Communications of the Association for Information Systems

Volume 34 Article 53

2-2014

Intelligent Knowledge Beyond Data Mining: Influences of Habitual Domains

Xiaodan Yu
School of Information Technology and Management, University of International Business and Economics (UIBE), Beijing, China

Yong Shi
CAS Research Center on Fictitious Economy and Data Science, UCAS, China, yshi@ucas.ac.cn

Lingling Zhang
School of Management, UCAS, China

Guangli Nie
Guanghua School of Management, Peking University, China

Anqiang Huang
School of Economics and Management, Beihang University, China

Communications of the Association for Information Systems

CAIS

Intelligent Knowledge Beyond Data Mining: Influences of Habitual Domains

Xiaodan Yu
School of Information Technology and Management, University of International Business and Economics (UIBE), Beijing, China
CAS Research Center on Fictitious Economy and Data Science, University of Chinese Academy of Sciences (UCAS), Beijing, China

Yong Shi
CAS Research Center on Fictitious Economy and Data Science, UCAS, China, yshi@ucas.ac.cn
University of Nebraska at Omaha, USA, yshi@ucas.ac.cn

Lingling Zhang
School of Management, UCAS, China
CAS Research Center on Fictitious Economy and Data Science, China

Guangli Nie
Guanghua School of Management, Peking University, China
Postdoctoral Program of Agricultural Bank of China

Anqiang Huang
School of Economics and Management, Beihang University, China

Abstract:

Data mining is a useful analytic method and has been increasingly used by organizations to gain insights from large-scale data. Prior studies of data mining have focused on developing automatic data mining models that belong to first-order data mining. Recently, researchers have called for more study of the second-order data mining process. Second-order data mining process is an important step to convert data mining results into intelligent knowledge, i.e., actionable knowledge. Specifically, second-order data mining refers to the post-stage of data mining projects in which humans collectively make judgments on data mining models' performance. Understanding the second-order data mining process is valuable in addressing how data mining can be used best by organizations in order to achieve competitive advantages. Drawing on the theory of habitual domains, this study developed a conceptual model for understanding the impact of human cognition characteristics on second-order data mining. Results from a field survey study showed significant correlations between habitual domain characteristics, such as educational level and prior experience with data mining, and human judgments on classifiers' performance.

Keywords: data mining, second-order data mining, collaborative intelligence, habitual domains, knowledge management

Editor's Note: The article was handled by the Department Editors for Information Technology and Systems

Volume 34, Article 53, pp. 985-1000, February 2014

I. INTRODUCTION

Data mining is an analytic technique, which automatically extracts novel and interesting rules or patterns from large-scale data by using data mining models [Han, Kamber, and Pei, 2011; Olson and Shi, 2005, p. 39; Tan, Steinbach, and Kumar, 2005; Shi et al., 2011]. The popularity of commercial data mining software expedites the process of using data mining as an important business intelligent tool in organizations to gain a competitive advantage. Although data mining was often considered to be a computer science discipline, in the past two decades, data mining (sometimes referred to as knowledge discovery from database), has gained increased attention in the information systems (IS) field. Instead of examining the complex mathematical data mining models, IS researchers focused on the topics of data mining implementation, data mining outcome evaluation, strategic use of data mining, and decision making related to data mining [Alavi and Leidner, 2001; Apte, Liu, Pednault and Smyth, 2002; Bendoly, 2003; Bose and Mahapatra, 2001; Jourdan, Rainer and Marshall, 2008; Overby, Bharadwaj and Sambamurthy, 2006]. Prior studies on data mining have focused mainly on examining first-order data mining, while little attention has been paid to studying the second-order data mining [Zhang, Li, Shi and Liu, 2009]. Researchers call for more study on the manner through which data mining assists better decision making [Brydon and Gemino, 2008; Jackson, 2002].

First-order data mining includes activities from developing data mining models to running these data mining models on data sets. The primary outcome of first-order data mining is the identification of rules or patterns. During the second-order data mining, domain experts and data mining experts collaboratively make judgments on data mining models' performance by following a set of explicit and/or implicit evaluation criteria. Consequently, human knowledge is incorporated with data mining results, and intelligent knowledge (i.e., a type of useful, actionable knowledge) is discovered [Fayyad, Piatetsky-Shapiro and Smyth, 1996a; Zhang et al., 2009]. The capability of converting data mining results into intelligent knowledge is critical to achieving specific data mining goals for organizations, such as increasing corporate performance, optimizing customer relationships, monitoring business activities, and supporting decision making [Negash, 2004].

Thus the overall goal of this article is to enhance our understanding of second-order data mining. In particular, 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 is discovered [Baker, Burkman and Jones, 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.[1] A field survey was administrated at a multidisciplinary research site. 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 of 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. Costly mistakes can be avoided when appropriate people are selected.

[1] Refer to a set of data mining classification models that are used to predict the target class for each case in the database.

The rest of the article is organized as follows. Section II introduces the theoretical bases by reviewing relevant literature in data mining, knowledge management, and the theory of habitual domains respectively. Hypotheses are developed at the end of Section II. Then the overall research design and experimental results are presented in Section III. Section IV discusses the limitations of the study. In Sections IV and V, we present the discussion and conclusion of our study. Future research suggestions are included.

