Chapter 3
Intelligent Knowledge and Habitual Domain

This paper is to enhance our understanding about the second-order data mining. In particularly, we examine the effect of human cognition on the creation of intelligent knowledge during the second-order data mining process. Prior studies have suggested that human cognition plays an important role in the second-order data mining process during which intelligent knowledge was discovered (Baker et al. 2009). Given the knowledge that no single data mining model outperforms others for all problems, a common practice in data mining projects is to run multiple data mining models at first and then invite a group of people to collaboratively make judgments on these data mining models' performance. These judgments often diverge. Little research exists to explain why these variations of human judgments occur.

The theory of habitual domains (Yu 1990, 1991, 2002; Yu and Chen 2010) provides a useful theoretical base for explaining the behavioral mechanism that guides human minds' operations. Drawing on the theory of habitual domains, in this article, we develop a theoretical model to explain the influence of habitual domains' characteristics on human judgments made on data mining models' performance. Specifically, among the many data mining models, this study chose to use the classifiers. A field survey was administrated at a multidisciplinary research. A social network data analysis technique was used to test the proposed relationships in the model. The specific research question of this study is:

What are the relationships between human habitual domain characteristics and the convergence of human judgments on data mining performance in the second-order data mining process?

Intelligent knowledge was created during second-order data mining through human judgments. A clear understanding about why people's judgments about classifiers diverge or converge will inform the design of the guidance for selecting appropriate people to evaluate/select data mining models for a particular problem. Thus, costly mistakes can be avoided when appropriate people are selected.

The rest of the chapter is organized as follows. Section 3.1 introduces the theory of habitual domains and related hypotheses. Then the overall research design and experimental results are presented in Sect. 3.2. Section 3.3 discusses the limitations of the study. In Sect. 3.4 and 3.5, we present the discussion and conclusion of our study.

Y. Shi et al., *Intelligent Knowledge*, SpringerBriefs in Business,
DOI 10.1007/978-3-662-46193-8_3

3.1 Theory of Habitual Domain

3.1.1 Basic Concepts of Habitual Domains

The analysis of intelligent knowledge, rough knowledge, and human knowledge leads us to wonder how various types of human knowledge along with the results from data mining classifiers contribute to the creation of intelligent knowledge. The theory of habitual domains provides us a theoretical foundation. The theory of habitual domains (Shi and Yu 1987; Yu 1990, 1991, 2002; Yu and Chen 2010) attempts to describe and explain the human's behavior mechanism that guides people in making decisions and judgments. The central proposition of habitual domain theory is that an individual thinks and acts in a habitual way, which is influenced by one's habitual domains. The theory of habitual domains builds on three necessary conditions: (1) our perceptions of the environment can be reached at steady states in our brain, (2) most of daily problems we counter happen regularly, and (3) human tends to take the most convenience way of dealing with daily problems (Yu 1990). In this chapter, we suggest that the theory of habitual domains is useful in explaining the elusive process involved in our minds in the process of intelligent knowledge creation.

Yu and Chen (2010, p. 11) defined the habitual domains as "the set of ideas and concepts which we encode and store in our brain can over a period of time gradually stabilize in certain domain". According to the theory of habitual domains, human attain knowledge or make decisions based on external stimulus and self suggestion. Unless there is an occurrence of extraordinary events, an individual tends to make decisions by following a stable mental model established in his/her mind. As a result, we can observe that each of us has his/her own set of habitual ways of doing cognitive related tasks, such as problem solving, decision making, and learning.

The theoretical building blocks of the habitual domains are ideas and operators. Ideas refer to specific thoughts resides in our minds. Operators are the actions, specifically the "thinking processes or judging methods" (Yu 1990, p. 118). The theory of habitual domains developed eight hypotheses to capture the basics to how our minds work. In particular, the analogy/association hypothesis is most relevant to this study. The analogy/association hypothesis is stated as follows:

"The perception of new events, subjects or ideas can be learned primarily by analogy and/or association with what is already known. When faced with a new event, subject or idea, the brain first investigates its features and attributes in order to establish a relationship with what is already known by analogy and/or association. Once the right relationship has been established, the whole of the past knowledge (preexisting memory structure) is automatically brought to bear on the interpretation and understanding of the new event, subject or idea (Yu and Chen 2010, p. 8)."

