n_{22}}{n_{+2}}$

a_i ($i = 1, 2, 3$) represent the attribute value pairs of $(X_m, 1)$, $(X_m, 2)$, and $(X_m, 3)$ respectively

Step 2: Classify a new observation using model

Given an observation:

$$s_u (s_u \in S ; s_u = [\bar{a}_n]) \quad (8)$$

where values only for independent variables are present, the naive Bayes classification model is used to assign the most likely dependent variable value y_j ($y_j \in Y$) to that observation. Most likely dependent variable value is chosen after the conditional probability $P(y_j | s_u)$, ($j = 1, \dots, k$) is calculated for every dependent value.

The calculation is given as:

$$P(y_j | s_u) \quad (9.0)$$

$$= P(y_j) P(s_u | y_j) \quad (9.1)$$

$$= P(y_j) P(\bar{a}_n | y_j) \quad (9.2)$$

$$= P(y_j) P(a_1, a_2, \dots, a_M | y_j) \quad (9.3)$$

$$= P(y_j) \prod_i P(a_i | y_j) (i = 1, \dots, M) \quad (9.4)$$

where (9.1) is given by Bayes Theorem formula (Mitchell 1997), (9.2) is obtained from (9.0), (9.3) is from (7), and (9.4) is a consequence of Assumption 1 and

Assumption 2. The observation s_u is assigned the dependent variable value, y_j , with the highest conditional probability $P(y_j | s_u)$ (Pazzani 1996).

4.3.1 Assessment of naive Bayes method

If the joint probability distribution $P(y_j, s_u)$ ($j = 1, \dots, k$) were easily measurable, then the desired conditional probability in (9.0) would require no calculation. Computing the joint distribution directly is frequently infeasible because of the exponential nature of the distribution; as the number of variables in consideration increases the cost of computing rises exponentially (Ezawa and Norton 1996). To overcome this, naive Bayes method is used with Assumption 1 and Assumption 2, thus replacing the joint distribution with observed pairs of independent variables and corresponding values.

Kohavi (Kohavi, Becker et al. 1998) indicate that the use of naive Bayes has benefits. The method was used to construct models that were robust to high correlation between independent variables; thus violations of assumptions did not alter the model. Using their experimental results, the authors suggest that the assumptions used in naive Bayes provide benefits when modeling real-life data. Researchers have also tested the method on real-life data. Results indicated similar robustness to high correlation between independent variables.

Weiss and Kulikowski (1991) present a second limitation of the use of naive Bayes. According to the authors, Bayes theorem is robust in defining first order dependency, which is of the form $P(a_1 | y_j)$. But in defining second order dependency of $P(a_1 \text{ AND } a_2 | y_j)$ and higher order dependencies the algorithm requires large

sample sizes to make approximations about the populations. Second order dependencies are likely to be sensitive to the assumption of no association between independent variables (Assumption 2). Further, with large sample sizes the efficiency and speed of the method to construct models is significantly reduced (SGI 1998).

4.4 Entropy-based decision tree method

A decision tree method is a classification procedure used to partition data into mutually exclusive, and exhaustive, subsets that best describe the dependent variable (Kass 1980). The subsets are constructed by using small groups of independent variables. The selected independent variables may then be used in further analysis, or in predicting values of the dependent variable. There are several approaches to construct decision tree functions from a set of given data, including CHAID, CART, and binary methods. Well-known applications of entropy-based decision trees technique are present in Quinlan's ID3 and C4.5 algorithms (Quinlan 1993). Entropy based decision tree methods are defined as follows.

Let S be set of N observations with k ($k \geq 2$) known groups. Let each observation in S be represented by the sets X and Y , where $X \subseteq S$ and $Y \subseteq S$. The set X denotes the set of independent variables:

$$X = \{X_1, \dots, X_M\} \quad (1)$$

and Y denotes the categorical dependent variable, with each value in Y denoting a group in S , such that:

$$Y = \{y_j\} \quad (2)$$

where $j = 1, \dots, k$. An observation in S takes exactly one dependent variable value,

that is:

$$y_g \cap y_j = \emptyset \quad (3)$$

where $y_g, y_j \in Y$, and $g \neq j$.

The construction of entropy-based decision tree involves the following three steps.

Step 1: Select the independent variable with strongest significance

Each independent variable in (1) is assumed to have some relationship with the dependent variable (2), such that, partitioning S according to one independent variable (holding all other independent variables constant) creates subsets of data where the dispersion for Y in each subset is less than the dispersion for Y in the original dataset.

The initial task is to measure dispersion in S , and to calculate subsequent decrease in this dispersion when each independent variable is applied. This task is uses the information theory concepts of entropy and information gain (Mitchell 1997).

Entropy is a measure used to compute the dispersion of observations in a given set of data, and is measured as:

$$\text{Entropy}(S) = - \sum p_j \log_2 p_j \quad (4)$$

where p_j is the proportion of observations in group j ($j = 1, \dots, k$) in S . The value of entropy is zero when all observations in S belong to group j , and the value of entropy is one when observations in S are equally distributed among all k groups.

Information gain measures the expected reduction in entropy caused by partitioning S according to one independent variable (Mitchell 1997). The information gain in partitioning S according to X_m , where $X_m \in X$, is given by:

$$\text{Gain}(S, X_m) = \text{Entropy}(S) - \sum_{v \in \text{Values}(X_m)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \quad (5)$$

where $\text{Values}(X_m)$ is the set of all possible values for X_m , and S_v is the subset of S that contains all observations that take value v for variable X_m . In (5), the term $\text{Entropy}(S)$ denotes entropy in the complete set S , and the subsequent term denotes expected value of entropy after S is partitioned using attribute X_m .

The use of information gain favours independent variables with a large number of values (Mitchell 1997). In order to mitigate the choice of variables with large number of values, a modified formula of information gain may be used. This formula is referred to as normalized mutual information, and is given as:

$$\text{normalized mutual information} = \frac{\text{Gain}(S, X_m)}{\log_2(V)} \quad (6)$$

where V is used to denote the number of partitions created by the application of X_m on set S .

The independent variable that maximizes information gain, or normalized mutual information, for the set of data has strongest significance to the dependent variable (Mitchell 1997).

Step 2: Partition S

The independent variable with strongest significance is used to partition S into subsets. Typically, each value of the selected variable is applied to create a corresponding subset of observations.

Step 3: Calculate entropy at the subsets

The entropy value is measured for each subset created by the application of the selected descriptor on S .

Step 4: Check for stoppage condition

If the value of entropy for a subset of data is zero, then that subset is denoted as a terminal node. If, instead, the value of entropy is non-zero, Steps 1, 2 and 3 are repeated.

4.4.1 Assessment of entropy-based decision tree

There is a natural bias in information gain that favours attributes with many values over those with few values (Mitchell 1997; Matwin 1998). This is because as the number of values increases, the number of sub-sets that each value creates also increases (without necessarily adding explainability to the results); with a creation of one extra subset of data, there is potential for reduction in entropy (dispersion) of dependent variable. The use of normalized mutual information reduces this bias; yet, the latter is still biased towards variables with many values. To prove this characteristic of the method, Mitchell (1997) presents an experiment using a unique Date value for every observation in S . His results indicated that the method consistently selected the Date variable as a significant variable, though the variable did not increase comprehension of the dependent variable. Mitchell summarized that the use of entropy-based decision trees with variables with many values should be avoided, or such variables should be recoded into variables with small number of categories.

4.5 CHAID decision tree method

Decision tree models can be constructed using the CHAID (Chi-square automatic

interaction detection) method (Kass 1980). This method uses Pearson's chi-square testing and likelihood-ratio statistic to determine association between the dependent and independent variable. The CHAID decision tree method is given in the following.

Let S be set of N observations with k ($k \geq 2$) known groups. Let each observation in S be represented by the sets X and Y , where $X \subseteq S$ and $Y \subseteq S$. The set X denotes the set of independent variables:

$$X = \{X_1, \dots, X_M\} \quad (1)$$

and Y denotes the categorical dependent variable, with each value in Y denoting a group in S , such that:

$$Y = \{y_j\} \quad (2)$$

where $j = 1, \dots, k$. An observation S takes exactly one dependent value, that is:

$$y_g \cap y_j = \emptyset \quad (3)$$

where $y_g, y_j \in Y$, and $g \neq j$.

The use of CHAID allows for determining significant categories in each independent variable. To determine significance, tests of conditional independence of Y on each pair of independent variable categories are conducted at *a priori* level denoted by α_m . Significant categories are subsequently used in construction of tree branches. The minimum level of significance for conducting chi-square tests between each descriptor and Y is denoted by α_s , which is also provided *a priori*.

The CHAID method is used to determine the levels of association between the dependent variable (2) and independent variables (1), through the following steps in (Ponnuthurai (1998) provides more details).

Step 1: Determine significant descriptor

Step 1-a: Determine significant categories in independent variable

Let X_m be an ordinal variable with c categories ($X_m \in X$). The $c \times k$ contingency table is constructed ($c \geq 2$ and $k \geq 2$). From this contingency table, a $2 \times k$ contingency sub-table is built for c_1 and c_2 (c_1, c_2 are contiguous pair of categories in X_m). The significance of the pair of categories, at level $\underline{\alpha_m}$, is calculated by chi-square valuation. If the conditional distributions of Y are identical for c_1 and c_2 , then these two categories in X_m are merged into one sub-group, denoted by c_{12} .

The tests of conditional independence of Y for all remaining categories of X_m are conducted at level $\underline{\alpha_m}$. The final version of X_m that contains only significant categories to Y is denoted as X'_m , and the contingency table between Y and X'_m is a $c' \times k$ table.

The above considerations are the same for nominal independent variables, with the only difference being that categories under consideration do not have to be contiguous.

Step 1-b: Determine significance of X'_m

The chi-square statistic is used to calculate the level of association, at level $\underline{\alpha_i}$, between Y and X'_m .

Step 2: Repeat Step 1

Step 1 is repeated for all independent variables in (1).

Step 3: Determine variable for splitting S

Independent variables (with significant categories) are considered. The independent variable with strongest significance to Y has highest level of observed association (determined by the p -value for corresponding chi-square test) (Kass 1980).

Step 4: Partition S

The set S is partitioned into mutually exclusive subsets of observations using the independent variable with strongest association to Y. Each subset is constructed by using significant categories in the selected independent variable.

Step 5: Calculate chi-square at the subsets of observations

Significant categories and descriptors are calculated for all subsets of S . Therefore, Steps 1 through 3 are conducted for each subset.

Step 6: Check for stoppage condition

If there are no significant descriptors for a subset of data (that is, variables under consideration produce p -value less than $\underline{\alpha}_s$) then that subset is denoted as a terminal node.

4.5.1 Assessment of CHAID

Pearson's chi-square (X^2) and likelihood ratio (G^2) are asymptotically similar (Agresti 1990). There is very little research conducted, though, that compares the two CHAID measures, and the compares generated models. Further, values of likelihood ratio, with formula:

$$G^2 = 2 \sum n_{ij} \log \frac{n_{ij}}{\mu_{ij}} \quad (4)$$

(n_{ij} is the observed cell count for row i , column j and μ_{ij} is the expected cell count for

a row i , column j) may be considered similar to the value of entropy:

$$\text{Gain}(S, X_m) = \text{Entropy}(S) - \sum_{v \in \text{Values}(X_m)} \frac{S_v}{S} \text{Entropy}(S_v) \quad (5)$$

where S_v denotes data observations with value of v for variable X_m (more detail is provided in Section 4.4). The similarity arises from the use of log-estimations of proportions and of residual values.

4.6 Recommended techniques for data mining web-based negotiations

Five data mining techniques have been presented in this chapter, and each method can be used for descriptive modeling (Agrawal, Mehta et al. 1997; Mitchell 1997; Kohavi, Becker et al. 1998). For this thesis it is proposed that association rules analysis, naive Bayes, entropy-based decision trees, and CHAID be used. The reasons for the selections of these methods are as follows:

1. Many response and intermediate variables in the collected dataset are either ordinal or nominal measurements. This feature of the dataset reduces the significance of models constructed with k -means clustering (which, due to Euclidean calculation of distances, is best used for variables calculated on a scale measurement (Chen, Mangiameli et al. 1995)).
2. Most research and development in data mining has been focused on classification and association rules tasks (Weiss and Indurkha 1998). Thus, research and academic literature contains definitions of techniques and results, and experiments that use the corresponding techniques; and

3. The four methods are proposed for descriptive modeling of web-based negotiations because software applications to perform these techniques are easily available (Harris-Jones 1997). Such software applications include SGI MineSet, SPSS AnswerTree, Cognos Scenario, and SAS EnterpriseMiner.

It is to be noted that association rules are typically used in situations where the objective is to develop rules that satisfy the thresholds for $\underline{c}$ and $\underline{p}$; there is no *a priori* selection of development and independent variables. In this study this method will be used with a constraint (that is, the dependent and independent variables are selected *a priori*). This constraint is placed on the method since it is believed that association rules may describe the associative relationship between the dependent variable (outcome of web-based negotiations or the status of agreement) and the independent variable (or antecedent).