II. CONCEPTUAL FOUNDATION

The conceptual findings for this research draw on theories from three areas: including data mining, knowledge management, and cognitive psychology. A detailed discussion follows.

Data Mining

Data mining refers to the "application of specific algorithms for extracting patterns from data" [Fayyad et al., 1996a, p. 39]. Similar to statistics, data mining has two primary functions: prediction and description. Commonly used data mining methods for prediction include regression and classification. Clustering and association analysis are two data mining methods for data description [Fayyad et al., 1996a]. Research in data mining is mostly application-driven. To date, data mining has received successfully applications in various areas, such as astronomy (i.e., image classification), marketing (i.e., customer segmentation), manufacturing (i.e., faults clustering), and Internet security (i.e., intrusion detection). The leading methodology used in data mining projects is the CRISP-DM process model [Shearer, 2000], which stands for *Cross Industry Standard Process for Data Mining*. The CRISP-DM model has six major phases (as is shown in Figure 1).

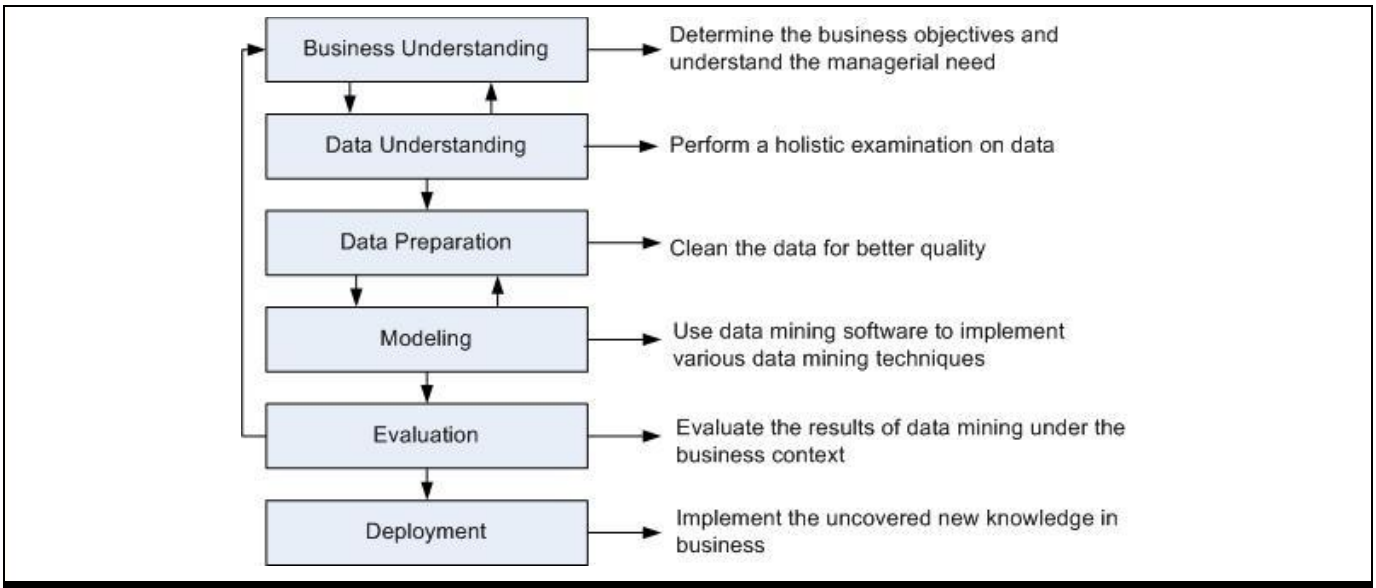

Figure 1. CRISP-DM Model [adapted from Shearer 2000]

The goal of data mining projects is to identify previously unknown, potentially useful, and easily understandable patterns in data [Fayyad, Piatetsky-Shapiro and Smyth, 1996b]. The interaction between humans and the data is highlighted in the CRISP-DM model. For example, during the data understanding stage, data mining analysts communicate with people who have the domain knowledge to find the meanings of the variables. In the evaluation stage, the results automatically generated by the data mining methods are consulted with the domain experts to ensure the validity and usefulness of the results. A data mining project's success requires a smooth communication among developers, users, and data mining models.

Prior studies in data mining were focused mostly on developing or improving data mining models. While Zhang et al. [2009] proposed that these studies on data mining models should be regarded as the "first-order" analytic process, and that the process of discovering intelligent knowledge from data mining is the "second-order" analytic process. Specifically, Zhang et al. [2009, p. 39] described the "first-order" data mining process as the process to *"find some existing phenomenological associations among specific data."* They proposed that the "second-order" analytic process *should translate the findings from the "first-order" analytic process—the rough knowledge—into the intelligent knowledge—a type of problem-solving knowledge*. Figure 2 presents the relationships between data mining and intelligent knowledge.