According to this hypothesis, analogy/association enables the brain to comprehend and interpret the new arriving information from the external environment. People with different habitual domain characteristics will perceive rough knowledge differently and thus make different judgments on the classifiers' performance.

Though there are a variety of variables constitute people's habitual domain characteristics, we choose these specific characteristics—level of education, areas of specialty, and prior experience with data mining—which are most relevant to the context of second-order data mining. The linkages between these three characteristics and the theory of habitual domains are explained in the next subsection. Hypotheses are developed.

3.1.2 Hypotheses of Habitual Domains for Intelligent Knowledge

The theory of habitual domains (Yu 1990) identifies four basic components of habitual domains. These four components are: potential domain, actual domain, activation probabilities, and reachable domain.

Potential domain is a collection of ideas and operators that can be potentially activated. Actual domain is the activated ideas and operators. The overall potentially reachable collection of ideas and operators based on the potential domain and the actual domain is called reachable domain. The activation probabilities define the degree to which subsets of potential domain can be actually activated at a particular time. Subsets of potential domain vary in the degree of their likelihood to be activated for given problems.

In most cases, a large size of potential domain is preferable. That is because holding all other things equal, the larger the potential domain, the more likely that a larger set of ideas, concepts or thoughts will be activated. Moreover, if the ideas, thoughts, and knowledge are stored in a systematical way and are integrated seamlessly, individuals are more likely to make judgments and cope with problems better.

The size of a potential domain is greatly contingent on an individual's habitual domain formation. The theory of habitual domains proposed eight approaches by which individuals form their habitual domains. The eight approaches are: active learning, projecting from a higher position, self awareness, active association, changing the relevant parameters, retreating, changing the environment, and brainstorming. Based on these eight approaches of habitual domains formation, this paper proposed that an individual's habitual domain's characteristics can be described by examining an individual's background in these eight areas. The assumption we made here is that for each of the eight approaches, if people follow different paths within the approach, then people's habitual domains will be formed differently. In other words, peoples' habitual domains' characteristics can be described by assessing peoples' background in each of the approaches by which s/he form the habitual domains.

Considering the purpose of this study along with the consideration of empirical assessment, this paper focused on the active learning dimension. The habitual domain is a multi-dimensional and complex concept. The theory of habitual domain has identified three dimensions of one's domain, namely behavior function, events, and external interaction. Each dimension has several specific components. Given the multidimensional nature of habitual domains, checking one's habitual domain thoroughly is challenging. Yu (1990) suggest that a study could only focus on one

component based on the study's purpose. Given the purpose of the study is to understand why people make different judgments on classifiers' performance on data sets and plus people's such decision making is to a large extent influenced by ones' learning experience, therefore, it is adequate to only check the active learning experience of people at this point. More approaches should be considered when different goals of the study are taken.

Active learning emphasizes on the various external sources (such as experts, media, and school education) around us. Active learning will not only give us a higher chance of getting new and innovative ideas but will also enable us to be able to more efficiently integrate ideas we have before and make those ideas more accessible.

We specifically identified three areas related to active learning. Those three areas are: level of education, areas of specialty, and prior experience with data mining. We posit that these three areas make up a significant high proportion of one's active learning experience. People who have similar background in each of the three areas of active learning will possess similar habitual domains and thus make similar judgments on data mining classifiers' performance. In the following paragraphs of this section, we will describe each of these three areas in details and develop hypotheses.

First, level of education is concerned with how many years of formal school education one has taken. From many years of education in school, each of us has been exposed to many new ideas and new knowledge from reading books, listening to lectures, and interacting with our classmates. Attending classes not only provides us new ideas and knowledge but also facilitates the absorption of these new ideas and knowledge in our minds by repetitions. In an experimental study, Macpherson (1996) found that educational background, specifically the number of years of education, has a significant positive effect on individuals' capabilities of generating insights. Another study reveals that education can decrease the anxiety toward the use of computer (Igbaria et al. 1989). Bower and Hilgard's study (1981) suggest that higher level of education would enhance individual's cognitive capabilities and thus accelerate the individual's learning process especially in novel situations. Considering the situations people face to the hidden patterns—which usually reveals unknown rules or hidden patterns, we construct the following hypothesis.

H1: The closer the levels of education between individuals, the higher the degree to which people agree on judging performance of classifiers for a particular database.