An important aspect of modeling with classification methods is model verification through estimates of classification error (Weiss and Kulikowski 1991; Kohavi 1995). A model is verified by measuring the number of observations that are assigned (through the use of constructed model) incorrect dependent variable values. The assignment of an incorrect dependent variable value to one observation is often referred to as a misclassification (Gordon 1981). The sum of all misclassifications in a model is termed as classification error. Two methods used to estimate classification error include holdout and cross-validation methods (Weiss and Kulikowski 1991; Kohavi 1995). (Note: Direct verification of association rules is not possible)

In the holdout method of model verification, dataset S is partitioned into two mutually exclusive subsets. One subset, denoted by S_t , is used to construct the

classification model, and the other subset, denoted by S_h , is used to estimate the model classification error. Classification error is estimated as the number of observations in S_h that are assigned incorrect dependent variable values as a proportion of total number of observations in S (Kohavi 1995). Classification error is the complement of classification accuracy.

In the cross-validation method, data is partitioned into m ($m \geq 2$) mutually exclusive subsets. The model is constructed in m iterations, where, in each iteration, a different subset is used for estimating classification error of the model constructed using the remaining $m - 1$ subsets. Thus, each observation is used exactly once in model verification and $m - 1$ times in model construction. Averaging classification error estimates, which are measured at each iteration, provides classification error for the final model.

It is to be noted that the techniques will be used through the following software implementations:

- Association rules analysis, naive Bayes, entropy-based decision trees

SGI MineSet version 3.0

- CHAID

SPSS AnswerTree version 1.0

5. Research Objective

In this thesis, Chapter 2 was used to present a case in social research — the conduct of web-based negotiations through net-centric tools. Chapters 3 and 4 were used to present data mining, data mining techniques, and provide an introduction to knowledge discovery in databases.

The primary research objective of this study is:

*"To describe the outcome of web-based negotiations by
conducting data mining on data from INSPIRE users"*

The secondary research objective of this study is:

*"To compare differences and similarities present in models of
web-based negotiations constructed with association rules,
naive Bayes, entropy-based decision tree, CHAID techniques"*

The secondary objective is addressed by the following investigative questions.

1. Are estimates of classification error (using naive Bayes, entropy-based decision tree, and CHAID methods) the same for all models?
2. Are differences between models of the selected dependent variable explainable by the methods that were used?
3. Do the techniques produce stable models when parameters are modified?
4. Do models constructed with the entropy-technique and the model constructed

with likelihood-ratio (G^2) produce similar estimates of classification accuracy?

5. Using CHAID, do models constructed with Pearson's chi-square (X^2) differ from models constructed with likelihood-ratio (G^2)?
6. Does the use of association rules analysis provide models that are similar to the model extracted using naive Bayes?

6. Methodology

The proposed methodology has been adapted from (Fayyad 1996), and is different from the latter in that Fayyad suggests domain understanding and data cleaning as unique stages in KDD. Domain understanding relates to the task of understanding the data and the process that is used to generate the data, while data cleaning relates to several preprocessing tasks such as variable normalization that would enhance data analysis. The task of domain understanding is already conducted as part of research activity in CCAM, and as such is not unique to this study's methodology. The task of data cleaning is presented in this methodology as a subtask of independent variable selection (see Section 6.2).

6.1 Select dependent variable

The dependent variable, outcome of web-based negotiations, will be selected.

6.2 Select independent variables and preprocess

Independent variables may describe the dependent variable. Emory (1991) states that independent variables explain some element of the distributions in the dependent variable. Since the goal of this thesis is exploratory and the nature of the modeling task is to explore new relationships in the domain of web-based negotiations, few restrictions will be placed on the selection of independent variables. These variables will be selected from the set of exogenous and intermediate variables.

Once selected, variables will be processed for "noisy" and out-of-range values.

6.3 Data mining

The four data mining techniques will be used to model the distribution between the dependent and independent variables. With the exception of association rules, cross-validation will be used to estimate the classification error of each model.

Estimates of classification error indicate the relative accuracy with which the model will predict unseen observations from the population (that is, population outside the given dataset) (Weiss and Kulikowski 1991). This estimation is, typically, used in predictive modeling. The use of model classification error is proposed in this study as a baseline measure to compare models. This baseline will be used when different independent variables are present in two different models of the same dependent variable.

Association rules will be constructed with *a priori* values of 20% for $\underline{p}$ (minimum prevalence) and 50% for $\underline{c}$ (minimum predictability).

Naive Bayes classification models will be constructed, and verified using cross-validation.

Entropy-based decision trees will be constructed using information gain and normalized mutual information as the measures of descriptor significance. Further, models will be constructed using the following parameter settings:

- Number of levels in tree. Trees will be restricted to either three or four levels. John (1996) observes that having few levels in decision trees improves comprehension of models while restricting the addition of non-important variables by chance.

- Minimum number of observations required in child nodes. Minimum number of observations required at a child node will be set to either 50 or 25 observations. These two values were selected after empirical pre-processing. Initially, values for minimum number of observations were randomly generated and corresponding models were generated. Based on these initial results, models that “over fit” the dataset or were too general were rejected. It was observed that decision trees with minimum observations of 25 and 50 were potentially more representative.

CHAID decision trees will be constructed using Pearson’s chi-square and the likelihood ratio. Further, models will be constructed using the following parameter settings:

- Minimum significance for merging categories (or α_m): The proposed values of α_m are 5%, 10%, and 20%.
- Minimum significance for selecting descriptors (or α_s): The proposed values of α_s are 5%, 10%, and 20%.

6.4 Model selection and presentation

Data mining will be used to construct several models. The primary and secondary objectives will be addressed using the constructed models.

7. Describing the outcome of web-based negotiations

Negotiations conducted through the INSPIRE negotiation support system can lead to one of three possible outcomes:

- a) Negotiator did not achieve an agreement with their counterpart; the reason may have been a termination by either negotiator or a lapse of the deadline
- b) Negotiator and counterpart agreed upon one package as their final agreement; this agreement was reached in the negotiation phase of bargaining
- c) Negotiator and counterpart agreed upon one package as final agreement; the agreement was improved in the post-settlement phase of bargaining

The variable that is used to measure the outcome of web-based negotiations is nominal. Forty-seven independent variables were considered as possible descriptive variables of this dependent variable — the outcome of web-based negotiations. (Note: the list of variables is presented in Appendix C). The selected dataset includes one thousand seven hundred and fifty-four (1,754) observations. Of these, eight hundred and thirty two (832) negotiators did not reach an agreement, eight hundred and fifty (850) negotiators reached an agreement in the negotiation phase, and seventy-two (72) reached an agreement in the post-settlement phase, typically aided by the post-settlement mechanism in the INSPIRE system.

7.1 Answering the primary objective

The research objectives and corresponding research questions are presented in

Chapter 5. The primary research objective of this study is to construct descriptive models of the dependent variable. Section 7.1 is used to respond to the primary objective, and Section 7.2 is used to respond to the secondary objectives.

The primary objective of this thesis is:

*"To describe the outcome of web-based negotiations by
conducting data mining on data of INSPIRE users"*

Section 7.1.1 presents association rules between independent variables and the dependent variable. Section 7.1.2 includes description of rules extracted using the naive Bayes technique. The results of modeling with entropy-based decision trees are presented in Section 7.1.3. Results from CHAID modeling are presented in Section 7.1.4. Models that describe the outcome of web-based negotiations are integrated and presented in Section 7.1.5.

7.1.1 Association rules analysis

Table 7.1 presents the one rule that describes observations where no agreement was reached, at level $\underline{p}$ (minimum prevalence) = 20% and $\underline{c}$ (minimum predictability) = 50%.

Table 7.1: Association rules that describe observations where no agreement was reached

1.01	Messages = 0	21.0	52.44
------	--------------	------	-------

Rule 2.01 indicates that in 21% of observations, users sent no messages to their counterparts and reached no agreement. This rule may explain that exchanging written messages are an important element in reaching agreements in web-based negotiations.

Association rules analysis was used to extract rules that describe observations where an agreement was reached. These rules are presented in Table 7.2.

Table 7.2: Association rules that describe observations where an agreement was reached

2.01	Gender = Male	20.14	52.69
2.02	Occupation = Student	26.71	56.18
2.03	Knew_opponent_country = No	41.55	54.49
2.04	Knew_opponent_id = No	42.00	54.08
2.05	NSS_before = No	38.92	53.96
2.06	Case_understand = 2	23.00	56.68
2.07	Optimal_score >= 92.75	34.93	53.78
2.08	Activity_deadline = Yes	23.51	55.01
2.09	Offers_2nd_last_day = 0	45.26	58.96
2.10	Offers_3rd_last_day = 0	43.20	57.26
2.11	Offers_5th_last_day = 0	36.18	56.01
2.12	Graph_used = Yes	23.23	68.98
2.13	Graph_inform = Yes	22.48	68.64
2.14	Disclose_country = No	24.65	65.36
2.15	Disclose_id = No	25.62	65.64
2.16	Knew_opponent = No	23.05	65.37
2.17	Surprised = No	22.37	69.88
2.18	Work_w_opponent = Yes	22.83	76.92

Rule 2.01 indicates that in 20% (approx.) observations of the dataset, the negotiator is male and an agreement is reached. This rule may be descriptive of the gender differences inherent in reaching agreements in web-based negotiations. Female users comprise 42% of valid observations (observations where valid responses are collected), while male users comprise 58% of valid observations. Chi-square test (degrees of freedom = 2) between the two variables indicates that no significant relationship exists at the 5% level of significance. Rule 2.01, therefore, may either indicate significant association for one value of the independent variable (Gender), or may be an anomaly.

Rule 2.02 indicates that in 26.7% of observations in the dataset, an agreement is reached and the negotiator is a student. Further, chi-square tests (degrees of freedom = 2) indicate that significant relationship exists between Occupation and Outcome of

web-based negotiations at the 5% level of significance. Therefore, further experiments and controlled laboratories (that allow the researcher to control influence of factors) may reveal stronger interactions.

Rules 2.03, 2.04, 2.14, 2.15, and 2.16 measure the influence of knowledge of opponent identity and nationality on the outcome of web-based negotiations. The five rules present that the negotiator reached an agreement despite being unaware of their opponent's identity or country. These rules may, therefore, be summarized as:

- The knowledge of opponent's identity or country is not an important factor in reaching agreement in web-based negotiations

Rule 2.05 indicates that in 38.9% of observations in the dataset, users have not used a negotiation support system prior to using INSPIRE, yet are able to reach an agreement with their opponent. This may be indicative that the INSPIRE system is "user-friendly" to amateur negotiators, and does not require prior use of negotiation support technology in order to be effectively used in a web-based negotiation environment.

Rule 2.06 indicates that in 23% of observations in the dataset, users reached an agreement and also indicated a high understanding of the case presented to them. This rule may describe that, as in face-to-face negotiations, reaching an agreement is contingent on the parties understanding their tasks — as explained in the Itex-Cypress negotiation case description.

Rule 2.08 indicates that in 23.5% of observations, the user was active forty-eight hours before negotiation deadline (by either sending or receiving messages or offers; or, by logging into the negotiation) and an agreement was reached. This rule may

describe that some agreements are reached by “frantic, last-minute” activities by either side. This hypothesis may be supported by chi-square tests (degrees of freedom = 4), which indicates significant relationship between Activity_deadline and Outcome of web-based negotiations at the 5% level of significance.

In contrast to Rule 2.08, Rule 2.09 presents inconsistency; one of the possible activities that can be conducted forty-eight hours prior to deadline is the transmission of offers. Rule 2.09 may therefore be indicative that transmission of offers forty-eight hours prior to deadline has no impact on reaching an agreement, while other activities such as exchanging written messages may have an impact on the dependent variable.

Rules 2.12 and 2.13 indicate the strength of using the history graph in reaching agreements. Therefore, future studies may focus on determining the strength of using the graph and using it to gather information to predict the outcome of web-based negotiations.

Rule 2.14 presents that in 24.65% of observations, the user’s opponent did not disclose their country and the user reached an agreement. This may indicate that, contrary to Graham et. al (1994), the non-disclosure of the opponent’s nationality does not impact reaching an agreement.

In 22.37% of observations, the user found that their opponent did not surprise them with their actions and an agreement was reached (Rule 2.17). This may suggest that in order to reach an agreement, negotiators must meet their opponent’s prior expectations and not surprise them opponents with actions or messages that seem “outside” those expectations.

Users who agree to work with their opponents on future engagements are also likely to have reached an agreement with them. Graham et al (Graham, Mintu et al. 1994) yield that personal attractiveness (physical appearance and level of comfort) of the opponent has a positive impact on reaching agreements in face-to-face negotiations. This is was deduced from observations that negotiators were more likely to reach an agreement with those they felt a personal connection with. In the case of web-based negotiations, the evaluation that users would be willing to work with their opponents may be an indirect indicator that they were sufficiently “attracted” to their opponent.

7.1.2 Naive Bayes

The naive Bayes method was used to extract rules to describe the outcome of negotiations. The model, developed using three fold cross-validation and with a corresponding estimated classification error of $4.11\% \pm 0.47\%$, is presented in Table 7.3. Only rules with conditions present in at least 20% of observations are considered relevant for the model.