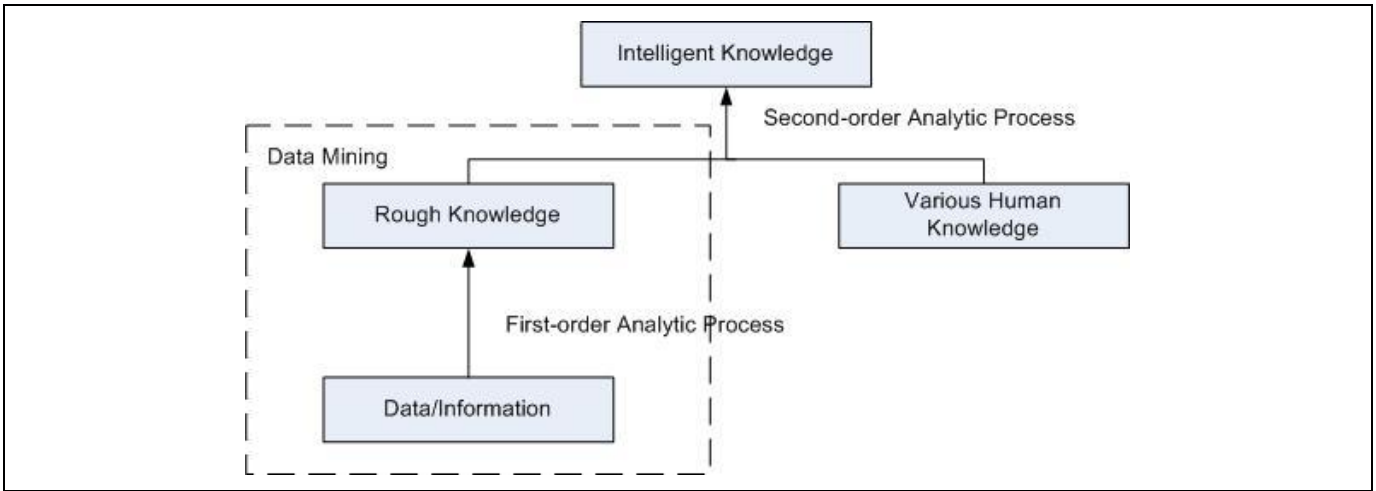

Figure 2. Data Mining to Intelligent Knowledge [adapted from Zhang et al., 2009]

In summary, data mining should enable an organization to (1) amass information stored in large-scale data, (2) identify hidden-patterns through applying various techniques, (3) generate meaningful knowledge (so-called intelligent knowledge) and ultimately (4) to use the intelligent knowledge to achieve a better performance and to meet the strategic objectives. Essentially, the capability of creating intelligent knowledge determines the success of a data mining project.

Rough Knowledge, Various Human Knowledge, and Intelligent Knowledge

Literature on knowledge management provides an in-depth understanding on rough knowledge, various human knowledge, intelligent knowledge (three concepts shown in Figure 2), and the relationships among them.

Consistent with previous literatures, we characterize knowledge into two major categories, namely, explicit knowledge and implicit (or tacit) knowledge [Polanyi, 1966]. Explicit knowledge refers to the codified knowledge that can be easily observed and transferred to the others. Implicit knowledge is the knowledge that has not been documented [Martin and Salomon, 2003]. Examples of explicit knowledge include business processes, manuals, procedures, and various documents. Implicit knowledge is harder to communicate than is explicit knowledge because implicit knowledge resides within a human's brain in an abstract form. Further, implicit knowledge is considered to be more valuable than explicit knowledge [Polanyi, 1966]. However, the creation of implicit knowledge usually involves a time-consuming process and requires context-specific experience. Since implicit knowledge cannot be written down, communicating implicit knowledge needs effective personal contact and trust [Collins, 2001].

Based on the taxonomy of explicit and implicit knowledge, we consider rough knowledge, which is the result of first-order data mining, to be a type of explicit knowledge. Various forms of human knowledge represent a type of implicit knowledge. Intelligent knowledge is a type of explicit knowledge, given the understanding that intelligent knowledge is actionable knowledge.

Prior studies have explained the knowledge creation process through a theoretical model, the knowledge creation model [Alavi and Leidner, 2001]. According to that model, the second-order data mining process can be viewed as a knowledge externalization process when tacit knowledge is converted into explicit knowledge. Specifically, the interactions among individuals and computers are critical to such a knowledge externalization process [Nonaka, Reinmoeller and Senoo, 1998].

Intelligent knowledge is created for certain purposes. The process of creating intelligent knowledge is a complex multi-criteria decision making process by humans. Discovering intelligent knowledge requires smooth interaction between humans and computers. Given a problem domain, humans use their domain knowledge and other specification knowledge to make judgments about the results from data mining classifiers.

Theory of Habitual Domain

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 humans' behavior

mechanisms that guide 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 the individual'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 the daily problems we encounter happen regularly, and (3) humans tend to take the most convenient way of dealing with daily problems [Yu, 1990]. In this article, 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] explained the habitual domains in this way: *"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, humans 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 that reside 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 of 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 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 among these three characteristics and the theory of habitual domains are explained in the next subsection. Hypotheses are developed.