Second, areas of specialty refer to the: (1) research areas and majors that individuals peruse in college (2) individuals' domain knowledge. Working or studying in a special area will provide one with relatively in-depth knowledge in that particular area. Further, working in a specific specialized area enables one communicate with a group of peers and can help one gain new knowledge and insights (Astin 1993). A study conducted by Paulsen and Wells (1998) found that students who studied in similar majors (according to hard-soft, pure-applied dimensions of Biglan's (1973) classification of academic fields) held similar epistemological beliefs, which are beliefs about the nature of knowledge and learning. Their study found that students majored in soft and pure fields were less likely than others to hold naïve beliefs in certain knowledge.

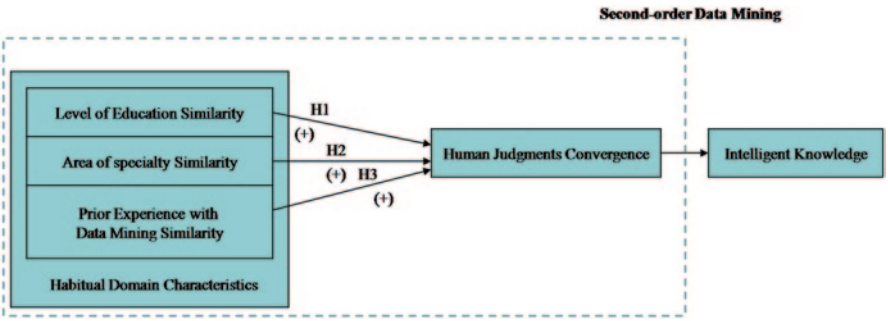

Fig. 3.1 Influence of Habitual Domains on Human Judgments Convergence

The importance of areas of specialty on the successful application of data mining has also been recognized in the field of data mining. For example, Ambrosino and Buchanan (1999) found that models that incorporated domain knowledge performed significantly better than models without considering domain knowledge in predicting the risk of mortality in patients with a specific disease. In a study of applying data mining to bank loans problem, Sinha and Zhao (2008) examined and compared performances of seven well-known classifiers. They found that models that incorporated the pre-derived expert rules outperformed models without those expert rules.

Thus, we have the following hypothesis.

H2: The closer the areas of specialty between individuals, the higher the degree to which people agree on judging performance of classifiers for a particular database.

Third, prior experience with data mining is about individuals' past experience related to data mining. Such experience can be gained by attending data mining related classes, leading or participating in data mining projects, using data mining software, developing data mining algorithms, and reading books or literatures related data mining. We suggest that an individual's experience with data mining greatly influences one's attitude toward various data mining classifiers. Empirical studies have found that previous experience with certain technologies can either hinder or foster one's adoption of a new technology (Harrison et al. 1992). For example, one study found that users resisted using an unfamiliar technology because of switching costs (Scholtz et al. 1990). Thus, we build the following hypothesis.

H3: The closer the experience with data mining between individuals, the higher the degree to which people agree on judging performance of classifiers for a particular database.

The research model is shown in Fig. 3.1. Building on the theory of habitual domains, the conceptual model describes the convergence of human judgments on data mining is positively influenced by the similarity of people's level of education, by the similarity of people's areas of specialty, and by the similarity of people's prior experience with data mining. The model is constructed and examined at the team level. The creation of intelligent knowledge from rough knowledge during second-order data mining is a complex process, this article focuses on studying the influence of habitual domain characteristics on the convergence of human judgments on classifiers' performance.

3.2 Research Method

The overall research design is a field survey. A pilot study was conducted to test the reliability and validity of the survey and the field procedure.

3.2.1 Participants and Data Collection

Considering the purpose of the study is to test if habitual domain characteristics affect people's judgments on data mining, it is necessary to have subjects with diverse background. Thus, the study collected data from members employed in a multidisciplinary research institute in China. The research institute has conducted several large data mining projects in the past. This research institute consists of a total of five research labs concentrating on various areas, ranging from e-commerce, green energy, to data mining. Researchers in the institute have backgrounds as varied as management information systems, computer science, economics, and biology. Of the 38 respondents, 42 percent of the respondents were male and 58 percent of the respondents were female. The distribution of respondents' age is shown in Table 3.1.

In the study, we first run eight classifiers[1] on two data sets and recorded the performance of each classifier given a set of measures. Then we administrated the survey questionnaire. The session lasted for a total of 4 h. An author of the paper gave an introduction to the background of the survey. The questionnaire collected participants' demographic information and also asked the participants to rate the performances of eight classifiers on the two large-scale data sets. The participants rated the performance of the classifiers on each of the two data sets according to the seven standard evaluation criteria (as is shown in Appendix A).