Table 7.3: Naive Bayes model

1	Offers_in_between ≥ 3	Agreement reached (69.3)	33.7
2	Met_expectations = <i>unknown</i>	No agreement (66.2)	35.2
3	Work_w_opponent = Yes	Agreement reached (76.9)	22.9
4	Offers_2 nd _last_day = <i>unknown</i>	No agreement (96.2)	20.0
5	Offers_2 nd _last_day = 1	No agreement (75.0)	57.6
6	Disclose_id = No	Agreement reached (65.6)	25.6
7	Disclose_country = No	Agreement reached (65.4)	24.7
8	Graph_used = Yes	Agreement reached (69.0)	23.3
9	Surprised = No	Agreement reached (69.9)	22.4
10	Negotiation_experience = 3	Agreement reached (58.2)	32.1
11	Knew_opponent_id = No	Agreement reached (54.1)	40.2
12	Case_understand = 2	Agreement reached (56.7)	23
13	Expected_friendliness = 2	Agreement reached (56.1)	20
14	NSS_before = No	Agreement reached (54.0)	39

Rule 1 indicates that a higher number of offers (measured from negotiation commencement until five days before negotiation deadline) are likely to result in an agreement being reached in the negotiation phase. Since interaction between the parties is increased as a result of offer submissions (and exchanges), this rule may indicate that increased interaction is favourable to reaching an agreement in the negotiation phase. The variable *Met_expectations*, presented in the post-

questionnaire, is compulsory; users cannot submit the questionnaire without answering this question. Therefore, a value of *unknown* for this variable identifies users who did not complete and submit their post-questionnaires; the naive Bayes method indicates 35.2% (or 620) such observations.

Users who responded to expect a “friendly” negotiation (prior to the negotiation phase) are also likely to reach an agreement (Rule 13). This may be indicative that prior expectations as to the atmosphere of the negotiation may influence the eventual outcome of the negotiations.

7.1.3 Entropy-based decision tree

Three independent variables are considered significant descriptors of the outcome of negotiations (with 3 levels and a minimum number of 50 observations per parent node). They are as follows (refer to Appendix A for the data dictionary):

- Met_expectations
- Offers_in_between
- Offers_2nd_last_day

Rules generated using the information gain method are presented in Table 7.4.

Table 7.4: Rules from information gain

Rule	Condition	Outcome of negotiations (% likelihood)	Observations in dataset (%)
1	Met_expectations = <i>unknown</i> AND Offers_in_between = (0,1)	No agreement (84.71%)	11.69
2	Met_expectations = <i>unknown</i> AND Offers_in_between ≥ 2 AND Offers_2 nd _last_day = 0	Agreement reached (61.1%)	14.0
3	Met_expectations = <i>unknown</i> AND Offers_in_between ≥ 2 AND Offers_2 nd _last_day ≥ 1	No agreement (72.2)	2.0
4	Met_expectations = 1	Agreement reached (82.83)	2.7
5	Met_expectations = 2	Agreement reached (86.56)	9.2
6	Met_expectations = 3	Agreement reached (82.83)	11
7	Met_expectations = 4	Agreement reached (60.56)	4.9
8	Met_expectations = 5	Agreement reached (63.5)	3.4
9	Met_expectations = 6	No agreement (53.7)	1.7
10	Met_expectations = 7	No agreement (88.89)	4.6

In Rule 1, the variable Met_expectations is presented to users in the post-questionnaire; since it is compulsory, users cannot submit the questionnaire without responding to the question. Observations that record a value of “*unknown*” for this variable are those where the user did not fill in the post-questionnaire. (Note: Users of INSPIRE can elect not to submit a post-questionnaire irrespective of whether they

reach an agreement, terminate negotiations, or allow the deadline date to pass).

Rule 2 presents users who send at least two offers from the beginning of negotiations with until five days remaining in negotiations, and send no offers on the second-to-last day of negotiations AND do not fill in the post-questionnaire, are likely to reach an agreement in the negotiation phase. This may indicate that ignoring to complete the post-questionnaire may not have a significant impact on reaching an agreement in the post-questionnaire; the presence of other variables in the rule may define the outcome of negotiations. Rule 3 may be misleading, especially in comparison to Rule 2. Since the rule is present in only 2% of observations, the collection of more observations may modify the rule.

Users who respond that their prior expectations were poorly met during negotiations are more likely to remain in impasse and not reach an agreement with their opponent (as presented in Rule 9 and Rule 10). Rule 4 through Rule 8 indicate that for other responses of the variable *Met_expectations*, users are likely to reach an agreement in the negotiation phase.

Rules generated using normalized mutual information (with 3 levels and a minimum number of 50 observations per parent node) are presented in Table 7.5. The use of this method indicates that the following four variables are significant in describing the outcome of a web-based negotiation:

- *Offers_in_between*
- *Work_w_opponent*
- *Offers_2nd_last_day*

- Opp_cooperate

Table 7.5: Rules from normalized mutual information

Rule	Antecedent	Outcome of negotiations (% likelihood)	Observations in dataset (%)
1	Offers_in_between ≤ 1 AND Work_w_opponent = <i>unknown</i>	No agreement reached (84.71)	11.7
2	Offers_in_between ≤ 1 AND Work_w_opponent = No	No agreement reached (68.18)	1.7
3	Offers_in_between ≤ 1 AND Work_w_opponent = Yes	Agreement reached (59.38)	2.17
4	Offers_in_between ≥ 2 AND Work_w_opponent = <i>unknown</i> AND Offers_2 nd _last_day = 0	Agreement reached (60.9)	13.85
5	Offers_in_between ≥ 2 AND Work_w_opponent = <i>unknown</i> AND Offers_2 nd _last_day ≥ 1	No agreement (72.22)	2.2
6	Offers_in_between ≥ 2 AND Work_w_opponent = No AND Opp_cooperate = 3	Agreement reached (62.5)	1.7
7	Offers_in_between ≥ 2 AND Work_w_opponent = No AND Opp_cooperate = 4	Agreement reached (52.0)	2.2
8	Offers_in_between ≥ 2 AND Work_w_opponent = No AND Opp_cooperate = 1	Agreement reached (68.75)	0.63
10	Offers_in_between ≥ 2 AND Work_w_opponent = No AND Opp_cooperate = 2	Agreement reached (77.5)	2.17
11	Offers_in_between ≥ 2 AND Work_w_opponent = No AND Opp_cooperate = 5	Agreement reached (70.15)	2.68
12	Offers_in_between ≥ 2 AND Work_w_opponent = Yes	Agreement reached (79.34)	20.58

The variable Work_w_opponent is present in the antecedent of Rule 1. This variable is presented to users in the post-questionnaire; since it is compulsory, users cannot submit the questionnaire without responding to the question. Observations

that record a value of “*unknown*” for this variable are those where the user did not fill in the post-questionnaire. Users who send one or no offer from the beginning with until five days remaining in the negotiations and do not fill in the post-questionnaire are more likely not to reach an agreement.

Rule 2 may indicate a hypothesis that users who send few offers combined with a desire to not work with their opponent (alluding a lack of collegiality between negotiators) are more likely not to reach an agreement. This compares with Rules 6 through 11 that indicate that a high number of offers in between (of at least two) are more likely to reach an agreement even though the user is not willing to work with their opponent in future engagements. In compliment to Rule 2, Rule 3 indicates that users who send few offers from the beginning with until five days remaining in the negotiations are likely to reach an agreement if they are willing to work with their opponent in a future engagement.

Rule 5 is misleading in comparison to Rule 4. Since this rule is present in only 2.2% of observations in the dataset, it may be a spurious. Further collection of observations may remove this rule from the list of significant rules.

Rules 12a and 12b are presented in combination, since there is an apparent rule that is present: Users who send at least 2 offers before the fifth-last day of negotiations and would like to work with their opponent on some other project, are likely to reach an agreement (either that agreement is reached in negotiation stage or it is reached in the post-settlement stage).

7.1.4 CHAID

Five independent variables are considered significant descriptors of Outcome of web-based negotiations. As presented in Table 7.6 these variables are as follows (refer to Appendix A for definitions):

- Offers_5th_last_day
- Instructions_easy
- Work_w_opponent
- Offers_in_between
- Activity_deadline

Seven rules are presented in Table 7.6. Only Rule 7 presented in Table 7.6 universally implies a single outcome — no agreement reached.

Table 7.6: Rules from Pearson's chi-square at level 5% for m and s

Rule	Rule antecedent	Outcome of negotiations (% likelihood)	Observations in dataset (%)
1	Offers_5 th _last_day = [0,1,2,3,4] AND Instructions_easy = [1,2,3,4,5,6,7] AND Work_w_opponent = No	Agreement reached (49.8%)	8.49%
2	Offers_5 th _last_day = [0,1,2,3,4] AND Instructions_easy = [1,2,3,4,5,6,7] AND Work_w_opponent = Yes	Agreement reached (76.88%)	22.75%
3	Offers_5 th _last_day = [0,1,2,3,4] AND Instructions_easy = <i>unknown</i> AND Offers_in_between = [0,1], <i>unknown</i>	No agreement (84.77%)	11.74%
4	Offers_5 th _last_day = [0,1,2,3,4] AND Instructions_easy = <i>unknown</i> AND Offers_in_between = (1,3]	Agreement reached (49.28%)	7.75%
5	Offers_5 th _last_day = [0,1,2,3,4] AND Instructions_easy = <i>unknown</i> AND Offers_in_between = (3,17]	Agreement reached (67.80%)	6.84%
6	Offers_5 th _last_day = <i>unknown</i> AND Activity_deadline = [No, Yes]	No agreement (93.53%)	7.41%
7	Offers_5 th _last_day = <i>unknown</i> AND Activity_deadline = <i>unknown</i>	No agreement (100.0%)	5.76%

Comparing Rule 1 and Rule 7 indicates that Pearson's chi-square identifies two distinct groups: (1) observations that have a valid value corresponding to the variable Offers_5th_last_day, and (2) observations that do not have a valid value corresponding to this variable. This variable is evaluated as *unknown* only if users have already completed negotiations with five days until the negotiation deadline (either through termination of negotiations or through reaching an agreement).

Comparison of Rule 1, Rule 2 against Rule 3, Rule 4, and Rule 5 presents a dichotomy of observations in relation to the variable Instructions_easy. This variable is measured in the post-settlement phase through the use of the post-questionnaire. The post-questionnaire is designed such that users have to provide valid values for

compulsory questions; a value *unknown* is not valid. Further, users cannot submit as final their post-questionnaire without having answered all compulsory questions. The variable *Instructions_easy* is used to measure the ease with which users found instructions in the INSPIRE system, and is a compulsory question. Therefore, the value *unknown* for this variable indicates users who did not fill in their post-questionnaires.

Table 7.7 presents rules extracted using the likelihood ratio method in CHAID, at level 5% significance for α_m and α . The use of this method indicates that the following five variables are significant in describing the outcome of a web-based negotiation.

- *Offers_5th_last_day*
- *Instructions_easy*
- *Work_w_opponent*
- *Offers_in_between*
- *Activity_deadline*

Table 7.7: Rules from likelihood ratio

Rule	Conditions	Outcome of negotiations (% likelihood)	Observations in dataset (%)
1	Offers_5th_last_day = [0,1,2,3,4] AND Instructions_easy = [1,2,3,4,5,6,7] AND Work_w_opponent = No	Agreement reached (49.8%)	8.49%
2	Offers_5th_last_day = [0,1,2,3,4] AND Instructions_easy = [1,2,3,4,5,6,7] AND Work_w_opponent = Yes	Agreement reached (76.88%)	22.75%
3	Offers_5th_last_day = [0,1,2,3,4] AND Instructions_easy = <i>unknown</i> AND Offers_in_between = [0,1]	Agreement reached (84.77%)	11.74%
4	Offers_5th_last_day = [0,1,2,3,4] AND Instructions_easy = <i>unknown</i> AND Offers_in_between = (1,3]	Agreement reached (49.28%)	7.75%
5	Offers_5th_last_day = [0,1,2,3,4] AND Instructions_easy = <i>unknown</i> AND Offers_in_between = (3,17]	Agreement reached (67.80%)	6.84%
6	Offers_5th_last_day = <i>unknown</i> AND Activity_deadline = [No, Yes]	No agreement (93.53%)	7.41%
7	Offers_5th_last_day = <i>unknown</i> AND Activity_deadline = <i>unknown</i>	No agreement (100.0%)	5.76%

The rules generated by likelihood ratio are identical to the rules extracted using Pearson's chi-square. According to Agresti (1996), the Pearson's X^2 and likelihood-ratio G^2 share many properties and commonly yield the same conclusions. When the sample size is large the two statistics have the same chi-squared distribution. This inspection is observed for the INSPIRE dataset; the rules extracted are identical as indicated in Table 7.6 and Table 7.7.

7.2 Integrating models of outcome of web-based negotiations

The use of data mining techniques has provided seven models. The different models provide several individual responses to a single modeling question, yet there is need

for assessing the information in a single model. In this section, these models are integrated into two comprehensive models:

1. integrated rules from association rules analysis and naive Bayes method, and
2. integrated rules from entropy-based decision trees and CHAID.