III. HABITUAL DOMAINS THEORY FURTHER EXPLORATION AND HYPOTHESES DEVELOPMENT

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 the *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, all other things being 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 article proposes that an individual's habitual domain's characteristics can be described by examining that person'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, people's habitual domains' characteristics can be described by assessing their background in each of the approaches by which they form the habitual domains.

Considering the purpose of this study along with the consideration of empirical assessment, this article focused on the active learning dimension. The habitual domain is a multidimensional 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] suggests that a study could focus on only one component, based on the study's purpose. Given the fact that the purpose of the study is to understand why people make different judgments on classifiers' performance on data sets and that people's decision making is to a large extent influenced by their learning experiences, it is adequate only to check the active learning experience of people at this point. More approaches should be considered when different goals of the study are considered.

Active learning emphasizes 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 also will enable us to more efficiently integrate previous ideas and make those ideas more accessible.

In this article, we specifically identify 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 significantly high proportion of one's active learning experience. People who have similar backgrounds 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 describe each of these three areas in detail and develop hypotheses.

First, *level of education* is concerned with how many years of formal school education one has. From many years of education, each of us has been exposed to many new ideas and new knowledge through books, lectures, and interactions with classmates. Attending classes not only provides us with new ideas and knowledge but also facilitates the absorption of these new ideas and knowledge in our minds by repetition. In an experimental study, Macpherson [1996] found that educational background, specifically the number of years of education, has a significantly positive effect on individuals' ability to generate insights. Another study reveals that education can decrease the anxiety toward the use of computers [Igbaria and Parsuraman, 1989]. Bower and Hilgard's study [1981] suggests that a higher level of education enhances an individual's cognitive capabilities and, thus, accelerates the individual's learning process, especially in novel situations. Considering the situations people face regarding 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 the performance of classifiers for a particular database.*

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

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 that did not consider domain knowledge in predicting the risk of mortality in patients with a specific disease. In a study that applied data mining to bank loan problems, Sinha and Zhao [2008] examined and compared the 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 the 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 literature related to data mining. We suggest that an individual's experience with data mining greatly influences his/her attitude toward various data mining classifiers. Empirical studies have found that previous experiences with certain technologies can either hinder or foster one's adoption of a new technology [Harrison and Rainer, 1992]. For example, one study found that users resisted using an unfamiliar technology because of the cost of switching [Scholtz and Wiedenbeck, 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 the performance of classifiers for a particular database.

The research model is shown in Figure 3. Building on the theory of habitual domains, the conceptual model describes the conclusion that 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 the influence of habitual domain characteristics on the convergence of human judgments on classifiers' performance.

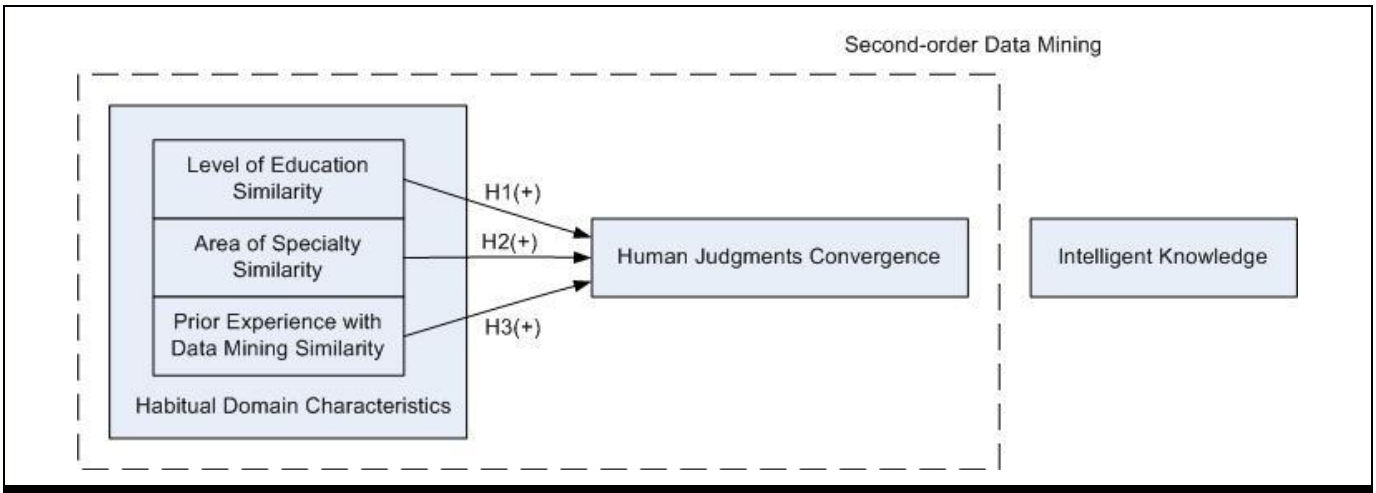

Figure 3. Influence of Habitual Domains on Human Judgment Convergence

IV. RESEARCH METHOD

The overall research design is a field survey[2]. A pilot study was conducted to test the reliability and validity of the survey and the field procedure.

Participants and Data Collection

Considering the purpose of the study is to test whether 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 to green energy to data mining. Researchers in the institute have backgrounds as varied as management information systems, computer science, economics, and biology. Of the thirty-eight respondents, 42 percent were male and 58 percent were female. The distribution of respondents' age is shown in Table 1.