The Nursery Database is a public data set from the Machine Learning Repository of the University of California, at Irvine (UCI). It was derived from a hierarchical decision model originally developed to rank applications for nursery schools. It was used during several years in 1980's when there was excessive enrollment to these schools in Ljubljana, Slovenia, and the rejected applications frequently needed an objective explanation[2]. PBC Dataset is a data set related to credit scoring from

Table 3.1 Frequency on Subjects' Age

Age	Frequency	Percentage (%)
20–30	28	73.68
30–40	6	15.79
40–50	3	7.89
Above 50	1	2.63

[1] The eight methods are J48, Nbtree, Baysnet, Naivebays, Logistic, Support Vector Machine (SVM), Multiple Criteria Linear Programming (MCLP), and Multiple Criteria Quadratic Programming (MCQP).

[2] Http://archive.ics.uci.edu/ml/datasets/Nursery

China. After preprocess, we got a data set with 1600 samples. 800 of them were classified as good customers and 800 of them were classified as bad customers. 80 variables were designed to reflect the behaviors of the customers.

3.2.2 Measures

3.2.2.1 Habitual Domains Characteristics

Measures for habitual domain characteristics—level of educational, prior experience with data mining, and areas of specialty—were enabled by asking participants to check the items that best describe their current status. Specifically, to assess subjects' educational background, we asked each participant to answer one multiple choice question that asks their highest degree (IV1). Second, area of specialty was measured by asking subjects' current major and research area (IV2). Third, to assess subjects' prior experience with data mining (IV3), we used multiple measures, including: their level of acquaintance with data mining, if ever participated in data mining related projects, if ever studied data mining related courses, level of acquaintance with data mining methods, and level of familiarity with data mining software.

3.2.2.2 Dependent Variables

Dependent variables in this study were participants' judgments on data mining classifiers' performance. Specifically, dependent variables consist of participant's ratings on performance of each of the eight classifiers on the data sets. We ran eight data mining classification algorithms on two large-set data sets. The second section of the questionnaire presented the results of performance of data mining algorithms on two datasets according to the selected standard measures. We asked subjects to evaluate the performance of the data mining algorithm on each of the seven measures, using a 10-point response scale (1 = very bad performance and 10 = outstanding performance).

3.2.3 Data Analysis and Results

3.2.3.1 Descriptive Analysis

We first analyze the psychometric properties of the acquaintance with data mining (IV3) by running a reliability analysis in SPSS. Results showed the subscales of IV3 have good internal consistency, $\alpha = 0.93$. Table 3.2 shows the frequency of individuals' educational background.

The descriptive statistic of the areas of specialty of individuals is shown in Table 3.3.

Table 3.2 Frequency on Subjects' Educational Background—Level of Study

Degree	Frequency	Percentage (%)
Master Graduate Student	14	36.8
Doctoral Graduate Student	14	36.8
Doctor	10	26.3
Total	38	100

Table 3.3 Frequency on Subjects' Educational Background—Major

Major	Frequency	Percentage (%)
Social Science	0	0
Management Science	28	73.7
Information Technology	10	26.3
Total	38	100

Table 3.4 Ratings on Classifiers' Performance on the Nursery Database

Classifier	Mean	SD
J48	8.11	1.29
Nbtree	7.78	1.61
Baysnet	6.11	1.90
Naivebays	6.22	1.61
Logistic	7.22	1.79
SVM	8.81	1.29
MCLP	8.46	1.69
MCQP	7.84	1.59

Table 3.5 Ratings on Classifiers' Performance on the PBC Database

Classifier	Mean	SD
J48	8.03	1.62
Nbtree	7.30	1.75
Baysnet	5.65	1.79
Naivebays	5.22	1.70
Logistic	7.11	1.52
SVM	5.41	1.84
MCLP	7.16	1.35
MCQP	7.65	1.46

Results showed that participants were generally somewhat familiar with data mining ($M=2$, $SD=0.81$).

The descriptive analysis of subjects' judgments on the eight classifiers' performance on Nursery Database indicated that SVM got the highest average score ($M=8.81$, $SD=1.29$) and Baysnet got the lowest average score ($M=6.11$, $SD=1.9$). Table 3.4 showed the descriptive statistics.