Two models are used given the different nature of rules from the different techniques. Association rules and naive Bayes rules are presented in the first order format (such as Condition X implies Y). Contrary to this, CHAID rules and entropy-based methods rules are presented in conjunct format (that is, Condition X AND Condition Z imply Y).

7.2.1 Model from association rules and naive Bayes

Kersten, Koszegi, and Vetschera (1999) provide a framework of constructs that can be used to measure negotiations. Their framework is comprised of six constructs based on three bipolar characteristics: exogenous vs. endogenous, subjective vs. objective, and individual vs. group. These constructs are:

1. Exogenous characteristics of the negotiator (exogenous, objective, individual);
2. Situational constraints of the negotiator (exogenous, objective, individual);
3. Atmosphere during negotiations (endogenous, subjective, group);
4. Negotiation process (endogenous, objective, group);
5. Results of negotiations (endogenous, objective, group); and
6. Negotiator's expectations and assessment of the process, results, opponent and

oneself (endogenous, subjective, individual).

These six constructs were used to combine models from association rules analysis and naive Bayes. Figure 1 presents the results of the combination; variables that appear in multiple models are considered only once in this graph. This figure presents rules that describe observations where no agreement is reached.

Figure 7.1: Association and Bayesian rules describing observations where no agreement reached

Figure 7.1 indicates that variables that measure the constructs "Negotiation Process" and "Negotiator's expectations and assessments" describe observations where no agreement is reached. There is no process to measure the impact of exogenous constructs ("Exogenous characteristics" and "Situational constraints of the

negotiator”) on these two constructs. This may indicate the additive influence of the former on the latter. This information would be necessary to model “Negotiation Process” and “Negotiator’s expectations and assessments” as intermediate constructs in a complete model of web-based negotiations.

Figure 7.2: Association and Bayesian rules describing observations where agreement reached in negotiation phase

The arrows in Figure 7.1 and Figure 7.2 indicate that “Negotiation Process” and “Negotiator’s expectations and assessments” are significant descriptors of the outcome of negotiations; the arrows do not indicate predictive relationships. Variables that measure the other three constructs do not have significant descriptive relationships with the outcome of negotiations. This result indicates difference in

other studies on negotiations. In experiments conducted by Graham et al. (Graham, Mintu et al. 1994) results indicated that the exogenous characteristics of a negotiator have significant descriptive relationship with the outcome of negotiations, and on whether an agreement was reached. In the results of a study conducted by ABC (year), the situational constraints of the negotiator (including the position being enacted — buyer or seller) determined the nature and scale of the negotiated outcome.

The results presented in Figure 7.1 and Figure 7.2 are subject to the following assumptions:

- The associative impact of “Exogenous characteristics”, “Situational constraints of the negotiator”, and “Atmosphere during negotiations” on “Negotiation process” “Negotiator’s expectations and assessment of the process, results, opponent and oneself” cannot be measured. Since the use of data mining stipulates the use of one dependent variable, it is not possible to evaluate the relationship between independent variables.
- Intermediate relationships (either moderating or intervening) cannot be measured directly. As indicated, data mining techniques and methodologies allow for measuring only the Independent variable :: Dependent variable relationship; further modeling and analysis would be required to determine intermediate relationships.

7.2.2 Models from CHAID and entropy-based decision tree

Rules from entropy-based decision trees and CHAID are presented in Table 7.8 and Table 7.9. These rules are induced in the m -order ($m \geq 2$) form (that is, X and Z imply Y, where X and Z are any two conditions and Y is a response or dependent variable value). Each such rule includes variables from different constructs. Therefore, it is not possible to separate m order rules into first order rules. The rules from these two techniques are presented in the format presented below.

Table 7.8: Decision tree rules when no agreement reached

Information Gain	Normalized Mutual Information	CHAID
Met_expectations = <i>unknown</i> AND Offers_in_between = (0,1)	Offers_in_between ≤ 1 AND Work_w_opponent = <i>unknown</i>	Offers_5th_last_day = [0,1,2,3,4] AND Instructions_easy = <i>unknown</i> AND Offers_in_between = [0,1], <i>unknown</i>
Met_expectations = <i>unknown</i> AND Offers_in_between ≥ 2 AND Offers_2 nd _last_day ≥ 1	Offers_in_between ≤ 1 AND Work_w_opponent = <i>unknown</i>	Offers_5th_last_day = <i>unknown</i> AND Activity_deadline = [No, Yes]
Met_expectations = 6	Offers_in_between ≥ 2 AND Work_w_opponent = <i>unknown</i> AND Offers_2 nd _last_day ≥ 1	Offers_5th_last_day = <i>unknown</i> AND Activity_deadline = <i>unknown</i>
Met_expectations = 7		

As presented in Table 7.8, variable Offers_in_between is included in rules generated from information gain and normalized mutual information. This indicates that the variable may have a statistically significant relationship with observations where no agreement is reached.

Table 7.9: Decision tree rules when agreement reached in negotiation phase

Information Gain	Normalized Mutual Information	CHAID
Met_expectations = <i>unknown</i> AND Offers_in_between ≥ 2 AND Offers_2 nd _last_day = 0	Offers_in_between ≤ 1 AND Work_w_opponent = Yes	Offers_5th_last_day = [0,1,2,3,4] AND Instructions_easy = [1,2,3,4,5,6,7] AND Work_w_opponent = No
Met_expectations = 1	Offers_in_between ≥ 2 AND Work_w_opponent = <i>unknown</i> AND Offers_2 nd _last_day = 0	Offers_5th_last_day = [0,1,2,3,4] AND Instructions_easy = [1,2,3,4,5,6,7] AND Work_w_opponent = Yes
Met_expectations = 2	Offers_in_between ≥ 2 AND Work_w_opponent = No AND Opp_cooperate = 3	Offers_5th_last_day = [0,1,2,3,4] AND Instructions_easy = <i>unknown</i> AND Offers_in_between = (1,3)
Met_expectations = 3	Offers_in_between ≥ 2 AND Work_w_opponent = No AND Opp_cooperate = 4	Offers_5th_last_day = [0,1,2,3,4] AND Instructions_easy = <i>unknown</i> AND Offers_in_between = (3,17)
Met_expectations = 4	Offers_in_between ≥ 2 AND Work_w_opponent = No AND Opp_cooperate = 1	
Met_expectations = 5	Offers_in_between ≥ 2 AND Work_w_opponent = No AND Opp_cooperate = 2	
	Offers_in_between ≥ 2 AND Work_w_opponent = No AND Opp_cooperate = 5	
	Offers_in_between ≥ 2 AND Work_w_opponent = Yes	

As indicated in Table 7.9, variable Work_w_opponent is included in models constructed using normalized mutual information and CHAID. This indicates that this variable is significant in describing observations where an agreement was reached in the negotiation phase. Variable Offers_in_between is included in all three models presented in Table 7.9. This indicates that this variable is also significant in describing observations where agreement was reached in the negotiation phase.

7.3 Answering the secondary objectives

The secondary objective of this study is to:

"To compare differences and similarities present in models of web-based negotiations constructed with association rules, naive Bayes, entropy-based decision tree, CHAID techniques"

Six investigative questions are presented in Chapter 5. The results of each question are presented in this Section.

Investigative Question 1: Are estimates of classification error equal for models generated using the three classification techniques?

Models extracted using naive Bayes, entropy-based decision tree, and CHAID produce different estimates of classification error. Table 7.10 is used present these estimates.

Table 7.10: Classification error rate estimates for naive Bayes, entropy-based decision tree and CHAID

Model	Method (corresponding parameters)	Classification error and standard error
1	naive Bayes	4.11% $\pm$ 0.47%
2	Entropy-based decision tree (information gain; 3 levels; 50 minimum observations per node)	25.4% $\pm$ 1.04%
3	Entropy-based decision tree (normalized mutual information; 3 levels; 50 minimum observations per node)	29.0% $\pm$ 1.08%
4	CHAID (Pearson's chi-square; $\alpha_m = 5\%$; $\alpha_s = 5\%$)	6.38% $\pm$ 0.58
5	CHAID (Likelihood-ratio; $\alpha_m = 5\%$; $\alpha_s = 5\%$)	9.75% $\pm$ 0.71%

The estimates presented in Table 7.10 are those from models presented in Section 7.1. The estimates of classification error are different for all three classification methods. Entropy-based decision trees (constructed using information gain and normalized mutual information) are associated with the highest of five classification errors. Mitchell (Mitchell 1997) indicates the entropy based decision may construct models with higher estimates of classification error since the methods are biased to selecting variables with large number of values (such as continuous variables); as such this may increase the error of accurately classifying new observations into the "correct" group.

From observing the estimates of classification error, the model generated through naive Bayes seems to develop the most accurate model. To determine whether the accuracy is valid, further research should conduct tests for predictive modeling. Given higher number of observations in the INSPIRE dataset, hold-out sample testing

should be conducted (creating two mutually exclusive sub-sets: one for creating models, and the other for measuring classification accuracy). Model accuracy measured through this method, will enable the researcher to validate the estimates of classification error presented in Table 7.10.

Investigative Question 2: Are differences between models explainable by the methods that were used?

Models constructed with entropy-based decision trees and CHAID are composed of different sets of independent variables. Table 7.11 presents these sets of independent variables.

Table 7.11: Independent variables

1	Entropy based decision trees	information gain maximum 3 levels minimum 50 observations per node	Met_expectations Offers_in_between Offers_2 nd _last_day
2	Entropy based decision trees	normalized mutual information maximum 3 levels minimum 50 observations per node	Offers_in_between Work_w_opponent Offers_2 nd _last_day Opp_cooperate
3	CHAID	Pearson's chi-square α_m and α_s at level 5% and likelihood ratio α_m and α_s at level 5%	Offers_5 th _last_day Instructions_easy Work_w_opponent Offers_in_between Activity_deadline

The models presented are for specific parameters in CHAID and entropy-based decision trees.

According to (Mitchell 1997) entropy-based methods are biased to selecting

variables measured on a continuous or ratio scale. Models constructed from information gain and normalized mutual information include `Offers_in_between`, `Offers_2nd_last_day` — variables that are measured on a ratio scale — and with fifteen and six values, respectively. In contrary to Mitchell's findings, the two entropy-based techniques also include `Opp_cooperate` (measured on an ordinal scale) and `Work_w_opponent` (a binary variable). The presence of these three variables, and absence of other continuous variables may either indicate inconsistencies in the dataset or exceptions to Mitchell's findings in (Mitchell 1997).

Kass (1980) states that CHAID methods are robust to the presence of variables measured on continuous scales, and are likely to generate models comprised of either ordinal or nominal scaled variables; this robustness attributed to CHAID based on chi-square analysis, itself based on the analysis of expected and observed cell counts in contingency tables. In contrary to Kass's findings, variables `Offers_5th_last_day` and `Offers_in_between` (ordinal variables with seven and fifteen values, respectively) are included in the CHAID model. This may be a result of inconsistencies in the dataset or exceptions to Kass's findings. These inconsistencies may be observed due to the smaller dataset. To validate/invalidate the presence of such inconsistencies, a dataset with more observations is required for data mining.

The classification models include three independent variables that appear in at least one entropy-based model and one CHAID model. These are:

- `Offers_2nd_last_day`
- `Offers_in_between`
- `Work_w_opponent`

These variables may indicate stronger relationship with the outcome of negotiations than other variables in the models. Data mining with a larger sample size is needed to validate the significance of these three variables. In addition, experimental research, controlling for these variables, would also indicate the significance of these three variables in describing the outcome of web-based negotiations.

Investigative Question 3: Do entropy-based decision trees and CHAID generate similar sets of independent variables when parameters are modified?

Information gain was used to construct decision trees by modifying the levels of the tree and the minimum number of observations acceptable in a tree node. The results of the different model generations are presented in Table 7.12. In this table, the classification error and standard error estimates presented for each permutation were yielded after the method was used to generate average the results from five iterations.

Table 7.12: Models from information gain

Number of Levels	Minimum number of observations	Independent variables	Classification error estimate and standard error estimate
3	25	Met_expectations Offers_in_between Offers_2 nd _last_day	23.14% $\pm$ 1.04%
3	50	Met_expectations Offers_in_between Offers_2 nd _last_day	25.4% $\pm$ 1.04%
4	25	Met_expectations Offers_in_between Offers_2 nd _last_day	25.4% $\pm$ 1.04%
4	50	Met_expectations Offers_in_between Offers_2 nd _last_day	24.89 $\pm$ 1.03%

The modifications of parameters do not alter the set of independent variables included in the model (Table 7.12). The estimates of classification error range from 23.14% and 25.4%. For the given dataset, it may be concluded that the information gain method is robust to parametric modifications.

