Decision Support Systems 51 (2011) 141–154

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models

Johan Huysmans [a], Karel Dejaeger [a], Christophe Mues [b], Jan Vanthienen [a], Bart Baesens [a,b,c,*]

[a] Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
[b] School of Management, University of Southampton, Southampton, SO17 1BJ, United Kingdom
[c] Vlerick Leuven Gent Management School, Leuven, Belgium

ARTICLE INFO

Article history:
Received 26 November 2008
Received in revised form 29 October 2010
Accepted 5 December 2010
Available online 10 December 2010

Keywords:
Data mining
Classification
Knowledge representation
Comprehensibility
Decision tables

ABSTRACT

An important objective of data mining is the development of predictive models. Based on a number of observations, a model is constructed that allows the analysts to provide classifications or predictions for new observations. Currently, most research focuses on improving the accuracy or precision of these models and comparatively little research has been undertaken to increase their comprehensibility to the analyst or end-user. This is mainly due to the subjective nature of 'comprehensibility', which depends on many factors outside the model, such as the user's experience and his/her prior knowledge. Despite this influence of the observer, some representation formats are generally considered to be more easily interpretable than others. In this paper, an empirical study is presented which investigates the suitability of a number of alternative representation formats for classification when interpretability is a key requirement. The formats under consideration are decision tables, (binary) decision trees, propositional rules, and oblique rules. An end-user experiment was designed to test the accuracy, response time, and answer confidence for a set of problem-solving tasks involving the former representations. Analysis of the results reveals that decision tables perform significantly better on all three criteria, while post-test voting also reveals a clear preference of users for decision tables in terms of ease of use.

1. Introduction

Predictive models are widely used in both research and business applications. For example, based on past applications, financial institutions construct credit scoring models to predict whether an applicant for a loan will be able to pay back the loan or will default on his/her loan obligations. The model is then used to decide which new credit applications should be granted or denied [2]. Similarly, insurance companies develop predictive models to identify the claims that are likely to be fraudulent [60], and in the medical sciences predictive techniques may be used to decide whether a particular medical condition is malign or benign.

In some of these applications, selection of the best predictive model is based solely on the ability to provide correct predictions for previously unseen examples. In other situations though, the interpretability of the model is equally important, i.e. one must be able to understand how the model reaches a particular decision. For example, in the domain of credit scoring, financial institutions often face the legal obligation of being able to motivate why a certain customer was denied credit [17]. Also, in the medical sciences, understanding how the model comes to its conclusions is often crucial, as it provides information about the variables that influence the disease and can therefore point to a potential cure or prevention strategy. An important criterion for selecting a model from a series of candidate models with similar performance is that it is in line with previous domain knowledge (see e.g. [45] for an overview of the literature on incorporating domain knowledge into data mining). This is especially true in many business settings, in particular where models support decision making or form the basis of policy development, and when there is a risk that sampling bias might cause the introduction of counter-intuitive relationships (e.g. [18]). Hence, there is often a trade-off between the predictive accuracy of powerful but essentially black box models such as neural networks or support vector machines and the good interpretability of other types of representations that may facilitate validation by an analyst or domain expert.

For these reasons, rule induction algorithms, which return a set of 'if–then' rules, or decision tree learners are often the preferred choice as they should offer the required level of interpretability. Alternatively, rule extraction [1,24,33] can be performed on black box models, such as neural networks and support vector machines, to extract a set of rules that approximate the black box as closely as

* Corresponding author. Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium. Tel.: +32 16 32 68 84; fax: +32 16 32 66 24.
E-mail addresses: Karel.Dejaeger@econ.kuleuven.be (K. Dejaeger), C.Mues@soton.ac.uk (C. Mues), Jan.Vanthienen@econ.kuleuven.be (J. Vanthienen), Bart.Baesens@econ.kuleuven.be (B. Baesens).

0167-9236/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2010.12.003

possible and at the same time provide a more understandable representation to the users. However, previous research concerning rule extraction techniques [25] indicated that some algorithms return models that closely approximate the underlying black box model, but at the cost of being very complex. For example, as reported in [27], the algorithms G-REX and CART return trees with an average of respectively 22 and 91 nodes when applied to some real-life data sets. Zhou et al. [63] applied the REFNE rule extraction algorithm and found the average number of rules to be 31.[1] It can be feared that such models might fail in their primary task of providing insight in the black box model from which they are extracted, as the black box is only replaced by a myriad of rules.