For classifiers' performance on the PBC database, results showed that J48 received the highest average score ($M=8.03$, $SD=1.62$). Naivebays received the lowest average score ($M=5.22$, $SD=1.70$). Table 3.5 presented the descriptive statistics for all classifiers' scores on the PBC database.

3.2.3.2 Geary's C Analysis

We identify Geary's C (1954) statistic as a perfect fit for testing the type of hypotheses in the present study. Geary's C is adapted for social network analysis from their origins in geography, where they were developed to measure the extent to which the similarity of the geographical features of any two places was related to the spatial distance between them (Geary 1954). Geary's C has been widely used in social network analysis for testing the homophily hypothesis which asks a question of: Is there a tendency for actors who have more similar attributes to be located closer to one another in network? Since the hypotheses of present study concerned about if the closeness of experts' habitual domain characteristics would affect their judgments on data mining algorithms' performance, thus it is obvious for us to use Geary's C for testing the hypotheses of this study. This social network data analysis method, Geary's C statistic has two advantages. First, it avoids merely focusing on subjects' answers to individual question, rather it provides a global view of the subject's responses to all of the questions. Second, it simplifies the dependent variables and makes it easy to conduct the correlation analysis.

It should be noted that although MANOVA method allows the analysis of the effects of more than one independent variable on two or more dependent variables, MANOVA method has strict assumptions on the data, such as normality of dependent variables, linearity of all pairs of dependent variables, and homogeneity of variances. The robustness of MANOVA results will be significantly affected when these important assumptions are violated. Unfortunately, we explored the two data sets on all the three assumptions of MANOVA. Two of the assumptions (normality and linearity of dependent variables) were violated, and only the homogeneity of variances assumption was met.

Therefore, we consider Geary's C statistic to test the effects of independent variables on dependent variables. To apply Geary's C statistic in our study, for each of the two datasets, we used the affiliation network method[3] in UCINET (Borgatti et al. 2002) to get an adjacency matrix[4] of all participants based on their judgments on data mining algorithm performance. This adjacency matrix thus described the "closeness" of each pair of participants on their overall perceptions on the data mining algorithm performance. Then, we create another attribute table that contains all information of participants' habitual domain characteristics. UCINET was used to calculate the Geary's C measure. Table 3.6 and 3.7 present the Geary's C statistic results.

Correlation results indicated that educational level is highly positively correlated with the closeness between individual's judgments on classifier's performance. To

[3] Affiliation network is a one mode network, which has been first applied to study the southern women and the social events in which they attended. The affiliation network describes how many same events each two of women have attended. Then affiliation network has been applied in many cases to establish the pairwise ties between actors Wasserman, S., and Faust, K. Social Network Analysis, 1995.

[4] The computational process to get the Geary's C and the adjacency was shown in Appendix C.

Table 3.6 Geary's C Correlation Analysis on the Nursery Database

IV	DV	Geary's C
LES	Closeness between individual's judgments on classifier's performance	0.99[a]
ASS	Closeness between individual's judgments on classifier's performance	1.004
PEDMS	Closeness between individual's judgments on classifier's performance	0.98[b]

IV independent vriable, *DV* dependent variable, *LES* level of education similarity, *ASS* area of specialty similarity, *PEDMS* prior experience with data mining similarity
[a] Indicates a correlation is significant at 0.1
[b] Indicates a correlation is significant at 0.01

Table 3.7 Geary's C Correlation Analysis on the PBC Database

IV	DV	Geary's C
LES	Closeness between individual's judgments on classifier's performance	0.99[a]
ASS	Closeness between individual's judgments on classifier's performance	1.005
PEDMS	Closeness between individual's judgments on classifier's performance	0.98[a]

IV independent vriable, *DV* dependent variable, *LES* level of education similarity, *ASS* area of specialty similarity, *PEDMS* prior experience with data mining similarity
[a] Indicates a correlation is significant at 0.1
[b] Indicates a correlation is significant at 0.01

put it another way, the degree to which individuals agree on classifier's performance is positively influenced by the similarity between individual's educational levels. Prior experience with data mining also indicated a significant influence on individual's agreements on data mining algorithms performance. However, on both two data sets, areas of specialty didn't show a significant relationship with people's judgments on classifier's performance. Overall, Hypothesis 1 and Hypothesis 3 were supported. Hypothesis 2 was rejected.