Table 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

[2] This present research is a part of an ongoing program in which we study the effect of human cognitive psychology characteristics on intelligent knowledge discovery from data mining.

In the study, we first ran eight classifiers[3] 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 four hours. One author of the article 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 for several years in the 1980s when there was excessive enrollment in these schools in Ljubljana, Slovenia, and the rejected applications frequently needed an objective explanation.[4] PBC Dataset is a data set related to credit scoring from China. After preprocess, we received a data set with 1600 samples; 800 of them were classified as good customers and 800 were classified as bad customers. Eighty variables were designed to reflect the behaviors of the customers.

Measures

Habitual Domains Characteristics

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

Dependent Variables

Dependent variables in this study were participants' judgments on data mining classifiers' performance. Specifically, dependent variables consist of participants' 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 the 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).

Data Analysis and Results

Descriptive Analysis

We first analyzed the psychometric properties of the acquaintance with data mining (IV3) by running a reliability analysis in SPSS. Results showed that the subscales of IV3 have good internal consistency, $\alpha = 0.93$. Table 2 shows the frequency of individuals' educational background.

Table 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

The descriptive statistic of the areas of specialty of individuals is shown in Table 3.

Table 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

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

[4] http://archive.ics.uci.edu/ml/datasets/Nursery

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

The descriptive analysis of the subjects' judgments on the eight classifiers' performance on the 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 4 shows the descriptive statistics.

Table 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

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 5 presented the descriptive statistics for all classifiers' scores on the PBC database.

Table 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

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[5] 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 the question: *Is there a tendency for actors who have more similar attributes to be located closer to one another in network?* Since the hypotheses of the present study is concerned with whether the closeness of experts' habitual domain characteristics would affect their judgments on data mining algorithms' performance, 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 an individual question; rather, it provides a global view of the subject's responses to all the questions. Second, it simplifies the dependent variables and makes it easy to conduct the correlation analysis.

It should be noted that although the MANOVA method allows the analysis of the effects of more than one independent variable on two or more dependent variables, the 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 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[6] in

[5] Values less than 1.0 indicate a positive association (somewhat confusingly), values greater than 1.0 indicate a negative association.

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

UCINET [Borgatti et al., 2002] to get an adjacency matrix[7] of all participants based on their judgments on data mining algorithm performance. This adjacency matrix described the "closeness" of each pair of participants on their overall perceptions on the data mining algorithm performance. Then we created another attribute table that contains all information of participants' habitual domain characteristics. UCINET was used to calculate the Geary's C measure. Tables 6 and 7 present the Geary's C statistic results.

Table 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	.99*
ASS	Closeness between individual's judgments on classifier's performance	1.004
PEDMS	Closeness between individual's judgments on classifier's performance	.98**

Notes: IV = Independent Variable, DV = Dependent Variable, LES = Level of Education Similarity, ASS = Area of Specialty Similarity, PEDMS = Prior Experience with Data Mining Similarity, * Indicates a correlation is significant at 0.1, ** Indicates a correlation is significant at .01.

Table 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	.99*
ASS	Closeness between individual's judgments on classifier's performance	1.005
PEDMS	Closeness between individual's judgments on classifier's performance	.98*

Notes: IV = Independent Variable, DV = Dependent Variable, LES = Level of Education Similarity, ASS = Area of Specialty Similarity, PEDMS = Prior Experience with Data Mining Similarity, * Indicates a correlation is significant at 0.1, ** Indicates a correlation is significant at .01.

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

V. 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 faculty members who worked in this institution. Thus, the generalizability of the respondents' behaviors to a more general population may be somewhat limited. One frequently mentioned comment noted that a drawback of using students as subjects is the significant differences between students and the targeting groups. In this study, the targeting groups would be the data mining customers who propose, sponsor, evaluate, and eventually implement a data mining project. The targeting groups may possess very different backgrounds in terms of education, areas of specialty, and previous experience, compared to students of this study.

Additionally, this study asks participants' opinions only on classifiers' performance on two data sets. Moreover, a data set is from UCI rather than a real-world data set. One major criticism of the UCI data set is that the data set in UCI is often biased because of preprocession of 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 detect only the association between subjects' attributes and subjects' responses on a set of questions. The impact of interactions among the subjects' attributes, such as level of education, areas of specialty, and prior experience with data mining, cannot be obtained. Future research can acquire a 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 at applying habitual domain theory to understand 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

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

specialty in this study, the study did not specify which areas of specialty should be considered in the assessment of individuals' habitual domains.