It was observed by Pazzani [35] that few papers actually aim to empirically assess comprehensibility beyond simply reporting the size of the resulting representations. Moreover, Pazzani notes that there is little understanding of the factors influencing comprehensibility (with even the effect of size remaining unexplored), and that there have been no attempts to show e.g. that users actually prefer certain visualizations over mere textual representations. Hence, he argues that data mining can benefit considerably from the interaction with cognitive science to increase the usefulness of knowledge discovery and the user's acceptance of the models obtained. Similarly, Freitas [19] reviews some recent concepts and approaches for discovering not just accurate but also comprehensible and/or 'interesting' (i.e. novel or surprising) knowledge from data. However, he echoes Pazzani's observation that there is no agreement on which of these representation types (e.g. if–then rules, decision trees, etc.) is the most comprehensible in general, and that there seems to be no study actually comparing their comprehensibility from the point of view of human users.

In this paper, we will present the results of an experiment that compares the impact of several representation formats on the aspect of comprehensibility. More specifically, an end-user experiment was set up to test the accuracy, response time, and answer confidence for a set of problem-solving tasks involving these representations. The experiment is run in a credit scoring context, which involves the use of predictive models for the task of assessing whether a loan applicant is likely to pay back or default on this loan and should therefore be accepted or rejected. The following formats are considered: decision tables, (binary) decision trees, propositional rules and oblique rules. Besides comparing these different representation formats, our study also investigates the influence of the size of each of these representations on their interpretability. A better understanding of this relation would provide an indication whether a model could be useful in practice as an explanation or validation aid, or should be avoided because it is too confusing for end-users or analysts and has little added value over a black box model.

In the next section, we describe in detail the four representation formats covered in the experiment. In Section 3, the theory underpinning the experiment is discussed and based on this theory, a series of research propositions are formulated. Section 4 discusses the empirical setup of the experiment while Section 5 presents the findings. Section 6 provides a discussion of the obtained results. Finally, the paper concludes with the key findings and interesting topics for future research.

2. Rule representation

In this section, an overview of the representation formats selected is provided. Given the wide range of representation schemes proposed in the literature, and their numerous variations, inevitably a selection of representation formats was required. Based on the ease of use for novice users, inclusion in prior studies, and prominence in the machine learning/data mining literature, we opted for the following representation formats.

Firstly, the most common type of rules is without any doubt *propositional if–then rules*. The condition part of a propositional rule consists of a combination of conditions on the input variables. While the condition part can contain conjunctions, disjunctions, and negations, most algorithms will return rules that only contain conjunctions.

Most algorithms will ensure that the condition parts of each rule demarcate separate areas in the input space: i.e., the rules are mutually exclusive. Therefore, only one rule is satisfied when a new observation is presented and that rule will be the only one used for making the classification decision. Other algorithms allow multiple rules to fire for the same observation. This requires an additional mechanism to combine the predictions of individual rules, such as assigning a confidence factor to each rule [9] or sorting the rules and allowing only the first firing rule to decide [50]. For this paper, it is assumed that all rules are mutually exclusive and therefore these mechanisms are not required.

Various formats can be used to represent propositional rules. The most straightforward approach is to simply write the rules down, as in the following example:

IF (INCOME > 400 AND GOAL = CAR) THEN ACCEPT
IF (INCOME > 900 AND GOAL = HOUSE) THEN ACCEPT
DEFAULT:REJECT

This fictitious example shows the credit policy of a financial institution which may be used to decide on loan applications. Based on this policy, the credit manager would accept all applications where the applicant has an income above 900 and the goal is the purchase of a house or where the income is above 400 and the goal is the purchase of a car. If these conditions are not satisfied, the default rule specifies that applications are to be rejected.

Other more graphical-oriented representations that are frequently used to depict conditional logic are decision tables and decision trees. A *decision table* [55] is a tabular representation that consists of four quadrants separated by horizontal and vertical double lines (see Fig. 1).