3.3 Limitation

Prior to discussing the findings of the study, limitations of the study must be acknowledged. First, the sample itself offers some important limitations. The setting for the study was a research institution and respondents were mostly students and a few faculties who worked in this institution. Thus, the generalizability of the respondents' behaviors to a more general population may be somewhat limited. One mostly mentioned drawback of using students as subjects is that the significant differences between students and the targeting groups. In this study, the targeting groups will be the data mining customers who propose, sponsor, evaluate, and

eventually implement a data mining project. The targeting groups may possess very different background in terms of educational background, areas of specialty, and previous experience compared to students of this study.

Additionally, this study only asks participants' opinions on classifiers' performance on two data sets. Moreover, a data set is from UCI rather than a real-world data set. One major criticism with UCI data set is that the data set in UCI is often biased because pre-processing of the data. Future study should provide classifiers' performance on more data sets so that the bias resulting from the data sets can be reduced.

Another limitation of the study comes from the type of data analysis we conducted. Geary's C analysis doesn't allow an interaction analysis of data. This autocorrelation method can only detect the association between subjects' attributes and subjects' responses on a set of questions. The impact of interactions among subjects' attributes, such as level of education, areas of specialty, and prior experience with data mining, cannot be obtained. Future research can acquire larger sample of data and conduct a MANOVA analysis to see if there are interaction effects of individuals' habitual domain characteristics on their judgments on data mining classifiers.

Finally, this study is a first attempt in applying habitual domain theory in understanding peoples' judgments made on data mining classifiers' performance. Therefore, the three constructs, level of education, areas of specialty, and experiences in data mining, need further refinement. For example, while we gave a formal description of areas of specialty in this study, the study did not specify which several areas of specialty should be considered in the assessment of individuals' habitual domains.

3.4 Discussion

People intend to take full advantage of data mining through discovering intelligent knowledge from the data mining results. Accordingly, data mining researchers have begun to explore deriving intelligent knowledge from data mining in this stage (Bendoly 2003; Zhang et al. 2009). Research activities that are interested in transforming data mining results into actionable intelligent knowledge are called "second-order" data mining. This paper proposed that the theory of habitual domain provides a useful theoretical lens to study "second-order" data mining. Habitual domain theory is proposed to account for the mechanism through which human make decisions and judgments. The theory of habitual domain operationalized habitual domain in four specific domains: potential domain, actual domain, activation probabilities, and reachable domain. Further, the theory proposed that such human habitual domains are expanded through active learning, specifically formal school education and important personal experience.

This paper derived empirically testable hypotheses based on the habitual domain theory. In our experiments, we found support for our hypotheses that people's judgments on data mining classifiers' performance are influenced by people's education

and prior experience with data mining. Education was found to be an important factor on peoples' perceptions on classifiers' performance. People's prior experience with data mining was also revealed as a predictor to peoples' evaluation on classifiers' performance with statistic significance.

The analysis, however, didn't confirm the hypothesized positive effect of areas of specialty similarity on people's convergence on classifiers' performance. To put it another way, this results indicated that individuals' judgments on classifier's performance will not be significantly influenced by individuals' majors. One possible explanation is that the majors of participants in the study were not diverse enough. This study only had individuals from these three majors: Computer Science, Financial Engineering, and Management Science. It is possible that students from these three majors show similar attitudes on data mining classifiers' performance on various data sets. A study conducted by Tikka (2000) found that students of majors related to technology and economics showed similar attitude toward the environment adopted a more negative attitude toward the environment and, on average, had fewer nature-related hobbies than students in general.

One key advantage of understanding what habitual domains characteristics influence people's judgments making on data mining methods is the opportunity it presents for training interventions to manipulate people's perceptions about a classifier. Since education and previous experience with data mining have significant effect on people's perceptions on classifiers, designing better training will increase the likelihood that novice data mining developers make quality judgments as data mining experts do.