Table 7.13: Model from normalized mutual information

Number of Levels	Minimum number of observations	Independent variables	Classification error estimate and standard error estimate
3	25	Offers_in_between Offers_2 nd _last_day Work_w_opponent Met_expectations	24.66% $\pm$ 1.03%
3	50	Offers_in_between Offers_2 nd _last_day Work_w_opponent Opp_cooperate	29.0% $\pm$ 1.08%
4	25	Offers_in_between Offers_2 nd _last_day Work_w_opponent Met_expectations	24.71% $\pm$ 1.03%
4	50	Offers_in_between Offers_2 nd _last_day Work_w_opponent Opp_cooperate	28.94% $\pm$ 1.08%

As presented in Table 7.13, Met_expectations is included in models where the parameter “Minimum number of observations (in each tree node)” is set to twenty-five. This variable is replaced by Opp_cooperate in models where the parameter “Minimum number of observations (in each tree node)” is set to fifty. Both variables are measured on a scale from {1,...7}; thus it does not lead that normalized mutual information is biased towards a variable with larger number of values (Mitchell 1997). This difference in models may be explained by the distribution of observations for each variable.

Table 7.14: Pearson's chi-square (X^2) models

		Independent variables	Classification error estimate and standard error estimate
5%	5%	Offers_in_between Instructions_easy Work_w_opponent Activity_deadline Offers_5 th _last_day	6.38% $\pm$ 0.58
5%	10%	Offers_in_between Instructions_easy Work_w_opponent Activity_deadline Offers_5 th _last_day	6.61% $\pm$ 0.59
5%	20%	Offers_in_between Instructions_easy Work_w_opponent Activity_deadline Offers_5 th _last_day	9.92% $\pm$ 0.71%
10%	5%	Offers_in_between Instructions_easy Work_w_opponent Activity_deadline Offers_5 th _last_day	5.07% $\pm$ 0.52%
10%	10%	Offers_in_between Instructions_easy Work_w_opponent Activity_deadline Offers_5 th _last_day	7.4% $\pm$ 0.63%
10%	20%	Offers_in_between Instructions_easy Work_w_opponent Activity_deadline Offers_5 th _last_day	11.8% $\pm$ 0.77%
20%	5%	Case_understand Graph_inform Performance Agreement_satisfaction	8.27% $\pm$ 0.66%
20%	10%	Case_understand Graph_inform Performance Agreement_satisfaction	12.3% $\pm$ 0.78%
20%	20%	Case_understand Graph_inform Performance Agreement_satisfaction	4.62% $\pm$ 0.50%

α_m denotes the minimum level of significance for merging categories
 α_m denotes the minimum level of significance for selecting descriptors

Table 7.14 presents models for Pearson's chi-square (X^2). As indicated, the model is robust to modifications to $\underline{\alpha}_m$ and $\underline{\alpha}_s$. The set of independent variables is different, though, for models with $\underline{\alpha}_m = 20\%$ (irrespective of the value for $\underline{\alpha}_s$). Since $\underline{\alpha}_m$ is used to determine significant categories in the independent variables, lower value(s) of this parameter may skew significance towards categories with spurious relationships with the dependent variable. Since $\underline{\alpha}_m$ is used in the CHAID process prior to $\underline{\alpha}_s$, the results presented in Table 7.14 may indicate that $\underline{\alpha}_m$ is a more important determinant (compared to $\underline{\alpha}_s$) of independent variables in a X^2 model.

Table 7.15 presents models for likelihood ratio (G^2). Similar to the models derived using X^2 , the set of independent variables is different for models with $\underline{\alpha}_m = 20\%$ (irrespective of the value for $\underline{\alpha}_s$). A similar conclusion, that $\underline{\alpha}_m$ may be a more important determinant (compared to $\underline{\alpha}_s$) of independent variables, may also be derived for G^2 models.

Table 7.15: Likelihood ratio (G^2) models

		Model variables	Classification error estimate and standard error estimate
5%	5%	Offers_in_between Instructions_easy Work_w_opponent Activity_deadline Offers_5 th _last_day	9.75% $\pm$ 0.71%
5%	10%	Offers_in_between Instructions_easy Work_w_opponent Activity_deadline Offers_5 th _last_day	13.5% $\pm$ 0.81%
5%	20%	Offers_in_between Instructions_easy Work_w_opponent Activity_deadline Offers_5 th _last_day	4.73% $\pm$ 0.51%
10%	5%	Offers_in_between Instructions_easy Work_w_opponent Activity_deadline Offers_5 th _last_day	13.91 $\pm$ 0.83%
10%	10%	Offers_in_between Instructions_easy Work_w_opponent Activity_deadline Offers_5 th _last_day	18.13% $\pm$ 0.92%
10%	20%	Offers_in_between Instructions_easy Work_w_opponent Activity_deadline Offers_5 th _last_day	9.58% $\pm$ 0.72%
20%	5%	Case_understand Graph_inform Performance Agreement_satisfaction	10.6% $\pm$ 0.74%
20%	10%	Case_understand Graph_inform Performance Agreement_satisfaction	14.48% $\pm$ 0.84%
20%	20%	Case_understand Graph_inform Performance Agreement_satisfaction	12.09% $\pm$ 0.78%

α_m denotes the minimum level of significance for merging categories

α_s denotes the minimum level of significance for selecting descriptors

Investigative Question 4: Are the estimates of classification accuracy similar for models constructed with entropy-based method and for models constructed with likelihood-ratio (G^2)?

The entropy method and likelihood ratio (G^2) use log derivatives as statistics to measure significance of independent variables in describing the dependent variable (Kass 1980; Mitchell 1997). The log measure is used in entropy calculation to evaluate the overall dispersion in a given dataset; the complementary information gain measure is used to evaluate the corresponding reduction in entropy when an independent variable is used to segment the dependent variable (Mitchell 1997). The likelihood ratio uses the log measure to evaluate the difference between observed cell count and expected cell count (Agresti 1996).

Table 7.16 presents the corresponding classification errors of models constructed through entropy-based decision trees (using both information gain and normalized mutual information) and CHAID (using likelihood ratio). As indicated, models constructed with entropy-based decision trees and the log-likelihood ratio include different estimates of classification error. Tables 7.12, 7.13, and 7.15 also indicate that the set of independent variables included in corresponding models are not equal.

Table 7.16: Classification error estimates for entropy methods and G^2

Model	Method (corresponding criterion)	Classification error and standard error (%)
1	Entropy-based decision tree (information gain; 3 levels; 50 minimum observations per node)	25.4% $\pm$ 1.04%
2	Entropy-based decision tree (normalized mutual information; 3 levels; 50 minimum observations per node)	29.0% $\pm$ 1.08%
3	CHAID (likelihood ratio; $\alpha_m = 5\%$; $\alpha_s = 5\%$)	9.75% $\pm$ 0.71%

Investigative Question 5: Using CHAID, do models constructed with Pearson's chi-square (X^2) differ from models constructed with likelihood-ratio (G^2)?

As presented in Section 7.1, the models constructed using X^2 and G^2 include the same set of independent variables. As presented in Table 7.14 and Table 7.15, corresponding estimates of classification error do differ; these differences may or not be significant. In order to determine the robustness of estimates of error, classification error rates were estimated over five iterations (using 3-fold cross-validation). Table 7.17 and Table 7.18 indicate the estimates of classification error (and standard error) for X^2 and G^2 , respectively, with α_m and α_s at level 5%.

Table 7.17: Estimated error rates and standard error using Pearson's chi-square(X^2) with α_m and α_s at level 5%

Run Number	Estimated classification error rate	Standard error on classification error rate
1.	6.38%	0.58%
2.	5.92%	0.56%
3.	7.01%	0.61%
4.	6.27%	0.58%
5.	5.81%	0.56%
Average	6.28%	0.58%
Median	6.27%	0.58%

Table 7.17 indicates the variance of classification error estimates between the five runs. In these five runs, the estimates of classification error differed between each run. Thus, the average indicator may be used as more reflective of the "true" error estimate for that model.

Table 7.18: Estimated error rates and Standard error using CHAID (Likelihood-ratio) with α_m and α_s at level 5%

Run Number	Estimated classification error rate	Standard error on classification error rate
1.	4.61%	0.50%
2.	4.61%	0.50%
3.	5.64%	0.55%
4.	12.59%	0.79%
5.	4.61%	0.50%
Average	6.41%	0.57%
Median	4.61%	0.50%

Table 7.18 presents the classification error rates and corresponding standard error on each estimate, using the likelihood ratio method (G^2). Comparing the average and median values for both X^2 indicates small difference between estimates.

Investigative Question 6: Does the use of association rules analysis provide models that are similar to the model extracted using naive Bayes?

Eight association and Bayesian rules were extracted that were identical. Table 7.19 presents these similar rules. It is important to note that all eight rules describe observations when the agreement was reached in the negotiation phase.

Table 7.19: Similar association and Bayesian rules

		Association	Dataset	Association	Naive Bayes
1	Knew_opponent_id = No	42.00	40.20	54.08	54.10
2	NSS_before = No	38.92	38.90	53.96	54.00
3	Case_understand = 2	23.00	23.00	56.68	56.70
4	Graph_used = Yes	23.23	23.30	68.98	69.00
5	Disclose_country = No	24.65	24.70	65.36	65.40
6	Disclose_id = No	25.62	25.60	65.64	65.60
7	Surprised = No	22.37	22.40	69.88	69.90
8	Work_w_opponent = Yes	22.83	22.90	76.92	76.90

In Table 7.19, prevalence is compared to percentage of observations in dataset, and predictability is compared to likelihood of condition. These results from Table 7.19 strongly indicate that association rules analysis and naive Bayes technique yield similar models, despite the inherent differences between the two techniques.

The similarities may arise from the assumptions of independence between independent variables and assumption of constant interaction between the dependent variable and the independent variable. These two assumptions are explicit parts of

naive Bayes; they are also implicit in association rules analysis. In association rules analysis, only 2-way associations are considered; thus, m -way ($m \geq 2$) associations are not considered. Thus, the use of association rules analysis implies the following assumption:

Assumption 1: Association between the dependent variable and each independent variable is constant of other independent variables

The above assumption is also implied in the use of naive Bayes analysis (Section 4.3). The use of the naive Bayes technique implies the use of a second assumption (Section 4.3):

Assumption 2: There is no association between independent variables

Assumption 2 is not implied in the use of association rules analysis. Consider two variables X_1 and X_2 (where $X_1 \in X$, $X_2 \in X$, X is set of variables in dataset S). Since association rules are extracted for all variables in S , and the dependent variable is not identified *a priori*, the task of the technique is to find all associations that have prevalence and predictability greater than or equal to $\underline{p}$ and $\underline{c}$, respectively. Thus, associations between X_1 and X_2 (either $X_1 \Rightarrow X_2$ or $X_1 \Leftarrow X_2$) can be extracted from S .

For future analysis, a method should be developed that enables the “chaining” of association rules: that is, enables the presentation of rules of the form $X_m \Rightarrow Y$ (where $X_m \in X$, X is set of independent variables in S , Y is dependent variable), and presents rules in the form $X_m \Leftarrow X_a$, $X_a \Leftarrow X_b$ (where $X_m, X_a, X_b \in X$, $X_m \neq X_a \neq X_b$).

8. Data mining to support statistics: Case of loglinear modeling

The results presented in Chapter 7 conclude the objectives of this thesis. That chapter presented rules that could be used for further research. In addition to providing to simple rules on the behaviour of web-based negotiators, the author believes that results from data mining may also be used to support statistical analysis.

The objective of this section is to present data mining as a possible preliminary and exploratory process to statistical analysis. Specifically, this section is used to demonstrate an example of how data mining can be used to support the task of loglinear modeling.

8.1 Introduction to Loglinear modeling

Loglinear modeling is a statistical method that tests for independence and interaction between variables. It is, typically, used to model interactions between three or more variables.

The simplest test for independence is conducted with Pearson's X^2 or likelihood ratio G^2 to measure independence between two variables.

Table 8.1: Sample two-way table

	1	$n_{11} (\pi_{11})$	$n_{12} (\pi_{12})$
X	2	$n_{21} (\pi_{21})$	$n_{22} (\pi_{22})$
Total		$n_{+1} (\pi_{+1})$	$n_{+2} (\pi_{+2})$

n_{ij} — Number of observations (counts) corresponding to i th row and j th column

n_{+j} and n_{i+} — Summation of counts across rows and columns, respectively

π_{ij} — (Population) Probability of observing the condition specified by the intersection of i and j .

π_{+j} — Summation of (Population) probabilities. Equals 1.

Variables X and Y are independent (that is, there is no association) if conditional distributions of Y are the same at each level of X. That is:

$$\left\{ \frac{\pi_{11}}{\pi_{1+}}, \frac{\pi_{12}}{\pi_{1+}} \right\} = \left\{ \frac{\pi_{21}}{\pi_{2+}}, \frac{\pi_{22}}{\pi_{2+}} \right\} \quad (1)$$

The formula presented in (1) also presents the conditional distribution of Y at X=1 (LHS) and the conditional distribution of Y at X=2 (RHS).

Another measure that is used to measure association is called the odds ratio. For X and Y presented in Table 8.1, the odds ratio is calculated as:

$$\text{Odds ratio} = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}} \quad (3)$$

A “no association” model between two variables is plausible when the odds ratio yields a value of one.

The loglinear model under null hypothesis (no association/independence) is given as:

$$\ln(\mu_{ij}) = \lambda + \lambda_i^x + \lambda_j^y \quad (3)$$

where λ_i represent Maximum Likelihood parameter estimates.