The horizontal line divides the table into a condition part (top) and an action part (bottom), whereas the vertical line separates subjects (left) from entries (right). Every column in the entry part corresponds to a rule, combining condition states with the appropriate action(s) to take. A dash symbol (-) in the condition part of the table indicates that the value is irrelevant in that condition and an "X" in the action part represents the correct conclusion to make if the conditions leading to that column are satisfied.

(a) Single-hit table

INCOME	< 1000		≥ 1000
AGE	< 25	≥ 25	-
ACCEPT	X		
REJECT		X	X

(b) Multiple-hit table

INCOME	≥ 1000	-	< 1000
AGE	-	< 25	< 25
ACCEPT			X
REJECT	X	X	

Fig. 1. Example decision tables.

[1] These techniques are only mentioned because the authors included complexity information in their papers; the above should not be considered as a criticism of these specific algorithms.

The decision table in Fig. 1(a) therefore corresponds to the following rules:

IF (INCOME < 1000 AND AGE < 25) THEN ACCEPT
IF (INCOME < 1000 AND AGE ≥ 25) THEN REJECT
IF (INCOME ≥ 1000) THEN REJECT

which we can rewrite as follows using a default rule:

IF (INCOME < 1000 AND AGE < 25) THEN ACCEPT
DEFAULT:REJECT

In this paper, the format of the decision tables is restricted to single-hit decision tables, i.e. decision tables for which the following criteria are satisfied:

- Completeness: all possible combinations of the condition states are included;
- Exclusivity: no combination is covered by more than one column;
- Lexicographical order: the condition entries within columns are lexicographically ordered, according to which entries at lower rows alternate first.

The fact that each possible combination of condition states occurs only in exactly one column is the key advantage of single hit tables [56]. Decision tables that are constructed according to these principles have proven advantageous with respect to verification and validation. In Fig. 1(b), a multiple-hit classification table is shown which violates the exclusivity criterion since a person with an income lower than 1000 and younger than 25 would satisfy both the second-last and last columns. Here, a particular evaluation scheme, e.g. a first-hit or an all-hits rule, would be used to decide which column(s) must be considered.

Decision trees are a third representation format often used to depict conditional logic (see Fig. 2). A decision tree consists of a number of internal nodes, specifying conditions to be tested, and a number of leaf nodes with a class label. New observations can be classified by traversing the tree from top to bottom where condition tests in the internal nodes indicate whether the left or right branch must be followed.

In the experiment, we restrict ourselves to binary decision trees, i.e. decision trees where each condition test has exactly two outcomes, yes or no, and where the left (right) branch must be followed if the outcome is positive (negative).

It can be shown that all the above representation formats are logically equivalent in the sense that one type of representation can be automatically translated into another (albeit in a simpler or more complex form), while preserving the predictive behavior of the original model [13,55]. Hence, propositional rules, decision tables and decision trees are essentially all propositional representations, in terms of their expressive power.

In this study, a second rule type, *oblique rules*, was also considered. In an oblique rule, linear combinations of attributes are jointly evaluated as rule conditions instead of simple attribute/value pairs. This type of rules was included as it is frequently encountered as the result of rule extraction (e.g. [32,41,42]). Also, various algorithms exist for the construction of oblique trees [7]. Oblique trees are a special class of decision trees in which multiple attributes for each test condition are evaluated instead of one attribute. As with decision trees and propositional rules, oblique trees and rules are logically equivalent.

In an oblique rule, multiple attributes are evaluated jointly in the test conditions. More specifically, oblique rules will create piece-wise discriminant functions, and could thus e.g. take the following form:

IF ($c_1X+c_2Y > c_3$) AND ($c_4X+c_5Z > c_6$) AND … THEN Class=1

with $c_1,\ldots,c_6,\ldots \in \mathbb{R}$. Oblique rules have the advantage that they can create decision boundaries that are non-parallel to the axes of the original input space. This can result in a rule set that consists of fewer rules than would be required by a propositional rule set. Nonetheless, oblique rules are usually considered more difficult to understand than propositional rules as shown by the following example.

IF ($5 \times$ INCOME $+ 12 \times$ AGE > 900) THEN ACCEPT
DEFAULT:REJECT

In this fictitious example, the credit policy of a company is shown in which a credit manager would make a trade-off between income and age. E.g., in the case of an older person, a lower income is accepted thus giving a non-parallel decision boundary as illustrated in Fig. 3.