Having a group of people with similar habitual domains characteristics can benefit data mining project teams in terms of reducing conflicts in data mining algorithms. Since 1980s, numerous data mining algorithms have been developed. But no single one data mining algorithm has been proved to be able to outperform all the other algorithms in all tasks. Therefore, in the real world data mining projects, data mining teams have to carefully compare among more than one data mining methods and choose one that has the best functioning performance. Depending upon ones' past educational background and experience with data mining, people will possess different views toward the data mining methods' performance. Having people with similar habitual domains characteristics will help the team establish a shared understanding about the data mining methods' advantages and disadvantages, and thus help the data mining project team to reach a convergent opinion on which data mining method to be used. But having people with similar habitual domains may also place a potential risk of entering a decision trap to the data mining project team. For instance, it is possible that all people converge on a wrong decision when the team faces an unusual problem of data mining. With the coming of the big data era (i.e. large scale of data and integration of both structured and unstructured data) (Chen et al. 2012), the chance of dealing with unfamiliar data mining task or using unfamiliar data mining tools increases significantly. Therefore, given unusual data mining tasks or unfamiliar data mining algorithms, it is important for the data mining project teams to choose team members with diverse educational background and data mining experience, so that the team can make an optimal decision on choosing a data mining method.

3.5 Remarks and Future Research

The broad goal of the chapter is to enhance our understanding about the second-order data mining, particularly the creation of intelligent knowledge by human from data mining results. This study drew on the theory of habitual domains to develop a conceptual model that explains why human judgments on data mining performance are different. The study further conducted a field survey to empirically test the model. The study adopted a social network analysis method, Geary's C, for analyzing the data to get a global view of the correlation between participants' attributes and their responses. The study findings support two of the three hypotheses proposed in the model. First, the hypothesis of education's influence on human judgments is supported. Second, the empirical study identifies a significant correlation between human's previous experience with data mining and human's judgments on data mining performance.

This chapter took the first step in empirically testing the effect of human cognitive psychology characteristics on the creation of intelligent knowledge at second-order data mining. The findings of this paper provide evidence for the variations of human judgments on classifiers' performance when human possess varied cognitive psychology characteristics. These findings are valuable in understanding the important role of human in the stage of second-order data mining. Most of present studies of data mining either ignore the role of human or symbolize human as agents in the post-stage of data mining. While it could be argued from this study that human's complex cognitive psychology characteristics play a significant role in the creation of intelligent knowledge from data mining results. It should be noted that intelligent knowledge is created based on human judgments made on rough knowledge. Such human judgments are a function of various human prior knowledge, rough knowledge, and human's habitual domain characteristics.

This research presents interesting directions for future research. Since there is no one data mining method outperform all the other data mining methods in all kinds of tasks, choosing a most appropriate data mining method for a given task is one important step influencing the overall data mining project success. Experts of data mining possess implicit knowledge that guides them in selecting the best data mining method. The findings of this research lead us to wonder that implicit knowledge of data mining experts can have linkages with experts' past experience and educational background. Understanding what type of experience and educational background are mostly founded in data mining experts is crucial in training data mining analysts. Future research could focus on understanding this issue thoroughly.

It is unknown from this study that what interaction effects there are between the habitual domain characteristics and the data mining methods' performance evaluation. Future research can conduct a survey with a larger sample size to test if the interaction effects exist.

Another future research direction is to apply the habitual domains theory in understanding the overall data mining project success. Just as the case with all types of project, a data mining project that is accepted and actually used by the end users is a true successful project. As is said thousands of times in the data mining literature,

customers of data mining want to discover innovative ideas from the hidden patterns of data mining. But, without domain knowledge or being lacking in the domain knowledge, it is challenging for data mining analysts to understand what ideas count for innovative ideas from the customers' perspective. Understanding the preferences of customers and being able to have a shared understanding with customers about what ideas are innovative ideas is of critical importance to the overall success of data mining project. The habitual domains theory not only conceptually describes how human obtain, store, process and apply information from the world in terms of concepts and propositions, but also prescribes ways of expanding humans' habitual domains and discussed the characteristics of information that would catch people's eyes. The theory of habitual domains possesses great potential in developing useful constructs to predict the acceptance and continuing usage of data mining.

Appendix A: Summary Of Data Sets, Classifiers and Measures (Table A)

Table A Data Sets, Classifiers, and Measures

Data sets	The Nursery Database
	The PBC Database
DMC	Decision Tree
	NbTree
	Baysnet
	Naivebays
	Logistic Regression
	SVM
	MCLP
	MCQP
Measures	Correctly Classified Instances
	Kappa Statistic
	Mean Absolute Error
	Negative—TP Rate
	Negative—FP Rate
	Positive—TP Rate
	Positive—FP Rate

DMC Data Mining Classifiers

Appendix B: Questionnaires for Measuring Dependent Variables (Table B-1 and B-2)