VI. 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 deal with transforming data mining results into actionable intelligent knowledge are called "second-order" data mining. This article 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 humans 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 article derived empirically testable hypotheses based on the habitual domain theory. In our experiments,[8] we found support for our hypotheses that people's judgments on data mining classifiers' performance are influenced by their education and prior experience with data mining. Education was found to be an important factor on people's perceptions on classifiers' performance. People's prior experience with data mining was also revealed as a predictor to their evaluation of 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, the results indicated that individuals' judgments on classifiers' performance will not be significantly influenced by the individuals' majors. One possible explanation is that the majors of participants in the study were not diverse enough. This study had only individuals from these three majors: Computer Science, Financial Engineering, and Management Science. It is possible that students from these majors show similar attitudes on data mining classifiers' performance on various data sets. A study conducted by Tikka [2000] found that students with majors related to technology and economics showed similar attitudes 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 on data mining methods is the opportunity for training interventions to manipulate people's perceptions about a classifier. Since education and previous experience with data mining have a significant effect on people's perceptions on classifiers, designing better training will increase the likelihood that novice data mining developers will make quality judgments such 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 the 1980s, numerous data mining algorithms have been developed. But no one data mining algorithm has proved to outperform other algorithms in all tasks. Therefore, in the real-world data mining projects, data mining teams have to compare more than one data mining method carefully and choose one that has the best functioning performance. Depending on their 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 reach a convergent opinion on which data mining method to use. But having people with similar habitual domains may also place a potential risk for the data mining project team of entering a decision trap. For instance, it is possible that all involved 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 an 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.

VII. CONCLUSIONS AND FUTURE RESEARCH

The broad goal of this article is to enhance our understanding about the second-order data mining, particularly the creation of intelligent knowledge by humans 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

[8] Each data set is considered as a separate experiment. Classifiers' performance on two data sets is independent.

Communications of the Association for Information Systems

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's 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 a person's previous experience with data mining and the person's judgments on data mining performance.

This article 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 article provide evidence for the variations of human judgments on classifiers' performance when people possess varied cognitive psychology characteristics. These findings are valuable in understanding the important role of humans in the stage of second-order data mining. Most present studies of data mining either ignore the role of people or symbolize people as agents in the post-stage of data mining. It could be argued from this study that 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 prior knowledge, rough knowledge, and habitual domain characteristics.

This research presents interesting directions for future research. Since there is no one data mining method that outperforms all the other data mining methods in all kinds of tasks, choosing a most appropriate data mining method for a given task is an important step that influences 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 if the implicit knowledge of data mining experts can have linkages with their past experience and educational background. Understanding what type of experiences and educational backgrounds are generally found 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 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 to understanding the overall data mining project success. As is the case with other types of projects, a data mining project that is accepted and actually used by the end users is a truly 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 what is lacking in the domain knowledge, it is challenging for data mining analysts to understand what ideas are innovative from the customers' perspective. Understanding the preferences of customers and being able to share an understanding with customers about what ideas are innovative is of critical importance to the overall success of a data mining project. The habitual domains theory not only conceptually describes how people obtain, store, process, and apply information from the world of concepts and propositions, but also prescribes ways to expand habitual domains and discuss the characteristics of information that would catch people's eyes. The theory of habitual domains possesses great potential for developing useful constructs in order to predict the acceptance and continuing usage of data mining.

ACKNOWLEDGMENTS

We gratefully thank the editor, associate editor, and the two anonymous reviewers for their constructive feedback. We appreciated Yibing Chen and Quan Chen's assistance in handling feedback of the survey. We would also like to thank those who have participated in the survey. This work was partially supported by National Natural Science Foundation of China (Grant No. 71331005, 71110107026, and 71201143), the CAS/SAFEA International Partnership Program for Creative Research Teams, and the President's Fund of the University of Chinese Academy of Sciences (GUCAS)(A) (Grant No.085102HN00).

REFERENCES

Alavi, M. and D. Leidner (2001) "Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issue", *MIS Quarterly,* (25)1, pp. 107–135.

Ambrosino, R. and B.G. Buchanan (1999) "The Use of Physician Domain Knowledge to Improve the Learning of Rule-based Models for Decision-support", *Proceedings of the AMIA Symposium,* Washington, D.C.

Apte, C., B. Liu, E.P.D. Pednault and P. Smyth (2002) "Business Applications of Data Mining", *Communications of the ACM,* (45)8, pp. 49–53.

Astin, A.W. (1993) *What Matters in College? Four Critical Years Revisited,* San Francisco, CA: Jossey-Bass.

Baker, J., J. Burkman and D.R. Jones (2009) “Using Visual Representations of Data to Enhance Sensemaking in Data Exploration Tasks”, *Journal of the Association for Information Systems,* (10)7, pp. 533–559.

Bendoly, E. (2003) “Theory and Support for Process Frameworks of Knowledge Discovery and Data Mining from ERP Systems”, *Information & Management,* (40)7, pp. 630–647.

Biglan, A. (1973) "The Characteristics of Subject Matter in Different Academic Areas", *Journal of Applied Psychology* (57:3), pp 195-203.

Borgatti, S.P., M.G. Everett and L.C. Freeman (2002) *Ucinet for Windows: Software for Social Network Analysis,* Cambridge, MA: Harvard.

Bose, I. and R.K. Mahapatra (2001) “Business Data Mining: A Machine Learning Perspective”, *Information & Management,* (39)3, pp. 211–225.

Bower, G. and E. Hilgard (1981) *Theories of Learning*, Englewood Cliffs, NJ: Prentice Hall.