In summary, the experiment covers three representation formats for propositional rules (textual description, decision tables and decision trees) and one representation format for oblique rules (textual description).

3. Theory development and propositions

Several studies empirically compared different conditional logic representation formats (e.g. [23,39,47,59]). Most of these studies focus on the construction of the models and measure which representation can be created with minimal semantic and syntactical errors by a domain expert. In this study however, it is assumed that the models are automatically generated by running a data mining or machine learning algorithm on a given data set of observations (e.g. C4.5 [37], CART [5] or Ripper [11]). As such, the focus lies on the aspect of comprehensibility by the end-user and his/her ability to perform certain problem solving tasks. Despite considerable literature concerning representation formats, no univocal conclusion has been reached as to which format is the most appropriate in a given situation [26,31,38].

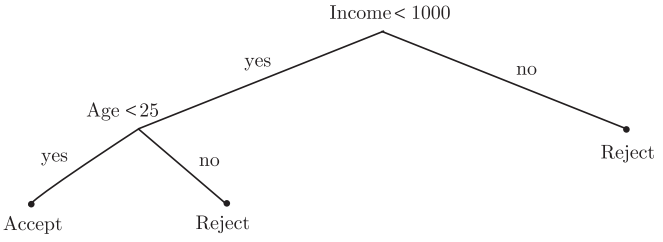


Fig. 2. A binary decision tree.

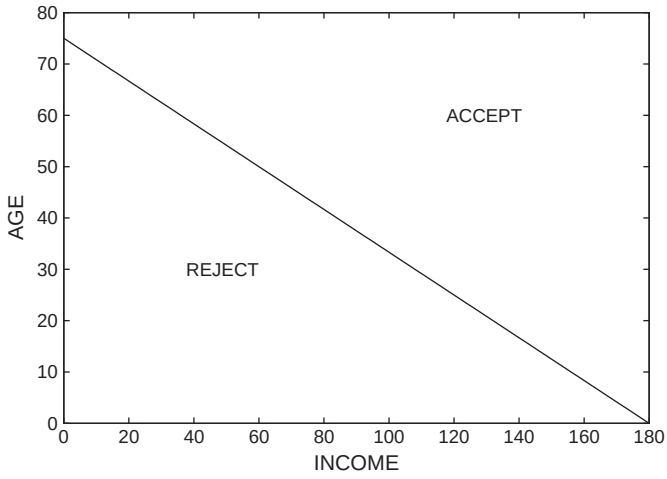


Fig. 3. Example decision boundary for oblique rules.

In a first part of this section, a theoretical model based on the cognitive fit theory is developed; subsequently, research propositions based on this model are formulated.

3.1. Research model

An important mechanism to improve problem-solving task performance is often considered the proper visualization of information [51,52,61]. While a study of Tufte suggests that graphical representations might be useful in all circumstances [52], other research did not confirm these results [15,57,58]. Moreover, different studies have been conducted to find out which representation format is most suited for a specific task (e.g. [4,16,21]). In presenting the theoretical underpinning of this study, we adopt the cognitive fit theory of Vessey [57] (see Fig. 4(a)).

The cognitive fit theory suggests that an appropriate fit between graphical representation and task requirements facilitates the creation of a mental representation of the problem. This in turn allows for higher performance in problem solving. More specifically, two types of tasks are defined, spatial and symbolic tasks, and two types of representations, spatial and symbolic representations. Spatial tasks are defined as those tasks where the problem area as a whole must be considered. These tasks require making associations or perceiving relations in the data. Symbolic tasks on the other hand do not require this general overview but focus instead on extracting discrete values from the data. Analogously, spatial representations are defined as those problem representations that preserve "information about the topological and geometric relations among the components of the problem" thus emphasizing the relations within the data itself [29], while symbolic representations focus on representing data that is symbolic by nature. The cognitive fit theory has been confirmed for simple tasks by different studies [14,53]. However, the thesis of a cognitive fit for more complex tasks (i.e. tasks requiring different processing steps and/or involving large amounts of data) is supported to a lesser extent. E.g., Frownfelter-Lohrke [20] and Speier [46] did not find empirical evidence for a cognitive fit in case of complex tasks.