Table B-1 Questionnaire Used For the Nursery Database

Score of Algorithm								
Measure		J48	Nbtree	Baysnet	Naivebays	Logistic	SVM	MCLP	MCQP
Correctly Classified Instances		0.97	0.97	0.9	0.9	0.93	0.99	0.99	0.97
Kappa Statistic		0.96	0.96	0.86	0.86	0.89	0.98	0.98	0.94
Mean Absolute Error		0.02	0.02	0.08	0.08	0.04	0.01	0.01	0.03
Not_Recom	TP Rate	1	1	1	1	1	1	0.98	0.99
	FP Rate	0	0	0	0	0	0	0	0.04
	F-Measure	1	1	1	1	1	1	0.98	0.96
Recommend	TP Rate	0	0	0	0	0	0	1	0.96
	FP Rate	0	0	0	0	0	0	0.02	0.01
	F-Measure	0	0	0	0	0	0	0.99	0.98
Priority	TP Rate	0.95	0.96	0.9	0.9	0.89	0.98		
	FP Rate	0.02	0.02	0.1	0.1	0.06	0.01		
	F-Measure	0.96	0.96	0.86	0.86	0.89	0.98		
Very_Recom	TP Rate	0.73	0.7	0.06	0.06	0.74	0.9		
	FP Rate	0.01	0	0	0	0.01	0		
	F-Measure	0.76	0.79	0.11	0.11	0.77	0.94		
Spec_Prior	TP Rate	0.98	0.99	0.87	0.87	0.9	0.99		
	FP Rate	0.02	0.02	0.05	0.05	0.05	0.01		
	F-Measure	0.97	0.98	0.88	0.88	0.9	0.98		

Table B-2 Questionnaire Used For the PBC Database

Score of Algorithm								
Measure		J48	Nbtree	Baysnet	Naivebays	Logistic	SVM	MCLP	MCQP
Correctly Classified Instances		0.87	0.86	0.75	0.70	0.84	0.71	0.84	0.86
Kappa statistic		0.74	0.72	0.50	0.39	0.69	0.43	0.68	0.84
Mean absolute error		0.18	0.16	0.25	0.30	0.21	0.29	0.16	0.16
Negative	TN rate	0.94	0.89	0.83	0.93	0.85	0.53	0.88	0.86
	FN rate	0.20	0.17	0.33	0.54	0.16	0.10	0.20	0.18
	F-measure	0.88	0.86	0.77	0.75	0.84	0.65	0.85	0.85
Positive	TP rate	0.80	0.83	0.67	0.46	0.84	0.90	0.80	0.82
	FP rate	0.06	0.11	0.17	0.07	0.15	0.47	0.12	0.14
	F-measure	0.86	0.85	0.73	0.60	0.84	0.76	0.83	0.83

APPENDIX C: Geary'S C Statistics

We illustrate how to manually compute the Geary's c measure using the following example.

Suppose we have three subjects x, y, z. For each of them, we measured three attributes A, B, C. Table C-1 shows the three subjects' attributes' values. We also computed an adjacency matrix W in Table 3.2 that describes the closeness for each pair of the three subjects.

Table C-1 Attributes' Values of Three Subjects

Subjects	Attribute A	Attribute B	Attribute C
x	3	4	5
y	5	3	6
z	4	7	8

Step 1: Construct the adjacency matrix, that is, the W, using the minimum method from affiliation network method.

The minimum method examines two subjects' values on each of the attributes, selects the lowest scores and then sums. For example, for subjects x and y, it yields $3+3+5=11$, it might means the extent to which subject x and y jointly agree on the three attributes A, B and C. Using this method, we filled out the adjacency matrix. (Table C-2)

Table C-2 Adjacency Matrix for Three Subjects

	x	y	z
x	12	11	12
y	11	14	13
z	12	13	19

Step 2: Calculate the Geary's c for each pair of subjects on each of the three attributes. First, let us calculate the Geary's c attribute A.

$$C = \frac{(N-1)\sum_i \sum_j \omega_{ij} \left(X_i - X_j\right)^2}{2W\sum_i \left(X_i - \bar{X}\right)^2}$$

$$N = 3, X_1 = 3, X_2 = 3, w_{12} = 11, w_{13} = 12, w_{23} = 13$$

$$C_A = \frac{(3-1)*2*[11(3-5)^2 + 12(3-4)^2 + 13(5-4)^2]}{2*107*[(3-4)^2 + (5-4)^2 + (4-4)^2]} = 0.65$$