Brydon, M. and A. Gemino (2008) “You’ve Data Mined: Now What?”, *Communications of the Association for Information Systems,* (22) Article 33, pp. 603–616.

Chen, H., R.H.L. Chiang and V.C. Storey (2012) "Business Intelligence and Analytics: From Big Data to Big Impact", *MIS Quarterly,* (36)4, pp.1165-1188.

Collins, H.M. (2001) “Tacit Knowledge, Trust and the Q of Sapphire”, *Social Studies of Science,* (31)1, pp. 71–85.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996a) “From Data Mining to Knowledge Discovery in Database”, *AI Magazine,* (17)3, pp. 37–54.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996b) “The KDD Process for Extracting Useful Knowledge from Volumes of Data”, *Communication of the ACM,* (39)11, pp. 27–34.

Geary, R.C. (1954) “The Contiguity Ratio and Statistical Mapping”, *The Incorporated Statistician,* (5)3, pp. 115–127, 129–146.

Han, J., M. Kamber and J. Pei (2011) *Data Mining: Concepts and Techniques*, *3rd edition,* Burlington, MA: Morgan Kaufmann.

Harrison, A.W. and R.K. Rainer, Jr. (1992) “The Influence of Individual Differences on Skill in End-user Computing”, *Journal of Management Information Systems,* (9)1, pp. 93–111.

Igbaria, M. and S. Parsuraman (1989) “A Path Analytic Study of Individual Characteristics, Computer Anxiety, and Attitudes Toward Microcomputers”, *Journal of Management,* (15)3, pp. 373–388.

Jackson, J. (2002) “Data Mining; A Conceptual Overview”, *Communications of the Association for Information Systems,* (8) Article 19, pp. 267–296.

Jourdan, Z., R.K. Rainer and T.E. Marshall (2008) “Business Intelligence: An Analysis of the Literature”, *Information Systems Management,* (25)2, pp. 121–131.

Macpherson R, B. Jerrom and H. A. Hughes (1996) "Relationship Between Insight, Educational Background and Cognition in Schizophrenia," *The British Journal of Psychiatry,* (168)6, pp. 718-722.

Martin, X. and R. Salomon (2003) “Tacitness, Learning, and International Expansion: A Study of Foreign Direct Investment in a Knowledge-intensive Industry”, *Organization Science,* (14)3, pp. 297–311.

Negash, S. (2004) “Business Intelligence”, *Communications of the Association for Information Systems,* (13) Article 15, pp. 177–195.

Nonaka, I., P. Reinmoeller and D. Senoo (1998) “The ‘ART’ of Knowledge: Systems to Capitalize on Market Knowledge”, *European Management Journal,* (16)6, pp. 673–684.

Olson, D. and Y. Shi (2005) *Introduction to Business Data Mining,* New York. NY: McGraw-Hill/Irwin.

Overby, E., A. Bharadwaj and V. Sambamurthy (2006) “Enterprise Agility and the Enabling Role of Information Technology”, *European Journal of Information Systems,* (15)2, pp. 120–131.

Paulsen, M.B., and C.T. Wells (1998) "Domain Differences in the Epistemological Beliefs of College Students", *Research in Higher Education,* (39)4, pp. 365-384.

Polanyi, M. (1966) *The Tacit Dimension*, Garden City, NY: Doubleday and Co.

Scholtz, J. and S. Wiedenbeck (1990) “Learning Second and Subsequent Programing Languages: A Problem of Transfer”, *International Journal of Human-Computer Interaction,* (2)1, pp. 51–72.

Shearer, C. (2000) “The CRISP-DM Model: The New Blueprint for Data Mining”, *Journal of Data Warehousing,* (5)4, pp. 13–22.

Shi, Y. and P.L. Yu (1987) "Habitual Domain Analysis for Effective Decision Making", *Asia-Pacific Journal of Operational Research,* (4)2, pp. 131-150.

Shi, Y., Y.J. Tian, G. Kou, Y. Peng, and J.P. Li (2011) *Optimization-based Data Mining: Theory and Applications,* New York, Springer.

Sinha, A.P. and H. Zhao (2008) “Incorporating Domain Knowledge into Data Mining Classifiers: An Application in Indirect Lending”, *Decision Support Systems,* (46)1, pp. 287–299.

Tan, P.-N., M. Steinbach and V. Kumar (2005) *Introduction to Data Mining,* Reading, MA: Addison Wesley.

Tikka, P. M., M.T. Kuitunen, and S.M. Tynys (2000) "Effects of Educational Background on Students' Attitudes, Activity Levels, and Knowledge Concerning the Environment", *The Journal of Environmental Education,* (31)3, pp. 12-19.

Wasserman, S. and K. Faust (1995) *Social Network Analysis*, New York, NY: Cambridge University Press.

Yu, P.-L. (1980) “Behavior Bases and Habitual Domains of Human Decision/Behavior-concepts and Applications” in Fandel, G. and T. Gal (eds.) *Multiple Criteria Decision-making, Theory and Applications,* New York, NY: Springer-Verlag, pp. 511–539.

Yu, P.L. (1990) *Forming Winning Strategies: An Integrated Theory of Habitual Domains,* Berlin, Germany: Springer-Verlag, p. 392.