Note that the cognitive fit theory has also been applied to other information systems research areas, most notably in the domain of information technology acceptance (e.g. [22,30]) and software maintenance (e.g. [43]).

Based on prior studies and the cognitive fit theory, a modified task-representation fit model is proposed (see Fig. 4(b)).

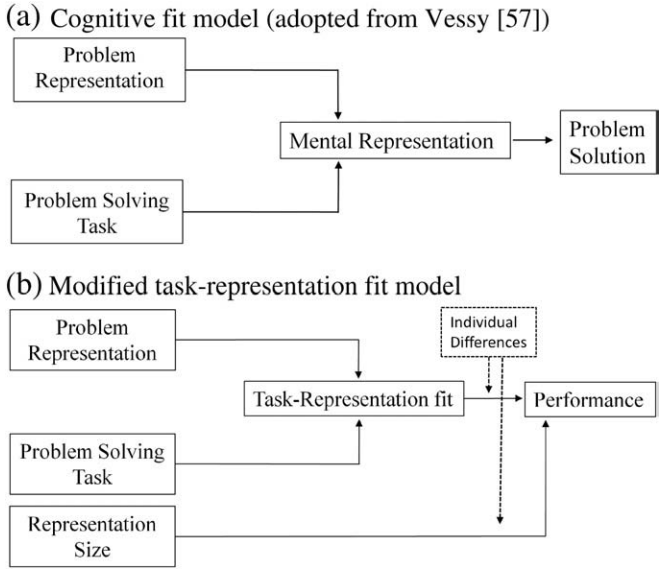


Fig. 4. Cognitive fit model (a) and proposed research model (b).

3.2. Constructs

Our proposed task-representation fit model (cf. Fig. 4(b)) is rooted in the cognitive fit theory and its distinction between spatial and symbolic tasks and representations. Although originally proposed in the context of tabular versus graphical representations (e.g. a table with data versus a bar chart), we believe this theory can be adapted to serve as a theoretic underpinning to our study of the comprehensibility of various representation formats for classification.

Firstly, concerning the type of task, participating respondents perform three problem-solving tasks (model-based classification, answering logical yes–no questions, and determining the equivalence of two models). In line with other research [8,12,46,57], questions requiring the combination of multiple paths in a decision tree or multiple rules in a rule set are considered spatial tasks. E.g., answering logical questions like *'Is it correct that applicants with a high income are more likely to be accepted than applicants with a low income?'* requires the respondents to consider the representation as a whole (i.e. taking into account multiple rules or paths simultaneously). On the other hand, classification-oriented questions like *'How does the representation classify observation X?'* seem to correspond closely to a symbolic task, as they require the user to follow one path in the representation. The experiment is more fully described in Section 4.

Secondly, concerning the type of representation, four different problem representation formats are under investigation in this study: decision trees, decision tables, propositional if–then rules, and oblique rule sets (see Section 2). In line with previous research [12,46,57,59], graphical representations such as decision trees and decision tables are assumed to be more suited for spatial tasks. Textual representation formats on the other hand are assumed to be less suited for these tasks as the textual representation does not graphically reflect the relationships within the knowledge domain. Subramanian et al. [47] further investigated the difference between both graphical representations (decision trees and decision tables) from an effectiveness perspective, measuring the number of correct decisions without taking decision time into account. It was concluded that decision trees perform significantly better than decision tables. Because retrieving a single optimal value corresponds to a symbolic task, one could hypothesize that decision trees are better suited for symbolic tasks than decision tables. However, as discussed in [54], one can cast serious doubt on these results, as the study of Subramanian et al. shows various shortcomings which severely disfavor decision tables. For example, the decision tables are neither complete, nor lexicographically ordered, depriving them from their most attractive properties. Hence, the issue of which graphical representation format is better suited for a specific task remains an open question.