Table 8.2: Sample 2-way table

		1	2
X	1	435	147
	2	375	134
Total		810	281
			1091

Table 8.2 cross-classifies a sample dataset¹. Given this sample dataset, the loglinear model is developed as follows (Thomas 1999).

Step 1: Calculate odds ratio

The odds ratio for the data is calculated as:

$$\begin{aligned} \text{Odds ratio} &= \frac{435 \times 134}{375 \times 147} \\ &= 1.057 \end{aligned}$$

Step 2: Calculate Maximum likelihood parameters

The loglinear model under null hypothesis is given in (3). This model can be further expanded as follows:

$$\ln(\mu_{ij}) = \lambda + \lambda_i^x + \lambda_j^y + \lambda_{i1}^{xy} + \lambda_{i2}^{xy} + \lambda_{11}^{xy} + \lambda_{12}^{xy} + \lambda_{21}^{xy} + \lambda_{22}^{xy} \quad (4)$$

All parameters with the exception of three — in this example $\lambda, \lambda_1^x, \lambda_1^y$ — are set to zero. Solving for the loglinear model yields the following values for the

¹ Data is derived from Belief in Afterlife dataset in Agresti (1996)

parameters:

$$\lambda = 4.898, \lambda_1^X = 0.92, \lambda_1^Y = 1.029, \lambda_{11}^{XY} = 0.56$$

Step 3: Determine Maximum likelihood estimates of cell frequencies (fitted values)

Table 8.3: Cell frequencies (fitted values)

		Y	
		1	2
X	1	432.1	149.9
	2	377.9	131.1
Total		810	281

Cell estimates for Table 8.2 are presented in Table 8.3 are calculated as follows:

$$\hat{\mu}_{ij} = \frac{n_{i+} n_{+j}}{n_{++}} \quad (5)$$

where $\hat{\mu}_{ij}$ is the expected count at cell i,j under null hypothesis.

The odds ratio calculated for the fitted values yields a value of one. Therefore, there is a close fit to the observed values. Thus, a “no association” model is plausible. The shorthand notation for a “no association” model is given by (A,B).

A model that presents no fit indicates that there may be an association between X and Y. In this situation, an association term is added to the model. That is,

$$\ln(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad (6)$$

The model presented in (6) is referred to as the “saturated model”.

Addendum: Case of three-way interaction

The previous steps presented the development of the loglinear model given two-way

interactions. The loglinear model (under no association) for three-way interaction is given as follows:

$$\ln(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \quad (7)$$

This model — denoted as the (AB, BC, AC) — specifies that the odds ratios for A-B are equal at each level of C, the odds ratio for B-C are equal at each level of A, and the odds ratio for A-C are equal at each level of B.

The saturated model (association) for three-way interaction includes a third term, as follows:

$$\ln(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \quad (8)$$

Therefore, when considering interaction between three categorical variables A, B, C, the A-B, A-C, B-C interactions have to be calculated. As the number of variables to analyze increase, the number of interactions to calculate increases. For example, modeling interactions between five variables involves calculating thirty interactions. Therefore, the task of loglinear modeling is made complex as the number of variables and categories increase.

The use of data mining techniques is, typically, not constrained by the number of variables considered. Such techniques can be used to perform summary-type analysis to identify a list of “significant” independent variables (it should be noted that independent variables deemed important by a data mining technique may or may not have a statistically significant relationship with the dependent variable). These significant variables may then be used for loglinear modeling

Section 8.2 presents the results of loglinear analysis when the results from CHAID and entropy-based methods are used to determine significant independent variables

for modeling.

8.2 Results from loglinear analysis

Significant independent variables extracted using entropy-based and CHAID techniques are considered for the task of loglinear modeling. These are as follows:

- Offers_2nd_last_day
- Offers_5th_last_day
- Offers_in_between
- Met_expectations
- Work_w_opponent
- Opp_cooperate
- Instructions_easy
- Activity_deadline

Variables with five or more values are categorized into four categories (using the quartile distributions of the original variable); all variables, with the exception of Work_w_opponent and Activity_deadline (binary variables), were categorized in this manner. For three variables, categorization yielded less than four groups since the distribution of observations in the original variable is skewed to a few values. (Refer to Appendix C for further information)

Loglinear interactions between the dependent variable and seventeen distinct 4-way and 6-way interactions were observed, such that each independent variable was included at least in one such m -way ($m = 4$ OR $m = 6$) interaction. For example, the

dependent variable interacts with the following 6-way interaction:

- Outcome of Negotiations * Instructions_easy * Met_expectations * Offers_2nd_last_day * Offers_5th_last_day * Offers_in_between

Each m -way interaction ($m \geq 2$) implies the existence of lower order interactions (Gilbert 1981; Agresti 1996). Such high-order interactions are difficult to interpret. An approach to reducing the number of high-order interactions is to divide the dataset according to a variable that appears in an m -way interaction. For this example, the dataset is divided into four subsets using the two binary variables Activity_deadline and Work_w_opponent, to produce four data sub-sets with the following characteristics:

1. Activity_deadline = No & Work_w_opponent = No
2. Activity_deadline = Yes & Work_w_opponent = No
3. Activity_deadline = No & Work_w_opponent = Yes
4. Activity_deadline = Yes & Work_w_opponent = Yes

Loglinear modeling was conducted for each of the four subsets, and resulting interactions between the dependent variable and the independent variable(s) were recorded. Table 8.4 presents the results of these modeling exercises.

Table 8.4: Loglinear modeling

		Outcome of negotiations	
Work_w_opponent	No	Outcome of negotiations * Offers_2 nd _last_day	Outcome of negotiations * Offers_5 th _last_day
	Yes	Outcome of negotiations * Offers_2 nd _last_day	No first order interaction detected

Results from loglinear analysis indicate that three variables (Instructions_easy, Offers_5th_last_day, and Offers_2nd_last_day) have direct interaction with the dependent variable; direct interactions cannot be yielded through the use of CHAID and entropy-based decision trees.

The division of the dataset into four subsets simplified the process of analysis. This sub-division of the dataset is distinct from the process followed in decision trees (for example, CHAID and entropy-based methods), since data subsets are mutually exclusive (data subsets maintain a hierarchical relationship in decision tree methods).

9. Conclusions

This thesis presented the case of constructing data mining models of the outcome of web-based negotiations. Web-based negotiations, through the development of the INSPIRE NSS, were presented in Chapter 2. This chapter was also used to present the need for constructing models of negotiated outcomes. Chapter 3 was used to present the emerging research of data mining and knowledge discovery. Five data mining techniques were presented and explained in Chapter 4. Chapter 5 presented the research objective of this thesis, and the research methodology was outlined in Chapter 6.

The use of data mining has resulted in the development of models. These models may act as exploratory models that can be used in further research of web-based negotiations. Tukey (1977) presented the use of exploration as a preliminary step before modeling and testing for relationships in a population. The models constructed with data mining may now be used in future experimental research, where relationships can be controlled for and validated on a larger population.

9.1 Key observations from this study

The study indicates the following key observations:

- 1) The outcome of web-based negotiations may be described by variables that may or not be obvious prior to modeling. For example, sending more offers in between (from beginning of negotiations to the beginning of the final week) may describe a higher likelihood to reach an agreement. Additionally, exchanging more offers may also contribute to reaching an agreement.
- 2) The eight variables considered for loglinear modeling (Offers_2nd_last_day, Offers_5th_last_day, Offers_in_between, Met_expectations, Work_w_opponent, Opp_cooperate, Instructions_easy, and Activity_deadline) are observed in multiple classification models. These variables should be considered important in future experiments of to predict the outcome of web-based negotiations.
- 3) The rules and models constructed in this thesis can be used in one of three ways:
 - As rules to be used in specific circumstances when the conditions of the rule (that is, the precedents) are observed and the likelihood of observing the status of agreement is required. In these specific cases, these rules may be used in similar way as rules are used in the area of decision sciences or decision-making
 - As models that explain the status of agreements in web-based

negotiations. These models may be used to provide the reader an overview of descriptors of agreement status.

- The set of independent variables presented may be considered as possible predictors of agreement status (that is, outcome of web-based negotiations) in further experimental studies. This would reduce activities typically required in experimental studies to determine variables of interest.

- 4) The evidence from this study should be used as results of the given dataset. Modeling was not conducted to measure the statistical generalizability of these models, and thus, these models cannot be used to comment on the behaviour of the population.
- 5) Different data mining techniques yielded models with similar estimates of classification error. Due to this similarity, it was not possible to determine which method yielded more accurate models. Therefore, further research may be needed to determine which technique yields more accurate models of the outcome of web-based negotiations.
- 6) The CHAID technique yielded similar models (included the same independent variables) when $\underline{\alpha}_m$ and $\underline{\alpha}_s$ were set to 5% or 10%; thus indicating the technique's possible robustness. Using CHAID at $\underline{\alpha}_s = 20\%$, though, significantly altered the model. This result may indicate that in CHAID (using either X^2 or G^2) the $\underline{\alpha}_m$ estimate is more relevant than the $\underline{\alpha}_s$ parameter. The result may also indicate that the technique's robustness is not observable

at $\alpha = 20\%$.

- 7) The use of association rules analysis and naive Bayes produced eight identical rules; such that values of prevalence and predictability were no different than at most 1%. A possible explanation of this similarity is that the assumption of a constant relationship between independent variables and dependent variable (Assumption 1 in Section 4.4) is implied in both techniques.
- 8) It is also not possible to directly account for high correlation between independent variables in data mining. Such measures have to be conducted *a priori* to data mining modeling through the use of univariate methods.
- 9) According to Fayyad (1996), data mining is used for exploratory modeling. Therefore, key assumptions are implied in the use of data mining techniques. First, in the use of data mining techniques (with the exception of association rules analysis) the assumption of no interaction between independent variables is implied. Second, two sets of variables are assumed in S (where S represents the considered dataset) — independent and dependent variable. Modeling for intermediate variables is constrained. Third, the use of data mining techniques does not allow for directly measuring, and therefore controlling, for interactions and correlations between independent variables. Therefore, multi-collinearity cannot be controlled for in the use of data mining techniques.

9.2 Suggestions for future research

The following suggestions are made for future research:

- 1) Research to compare association rules analysis and naive Bayes method. Experiments should be conducted to determine under what terms these two methods would yield identical rules.
- 2) Develop framework to “chain” association rules. Association rules present the use of identifying interactions between dependent and independent variables, as well as between independent variables. A “chain” method should be constructed that allows for constructing a link of interactions between independent variables. This may help in the analysis of cause-effect relationships.
- 3) Future research should use other data mining techniques in modeling web-based negotiations. The nature of the selected dependent variable for this study (Outcome of negotiations) and the nature of most independent variables indicated the use of methods that were robust to the presence of ordinal and nominal variables. Future research should consider dependent variables measured on ratio or continuous scales and should consider the use of methods such as CART (Classification and Regression Trees), optimal decision trees, and ANN (Artificial Neural Networks).
- 4) Empirical models may be constructed to describe:
 - a) Bargainers that reach an improve agreement through the post-settlement mechanism, and compare results with bargainers that do not improve agreements through the post-settlement mechanism.

- b) Identify factors that determine whether a user is active late into the negotiation. Modeling from this study indicated several observations where negotiators were able to postpone negotiating until the very last days of the process.
 - c) Compare results from web-based negotiations with results from face-to-face negotiations. A design that ensures negotiators in both groups use the same negotiations instruments would allow for results to highlight differences between face-to-face and web-based negotiations.
- 5) This study presented an example in Chapter 8 to present one potential use of data mining results to support loglinear modeling. Future research should study other areas of statistics that can be supported through data mining.

10. References

- Adler, N. J., R. Brahm, et al. (1992). "Strategy Implementation: A comparison of face-to-face negotiations in the People's Republic of China and the United States" *Strategic Management Journal* 13: 449-466.
- Agrawal, R., T. Imielinski, et al. (1993). "Mining association rules between sets of items in large databases". *Proceedings of the 1993 SIGMOD Conference*, Washington DC.
- Agrawal, R., M. Mehta, et al. (1997). "The Quest data mining system". San Jose, IBM Almaden Research Center, <http://www.ibm.com>.
- Agresti, A. (1990). *Categorical data analysis*, John Wiley & Sons.
- Agresti, A. (1996). *An introduction to categorical data analysis*, John Wiley & Sons, Inc.
- Alemi, F. and P. Fos (1990). "A demonstration of methods for studying negotiations between physicians and health care managers" *Decision Sciences* 21(3): 633-641.
- Balakrishnan, P. and J. Eliashberg (1995). "An analytical process model of two-party negotiations" *Management Science* 41(2): 226-243.
- Bell, D. E., H. Raiffa, et al. (1988). *Decision making. Descriptive, normative and prescriptive interactions*. Cambridge, MA, Cambridge Univ. Press.
- Berry, M. and G. Linoff (1997). *Data mining techniques: For marketing, sales and customer support*, Wiley Computer Publishing.
- Brin, S., R. Motwani, et al. (1997). "Beyond market baskets: Generalizing association rules to correlations". *ACM SIGMOD Conference on Management of data*.
- Bui, T. (1994). "Designing multiple criteria negotiation support systems:

- Frameworks, issues and implementation". *Tenth International Conference: Expand and Enrich the Domains of Thinking and Applications*, Springer-Verlag.
- Carroll, M. L. (1996). *Cyberstrategies: How to build an Internet-based information system*, Van Nostrand Reinhold Publishing.
- Chatterjee, K. and G. L. Lilien (1984). "Efficiency of alternative bargaining procedures: An experimental study" *Journal of Conflict Resolution* 28(2): 270-295.
- Chen, M.-S., J. Han, et al. (1996). "Data Mining: An overview from a database perspective" *IEEE Transactions on Knowledge and Data Engineering* 8(6): 866-883.
- Chen, S. K., P. Mangiameli, et al. (1995). "The comparative ability of self-organizing neural networks to define cluster structure" *Omega International Journal of Management Science* 23(3): 271-279.
- Cognos (1999). "Scenario - Data mining at your fingertips". Ottawa, Canada, Cognos.
- Drake, L. E. (1995). "Negotiation Styles in Intercultural Communication" *International Journal of Conflict Management* 6(1): 72 - 90.
- Eisenbeis, R. A. and R. B. Avery (1972). *Discriminant analysis and classification procedures: Theory and applications*. Lexington, Massachusetts, Lexington Books.
- Emory, W. C. and D. R. Cooper (1991). *Business Research Methods*. Boston, Irwin Publishing.
- Ezawa, K. J. and S. W. Norton (1996). "Constructing Bayesian networks to predict uncollectible telecommunications accounts" *IEEE Expert* 11(5): 45-52.
- Fayyad, U. M. (1996). "Data Mining and Knowledge Discovery: Making Sense Out of Data" *IEEE Expert* 11(5): 20-25.