Yu, P.L. (1991) "Habitual Domains", *Operations Research* (39)6, pp. 869-876.

Yu, P.L. (2002) *Habitual Domains and Forming Winning Strategies,* Hsinchu, Taiwan: NCTU Press.

Yu, P.-L. and Y.-C. Chen (eds.) (2010) *Dynamic MCDM, Habitual Domains and Competence Set Analysis for Effective Decision Making in Changeable Spaces,* New York, NY: Springer.

Zhang, L., J. Li, Y. Shi and X. Liu (2009) “Foundations of Intelligent Knowledge Management”, *Human Systems Management* (28), pp. 145–161.

APPENDIX A: SUMMARY OF DATA SETS, CLASSIFIERS, AND MEASURES

Table A–1: 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: Questionnaire Used for the Nursery Database

Score of algorithm													
Measure		J48		Nbtree		Baysnet		Naivebays		logistic		SVM		MCLP		MCQP
Correctly		0.97		0.97		0.90		0.90		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	1.00		1.00		1.00		1.00		1.00		1.00		0.98		0.99
	FP rate	0.00		0.00		0.00		0.00		0.00		0.00		0.00		0.04
	F-Measure	1.00		1.00		1.00		1.00		1.00		1.00		0.98		0.96
recommend	TP rate	0.00		0.00		0.00		0.00		0.00		0.00		1.00		0.96
	FP rate	0.00		0.00		0.00		0.00		0.00		0.00		0.02		0.01
	F-Measure	0.00		0.00		0.00		0.00		0.00		0.00		0.99		0.98
priority	TP rate	0.95		0.96		0.90		0.90		0.89		0.98				
	FP rate	0.02		0.02		0.10		0.10		0.06		0.01				
	F-Measure	0.96		0.96		0.86		0.86		0.89		0.98				
very_recom	TP rate	0.73		0.70		0.06		0.06		0.74		0.90				
	FP rate	0.01		0.00		0.00		0.00		0.01		0.00				
	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		0.90		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		0.90		0.98				

Table B–2: Questionnaire Used for the PBC Database

Score of algorithm													
Measure		J48		Nbtree		Baysnet		Naivebays		logistic		SVM		MCLP		MCQP
Correctly		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 C–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 attribute, selects the lowest scores, and then sums. For example, for subjects x and y, 3 + 3 + 5 = 11, might mean the extent to which subjects 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: 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, calculate the Geary's C attribute A.

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

$$N = 3,\ X_1 = 3,\ X_2 = 3,\ X_3 = 3,\ w_{12} = 11,\ w_{13} = 12,\ w_{23} = 13$$

$$C_A = \frac{(3-1)*2*[11(3-5)^2 + 12(3-4)^2 + 13(5-4)^2]}{2*107*[(3-4)^2 + (5-4)^2 + (4-4)^2]} = .65$$

ABOUT THE AUTHORS

Xiaodan Yu is an Assistant Professor of Information Systems in the School of Information Technology and Management at the University of International Business and Economics (UIBE) in Beijing, China. She is also currently an adjunct researcher in the Research Center on Fictitious Economy and Data Science at the University of Chinese Academy of Sciences in Beijing, China. Her research interests include IT use, business intelligence, virtual team, and agile software development. She earned her PhD in Information Technology from University of Nebraska at Omaha and an MS in Management Science and Engineering from Beijing Institute of Technology.

Yong Shi, Senior Member of IEEE, has served as the Executive Deputy Director of Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, China, since 2007. He has been the Union Pacific Chair of Information Technology, College of Information Science and Technology, Peter Kiewit Institute, University of Nebraska at Omaha. Shi's research interests include business intelligence, data mining, multiple criteria decision making, and their applications in business and management. He has published more than twenty books, over 200 papers in various journals, and numerous conferences/proceedings papers. Shi has received many distinguished awards, including the Georg Cantor Award of the International Society on Multiple Criteria Decision Making (MCDM), 2009, and Fudan Prize of Distinguished Contribution in Management, Fudan Premium Fund of Management, China, 2009.

Lingling Zhang currently is Associate Professor at School of Management, University of Chinese Academy of Sciences. She also works as a researcher at Research Centre on Fictitious Economy & Data Science, Chinese Academy of Sciences. She was once visiting scholar of Stanford University in the USA. Her research interests include data mining, intelligent knowledge management, and management information system. She has received two grant supported by the Natural Science Foundation of China (NSFC), published four books, more than forty papers in various journals.

Guangli Nie received his BS in Information Management from Shandong University in 2002 and a PhD in Management Science from the University of Chinese Academy of Sciences and is doing research for the Agricultural Bank of China. He has published more than twenty papers in various journals and conferences, including *Expert Systems with Applications*. His research interests include data mining and credit portfolio.

Anquan Huang received his Master's degree from the University of Chinese Academy of Sciences, Beijing, China, in 2010. Currently, he is a PhD candidate at Beihang University. His research interests include intelligent knowledge and data mining, economy forecast theory and method, and knowledge management.

Communications of the Association for Information Systems

ISSN: 1529-3181

Communications of the Association for Information Systems