- Fayyad, U. M., G. Piatetsky-Shapiro, et al. (1996). "From data mining to knowledge discovery: An overview". *Advances in knowledge discovery and data mining*. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurisamy. Menlo Park, MIT Press.
- Gilbert, G. N. (1981). *Modelling society: An introduction to loglinear analysis for social researchers*. London, George Allen & Unwin.
- Glaser, B. and A. L. Strauss (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, Aldine.
- Glymour, C., D. Madigan, et al. (1997). "Statistical Themes and Lessons for Data Mining" *Data Mining and Knowledge Discovery* 1(1): 11-28.
- Gordon, A. H. (1981). *Classification*, Weissley Publications.
- Graham, J. L., A. T. Mintu, et al. (1994). "Explorations of Negotiation Behaviors in Ten Foreign Cultures Using a Model Developed in the United States" *Management Science* 40(1): 72-95.
- Harris-Jones, C. (1997). "Data Mining Tools Survey". London, AMS Center for Advanced Technologies.
- Hendon, D. W., R. A. Hendon, et al. (1996). *Cross-cultural negotiations*, Quorum Books.
- James, K. (1985). *Classification Algorithms*, Thompson Publishing.
- John, G. H. (1996). "Robust linear discriminant trees". *Learning from data: Artificial intelligence and statistics V*. D. Fisher and H. J. Lenz. New York, Springer: 375-386.
- John, G. H. (1997). "Enhancements to the Data Mining Process". *Computer Science*

- Department, School of Engineering, Stanford University: 194pp.
- Kass, G. V. (1980). "An exploratory technique for investigating large quantities of categorical data" *Applied Statistics* 29(2): 119-127.
- Kersten, G., S. Koszegi, et al. (1999). "Cultural Influences in Web-based Business Negotiations". Ottawa, InterNeg: 15.
- Kersten, G. E. and S. J. Noronha (1999). "Negotiation Via the World Wide Web: A Cross-Cultural Study of Decision Making" *Group Decision and Negotiation* 8: 251-279.
- Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". *International Joint Conference on Artificial Intelligence (IJCAI)*.
- Kohavi, R., B. Becker, et al. (1998). "Improving simple Bayes". Mountain View, Data Mining and Visualization group, Silicon Graphics, Inc.
- Lax, D. and J. K. Sebenius (1986). *The manager as negotiator: Bargaining for cooperation and competitive gain*. New York, The Free Press.
- Limb, P. R. and G. J. Meggs (1994). "Data Mining - tools and techniques" *BT Technology* 12(4): 32-41.
- Mahadevan, A., K. Ponnudurai, et al. (1999). "Knowledge discovery in databases and decision support". *Decision Support for Sustainable Development*. G. Kersten, A. Yeh and Z. Mikolaguk, Kluwer Academic.
- Matwin, S. (1998). "Data Mining and Concept Learning", School of Information Technology and Engineering, University of Ottawa.
- Megiddo, N. and R. Srikant (1998). "Discovering predictive association rules". San

- Jose, CA, IBM Almaden Research Center.
- Mitchell, T. M. (1997). *Machine Learning*. New York, Mc-Graw-Hill.
- Noronha, S. and S. Szpakowicz (1996). "Negoplan: A system for logic-based decision modelling". *Lecture Notes in Computer Science*.
- Pazzani, M. J. (1996). "Searching for dependencies in Bayesian classifiers". *Learning from data: Artificial intelligence and statistics V*. D. Fisher and H. J. Lenz. New York, Springer: 239-248.
- Ponnuthurai, K. (1998). "Data mining and organizations". School of Business. Ottawa, Carleton University: 96.
- Prasnikar, V. and A. E. Roth (1992). "Consideration of fairness and strategy: Experimental data from sequential games" *Quarterly Journal of Economics* (August): 865-888.
- Pregibon, D. (1998). "Data Mining: What's at stake for the field of statistics". *Talks in statistics*, Statistics Canada, Ottawa, Canada.
- Quinlan, J. R. (1993). *C4.5: Programs for machine learning*, Morgan Kaufmann Publishers.
- Rangaswamy, A. and G. R. Shell (1994). "Using Computers to Realize Joint Gains in Negotiations: Toward an Electronic Bargaining Table". *Computer-Assisted Negotiations and Mediation Symposium*, Harvard Law School, Cambridge, MA.
- Romesburg, H. C. (1990). *Cluster Analysis for Researchers*. Malabar, Robert E. Krieger Publishing Company.

- SAS (1999). "Data mining", SAS Institute Inc. 1999.
- SGI (1998). "Online MineSet 2.5 User's Guide", Silicon Graphics Inc., <http://www.sgi.com/Products/software/MineSet/>.
- SPSS (1998). "SPSS in Data Mining". Chicago, SPSS Inc., <http://www.spss.com>.
- Stavanhagen, R. (1996). *Ethnic conflicts and the Nation-State*. London, Macmillan.
- Strauss, A. and J. Corbin (1994). "Grounded Theory Methodolgy". *Handbook of qualitative research*. N. K. Denzin and Y. S. Lincoln, Sage.
- Sun, L., F. Xu, et al. (1994). "Cluster analysis by the k-means algorithm and simulated annealing" *Chemometrics and Intelligent Laboratory Systems* 25: 51-60.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, Addison-Wesley.
- Thomas, R. (1999). *Class Notes for 42.693*
- Weiss, S. M. and N. Indurkha (1998). *Predictive data mining: A practical guide*. San Francisco, Morgan Kaufman Publishers.
- Weiss, S. M. and C. L. Kulikowski (1991). *Computer Systems that Learn: Classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. San Mateo, Morgan Kaufman Publishers.
- Wirth, R., C. Shearer, et al. (1997). "Towards process-oriented tool support for Knowledge Discovery in databases". *Principles of Data Mining and Knowledge Discovery: First European Symposium*. J. Komorowski and J. Zytkow, Springer-Verlag.

Appendix A: Data Dictionary

Note: All missing values in the data set are represented as 'NA'.

Variables captured in analysis phase

Table A.1: Representation of variables captured in pre-questionnaire

Year_of_Birth	User's year of birth.	Discrete	Higher than 1900
Gender	User's gender.	Binary	{M, F}
Opp_gender	Counterpart's gender.	Binary	{M, F}
Occupation	User's occupation	Nominal	{Student, Professional, Other}
Cborn	User's country of birth	Nominal	
Opp_cbon	Counterpart's country of birth	Nominal	
Creside	User's country of residence	Nominal	
Opp_cres	Counterpart's country of residence	Nominal	
First_language	User's mother tongue	Nominal	
Knew_opponent_country	User's knowledge of counterpart's country prior to commencement of bargaining	Binary	{No, Yes}
Knew_opponent_id	User's knowledge of counterpart's identity prior to commencement of bargaining	Binary	{No, Yes}
Internet_access	User's level of Internet access	Ordinal	{0,6}
Imore	If user's level of Internet access is ≥ 4 , would they expect to increase use after INSPIRE	Binary	{No, Yes}
Negotiation_experience	Prior negotiation experience	Ordinal	{1,5}
NSS_before	Prior use of a negotiation support system (NSS)	Binary	{No, Yes}
Case_understand	Ease in understanding INSPIRE case	Ordinal	{1,5}
Expected_score	Utility score on user's expected package	Continuous	{0,100}
Reservation_score	Utility score on user's reservation package	Continuous	{0,100}
Expected_friendliness	User's expectations on friendliness in negotiations	Ordinal	{1,5}

ess	friendliness in negotiations		
Weighing_issues	User's ease in weighing negotiation issues	Ordinal	{1,5}
Weighing_options	User's ease in weighing negotiation options	Ordinal	{1,5}
Optimal_score	Measures the user's optimal score	Continuous	{0,100}
Negotiation_case	Identifies the user as buyer or seller	Nominal	{Itex, Cypress}

Variables captured in negotiation phase

Table A.2: Representation of variables captured by INSPIRE history recording mechanism

Offers	Number of offers <i>conveyed</i> by user during negotiations	Discrete	{0, ∞ }
Opp_offers	Number of offers <i>received</i> by user during negotiations	Discrete	{0, ∞ }
Offers_w_messages	Number of offers <i>conveyed</i> by user during negotiations that included written messages	Discrete	{0, ∞ }
Opp_Offers_w_messages	Number of offers <i>received</i> by user during negotiations from opponent	Discrete	{0, ∞ }
Messages	Number of written message <i>sent</i> by user during negotiations	Discrete	{0, ∞ }
Opp_messages	Number of written messages <i>received</i> by user from opponent	Discrete	{0, ∞ }
Outcome_negotiations	Outcome of negotiations	Nominal	1) No agreement reached 2) Agreement reached in negotiation phase 3) Agreement reached in the post-settlement phase
Score	Utility rating (based on user defined utility function) of agreed upon package	Continuous	{0, 100}
Optimal	Was final agreement Pareto-optimal.	Binary	{No, Yes}

Postsettlement_suggested	Did INSPIRE suggest post-settlement to user	Binary	{No, Yes}
Postsettlement_used	Did user utilize post-settlement packages	Binary	{No, Yes}
Opp_Postsettlement_used	Did counterpart utilize post-settlement packages	Binary	{No, Yes}
Postsettlement_offers	Number of offers <i>sent</i> in post-settlement stage by user	Discrete	{0, ∞ }
Opp_Postsettlement_offers	Number of offers <i>received</i> by user from opponent in post-settlement	Discrete	{0, ∞ }
Negotiation_length	Length of user's negotiations	Continuous	{0, ∞ }
Opp_Negotiation_length	Length of counterpart's negotiations	Continuous	{0, ∞ }
Deadline_expiration	Did deadline expire	Binary	{No, Yes}
Score_imbalance	Was user's score < rscore	Binary	{No, Yes}
Mean_offers_gap	Average time between each offer (sent by user)	Continuous	{0, ∞ }
Opp_Meanoffers_gap	Average time between each offer (sent by opponent)	Continuous	{0, ∞ }
MeanMessages_gap	Average length of each message (sent by user)	Discrete	{0, ∞ }
Opp_MeanMessages_gap	Average length of each message (sent by opponent)	Discrete	{0, ∞ }
Activity_deadline	Activity 48 hours prior to end of negotiations	Binary	{No, Yes}
Offers_2 nd _last_day	Number of offers <i>sent</i> by user on the 2 nd -last day of negotiations	Discrete	{0, ∞ }
Opp_Offers_2 nd _last_day	Number of offers <i>received</i> by user on the 2 nd -last day of negotiations	Discrete	{0, ∞ }
Offers_3 rd _last_day	Number of offers <i>sent</i> by user on 3 rd -last day of negotiations	Discrete	{0, ∞ }
Opp_Offers_3 rd _last_day	Number of offers <i>received</i> by user on 3 rd -last day of negotiations	Discrete	{0, ∞ }
Offers_5 th _last_day	Number of offers <i>sent</i> by user on 5 th -last day of negotiations	Discrete	{0, ∞ }
Opp_Offers_5 th _last_day	Number of offers <i>received</i> by user on 5 th -last day of negotiations	Discrete	{0, ∞ }
Offers_in_between	Number of offers <i>sent</i> by user from the beginning of negotiations till the 5 th -last day of negotiations	Discrete	{0, ∞ }

Time_distance	Time distance (hours) between agreement and deadline	Discrete	{0, ∞}
----------------------	---	-----------------	---------------

Variables captured in post-settlement phase

Table A.3: Representation of variables captured in post-questionnaire

After	User's desire to increase Internet use after INSPIRE	Binary	{No, Yes}
INeasy	User's ease of use with INSPIRE	Ordinal	{1,7}
Instructions_easy	User's ease of understanding system instructions	Ordinal	{1,7}
Messages_helpful	Level of aid messages provided user	Ordinal	{1,7}
Utility_values	Level of aid displayed utility values provided user	Ordinal	{1,7}
Graph_used	Was history graph used	Binary	{No, Yes}
Graph_inform	If used, was graph informative	Binary	{No, Yes}
Graph_infl_me	If used, did graph influence user's actions	Binary	{No, Yes}
Graph_infl_assessment	If used, did graph influence user's assessments of counterpart	Binary	{No, Yes}
Postsettlement_used_question	Did user use post-settlement mechanism	Binary	{No, Yes}
Difforig	If used, did user accept package different from original one	Binary	{No, Yes}
Postsettlement_mechanism_real	In real life, would user accept something similar to post-settlement mechanism	Binary	{No, Yes}
Agreement_satisfaction	User's satisfaction with agreement (if one reached)	Ordinal	{1,7}
Met_expectations	Whether negotiations met user's prior expectations	Ordinal	{1,7}
Control	Level of control user had during negotiations	Ordinal	{1,7}

Friendly	Level of friendliness user experienced during negotiations	Ordinal	{ 1,7 }
Performance	Level of satisfaction user has with their performance during negotiations	Ordinal	{ 1,7 }
Disclose_country	Counterpart revealed their country	Binary	{ No, Yes }
Disclose_id	Counterpart revealed their ID	Binary	{ No, Yes }
Country_guess	User's guess on counterpart's country	Nominal	
Knew_opponent	User knew with whom they were negotiating	Binary	{ No, Yes }
Opp_informative	Counterpart was informative	Ordinal	{ 1,5 }
Opp_persuasive	Counterpart was persuasive	Ordinal	{ 1,5 }
Opp_honest	Counterpart was honest	Ordinal	{ 1,5 }
Opp_exploit	Counterpart was exploitative	Ordinal	{ 1,5 }
Opp_cooperate	Counterpart was co-operative	Ordinal	{ 1,5 }
Surprised	Counterpart's actions surprised user	Binary	{ No, Yes }
Work_w_opponent	Like to work with negotiation partner on some other project	Binary	{ No, Yes }
See_opponent	Agreement to see counterpart again	Ordinal	{ 1,5 }
Predict_opponent	Predict counterpart's next moves	Ordinal	{ 1,5 }
Understood_opponent_priorities	Understood counterpart's priorities	Ordinal	{ 1,5 }
Practice	Will use NSS to practice negotiation skills in future	Binary	{ No, Yes }
Prepare	Will use NSS to prepare for a negotiation in future	Binary	{ No, Yes }
Nego	Will use NSS to conduct actual negotiation in future	Binary	{ No, Yes }

Appendix B: List of possible independent variables

The following forty-seven independent variables were considered important towards describing “Outcome of negotiations”.

Table B.1: Independent variables in modeling outcome of negotiations

1.	Age of negotiator (derived from Date of Birth)	2.	Gender	3.	Occupation
4.	Knew_opponent_country	5	Knew_opponent_id	6	Internet_access
7	Negotiation_experience	8	NSS_before	9	Case_understand
10	Expected_score	11	Reservation_score	12	Expected_friendliness
13	Weighing_issues	14	Weighing_options	15	Optimal_score
16	Messages	17	Activity_deadline	18	Offers_2 nd _last_day
19	Offers_3 rd _last_day	20	Offers_5 th _last_day	21	Offers_in_between
22	INeasy	23	Instructions_easy	24	Messages_helpful
25	Utility_values	26	Graph_used	27	Graph_inform
28	Graph_infl_me	29	Graph_infl_assessment	30	Met_expectations
31	Control	32	Performance	33	Disclose_country
34	Disclose_id	35	Country_guess	36	Knew_opponent

37	Opp_informative	38	Opp_persuasive	39	Opp_honest
40	Opp_exploit	41	Opp_cooperate	42	Surprised
43	Work_w_opponent	44	See_opponent	45	Predict_opponent
46	Understood_opponent_priorities	47	Negotiation_case		

Appendix C: Categorization of variables for loglinear modeling

Offers_2nd_last_day	1 2 3	0 2 3...7
Offers_5th_last_day	1 2	0 1...7
Offers_in_between	1 2 3 4	0,1 2 3 4...17
Met_expectations	1 2 3 4	1,2 3 4 3...7
Opp_cooperate	1 2 3 4	1 2 3 4,5
Instructions_easy	1 2 3	1 2 3...7

Appendix D: Sample INSPIRE negotiation

Negotiations between Misty and Smiley

The parties and the problem

Both Misty and Smiley have carefully read the information about their respective organizations to understand the problem and its issues. Since each negotiator can access the system and make a proposal independently, we will follow Misty's side of the negotiation.

There are only two *issues* in this simple negotiation: the *price* of the aircraft and the terms of the *warranty*. It has been established that the normal price of this aircraft is in the range of \$300 000 to \$320 000. The sensible increase is of \$10 000. Thus, the *price options* are \$300 000, \$310 000, and \$320 000. In this industry there are four types of warranty typically available. The options are: no warranty, a 6-month, one year, and a 2-year warranty.

Both negotiators analyze the two issues and their associated options in terms of their relevance to their respective organizations and move to the pre-negotiation phase.

Preparation

Issue rating

To prepare for the negotiations Misty and Smiley each rated the two issues. Note, that the pre-negotiation steps are conducted independently; one negotiator can *never* see the information (ratings) that the other negotiator enters.

Misty feels that price is far more important than warranty. Therefore, she assigns 70 points to price and 30 to warranty. Although Misty does not know it, Smiley feels that each issue is equally important and so Smiley assigns 50 points to each.

Misty's Issue Ratings	
Negotiation Issue	Rating
Price	70
Warranty	30

Smiley's Issue Ratings (Misty does not see this)	
Negotiation Issue	Rating
Price	50
Warranty	50

Option rating

Each issue has one or more options, for example, price has three options: 300 000 \$, 310 000 \$, 320 000 \$. After rating the issues, the options in each issue must also be rated similarly. In the INSPIRE system, for each issue at least one option must be assigned the maximum rating for the issue and at least one option must be assigned a rating of zero.

Misty considers the three options for the price of the aircraft and assigns the maximum rating (that is, 70) to the price of \$320 000 because Misty represents Rosa which wants to sell the aircraft. The lowest possible price is assigned a rating of zero. Misty considers the price of \$ 310 000 as somewhat acceptable and assigns a rating of 45.

Misty assigns ratings to each warranty option in a similar way. Note that for Misty, "no warranty" has the same maximum rating (30) as the 6 months warranty perhaps because the organization does not think it is possible that this plane will fail during the first 6 months of operation.

Price	Rating (Max = 70)	Warranty	Rating (Max = 30)
300 000 \$	0	No warranty	30
310 000 \$	45	6 months	30
320 000 \$	70	One year	10
		Two years	0

Package evaluation

Given user's ratings for each issue and each option, INSPIRE calculates ratings for complete packages that are the subject of negotiations. A package consists of price option and warranty option, for example, 320 000 \$ and No warranty is one complete package. INSPIRE presents a few packages and their ratings so that the user can assess if the INSPIRE's results accurately describe the user's preferences.

Misty's package evaluations are below.

Price	Warranty	Rating
320 000 \$	No warranty	100
320 000 \$	One year	80
310 000 \$	6 months	75
320 000 \$	Two years	70

The above ratings indicate how good the packages are given Misty's ratings of the issues and options. Comparing the packages Misty decided to change two ratings, for the second and fourth package. The second package while worse than the first (best) package is still quite good so Misty increases its rating from 80 to 82. The fourth package is not nearly as good as the third package and Misty downgrades its rating to 69.

Price	Warranty	Rating
320 000 \$	No warranty	100
320 000 \$	One year	82
310 000 \$	6 months	75
320 000 \$	Two years	69

INSPIRE uses now the issue, option and package ratings and it determines the utility function which closely reflects Misty's preferences. These function will be used to provide a rating for every considered package.

Negotiations

Exchange of offers and messages

Misty thought a while on how to begin negotiations. In the meantime Smiley prepared and sent the following offer together with a short message:

Smiley's opening offer and message						
<table border="1"> <tr> <td>Price</td> <td>300 000 \$</td> </tr> <tr> <td>Warranty</td> <td>Two years</td> </tr> </table>		Price	300 000 \$	Warranty	Two years	<p>Hi Misty, I'm Smiley and I represent Casa Ltd. in this negotiation. After extensive deliberations we have prepared an offer to purchase the aircraft. We think that it is a very good offer and hope that you will find it acceptable.</p>
Price	300 000 \$					
Warranty	Two years					
Misty's rating: 0						

Misty considers Smiley's offer and -- contrary to Smiley's expectations -- finds it unacceptable. Note that the offer's rating is 0. Misty prepares a counter-offer that better reflects Rose Inc. requirements. A message is attached to Misty's offer.

Misty's response

<table border="1"> <tr> <td>Price</td> <td>320 000 \$</td> </tr> <tr> <td>Warranty</td> <td>6 months</td> </tr> </table>		Price	320 000 \$	Warranty	6 months	<p>Dear Smiley, Thanks for your offer. Unfortunately Rosa Inc. cannot accept it. However, there is an offer that we propose and hope that you will be able to accept. Thank you, Misty</p>
Price	320 000 \$					
Warranty	6 months					
<p>Misty's rating: 95</p>						

Note that the rating of the above offer is 95. From Misty's ratings of Price and Warranty (see section "Option rating") it would appear that Misty's offer "320 000 \$ and 6 months" should be rated 100. However, the adjustments that Misty made in the Package evaluation step caused a drop in rating to 95. Although these adjustments were made only to two packages they affect ratings of all possible packages because they modify Misty's utility function.

Graphing negotiations

At anytime during the negotiation, a graphical overview of the history of offers can be viewed. The graph plots the users' ratings of offers sent (in green) and received (in red) and the time that the offers were sent. The graph can be viewed from the "View offer and message history" link.

Below are two graphs depicting the history of negotiations. The graph on the left-hand side represents the history from Misty's point of view because Misty's ratings are used. *Note that Smiley cannot see this graph.* The graph that represents Smiley's perspective is shown on the right-hand side.

We show the two graphs to indicate that:

1. the two graphs are *different* because each represents a different point of view;
2. sometimes what one side considers as a concession is not a concession from the other side's point of view.

Compare Smiley's two offers: Offer 2 and Offer 3. In Misty's graph Offer 3 is worse than Offer 2. This is because in terms of Misty's utility function Offer 3 has lower rating than Offer 2. Misty may think that Smiley does not want to negotiate because Smiley makes a negative or reverse concession (proposes an offer that is worse than proposed previously).

While Misty cannot see the graph on the right-hand side, Misty knows that such a situation is possible. Indeed, Smiley thinks that a small concession has been made on behalf on Casa Ltd. because the rating in terms of Smiley's utility function for Offer 3 is lower than for Offer 2. The graph on the right-hand side clearly indicates this.

Both of our negotiators should be well aware that such a situation may happen. What one side perceives as a concession, the other side may not see as such. Worse, the other side may consider this as a "reverse concession" or a hardening of one's position.

A message

Misty and Smiley continue to exchange offers and messages. On one occasion Misty decided not to send an offer but ask Smiley for clarification. Thus, Misty sends only the following message:

The compromise

Misty and Smiley exchange the offers and messages until, in the fourth round, Misty presented the following offer:

Misty's rating: 55

Smiley reviews Misty's last offer, which is shown below. (Note that the offer's rating now reflects Smiley's preferences).

<table border="1"><tr><td>Price</td><td>310 000 \$</td></tr><tr><td>Warranty</td><td>One year</td></tr></table>		Price	310 000 \$	Warranty	One year	<p>Smiley, On behalf of the Rosa Inc. I want to thank you for your offer. We have made only one change to it, that is, we propose a one year warranty instead of two years. I hope that you will find this new offer acceptable. Misty</p>
Price	310 000 \$					
Warranty	One year					
		Smiley's rating: 53				

After short consideration Smiley accepts Misty's last offer. Smiley does it by selecting, on the INSPIRE page, the option:

Yes, I accept my counterpart's most recent offer as listed above.

Post-settlement

Efficient packages

In some negotiations it may happen that the parties reach an agreement but there is one or more packages that are better than the accepted offer for *both* sides. Note, that *better* is measured with the parties' utility functions. Thus, there may be a package for which the two ratings are higher than the package that has been accepted.

INSPIRE has a post-settlement stage, during which it uses the preference information provided by each user to determine whether it is possible to construct packages that are better for the two parties. In this negotiation INSPIRE determined that Misty's and Smiley's settlement could be improved. In this simple negotiation there is only one such package:

Price	320 000 \$
Warranty	Two years

Misty asked Smiley about this new offer and after a short exchange they both agreed that this last package is superior to the compromise package. Thus the negotiation is completed.

Negotiation graph

The complete graph of the history of negotiation shows the effect of accepting this alternative. To show the differences between Misty's and Smiley's view points we

present both graphs *noting again that each negotiator can see only his or her graph and cannot see the opponent's graph.*

The graph that Misty sees.

The graph that Smiley sees.