In addition to recognizing the potential impact of task-representation fit on task performance, previous research has also argued that the complexity of the presented model may impact the task performance [6]. This thesis is supported by the cognitive load theory which states that humans are limited in their cognitive abilities [49]. Based on these observations, the construct of 'representation size' was added to our proposed task-representation fit model. It should be noted that the concept of complexity is typically understood as comprising of several (internal and external) aspects, including elements such as size of the model, but also the specific domain context and subjective opinion and prior expectations of the user [6,62]. Therefore, complexity is hard to define and measure exactly. Instead of complexity, we focussed on representation size and its relation to task performance, since no specific domain expertise is assumed. More precisely, the number of variables in the model and the number of rules, nodes in a tree or columns in a decision table are considered. Thus, it will be hypothesized that a larger representation size will affect task performance.

A number of different problem-solving task performance aspects can be measured to assess the task-representation fit and the impact

of representation size. In this study, accuracy, answer confidence, and answer time are investigated; these measures are further detailed in Section 4.3.

In addition to being affected by task-representation fit and representation size, task performance may be further influenced by the individual characteristics of the user. This idea is supported by various studies of the impact of individual differences on the problem-solving process (see e.g. [3,30]). Based on these findings, we postulate that age, prior experience, and education may act as intermediating factors on the effect that the earlier constructs have on task performance. The impact of individual differences are however not the aim of the study as our group of subjects is rather homogenous as far as age, experience, and education is concerned. The subjects participating in the study are a mixture of both doctoral students not involved in the project and graduate business students.

In summary, it seems appropriate to construct an experiment that consists of multiple parts, each with questions that require the use of different types of information. Besides comparing different representation formats, the relation between representation size and answer performance is investigated. More specifically, we want to measure if and to what extent the answer performance of the subjects deteriorates when faced with more complex models.

3.3. Associations and propositions

A fundamental assumption of various machine learning algorithms is that smaller models are more comprehensible than larger ones, and that *representation size* therefore can be used as a proxy for model comprehensibility. This assumption, supported by the cognitive load theory, forms the basis of different rule extraction and pruning algorithms (see e.g. [36,40]). It is believed that the ability to better understand a model will increase task performance. Hence, we formulate the following proposition.

Proposition 1. Representation size.

Independent of the representation format, a larger representation size will result in a decreased task performance, i.e. lower accuracy and confidence, and longer answer times.

Additionally, we expect the validity of the above proposition to depend on the design of the experiment. In the experiment performed in this paper, the respondents were instructed to answer both fast and accurately, with no specific priority for either goal. In other experiments (e.g. [39,59]), respondents were invited to work as fast as possible, but to always try to ensure that the answers are correct. Prior studies indicate that decision-makers will trade off accuracy for effort [26], hence focusing on accuracy could potentially lead to biased results.

Besides measuring the relation between representation size and comprehensibility, this paper also aims to find out under what circumstances one representation format outperforms the others. In line with the proposed task-representation fit model, we theorize that *spatial tasks* are more easily solved by graphical representations, i.e. decision trees and decision tables, as they provide a closer cognitive match for spatial tasks. The performance of these representations is therefore expected to be better than that of a textual representation if the questions require the users to consider the relationships within the model. This is for instance the case for questions requiring the respondents to combine information from multiple rules or to follow multiple paths in the tree, as commonly required during the domain expert's validation of the model. We also believe that the observed differences will be more important for larger representations. Therefore, the following proposition is formulated.

Proposition 2. Spatial tasks.

Graphical representations, decision tables and decision trees, achieve a better performance on spatial tasks than textual representations. Performance differences will be bigger for larger, more complex models.

As for *symbolic tasks*, e.g. the retrieval of the class label predicted by the model for a given observation, we believe that the hierarchical structure of the graphical models has a positive influence on task performance, because their hierarchical organization allows for a very efficient top-down search strategy. On the other hand, because the textual description contains a default rule, such a model is usually relatively small compared to the logically equivalent decision tree or table. We expect that these advantages balance each other out for simple models, but that for elaborate models the efficiency of the search strategy for decision trees and tables will outweigh the smaller model size of the textual description, in line with the study of Speier [46]. Based on these considerations, the following proposition is stated.

Proposition 3. Symbolic tasks.

All three representation formats achieve similar performance on simple symbolic tasks. For larger symbolic tasks, decision trees and tables outperform the textual description.

One could question the usefulness to measure whether a performance difference between the different representation formats can also be observed for symbolic tasks. After all, once the model has been validated by the analyst, its everyday use is usually automated and only seldom will an end-user have to look explicitly at the tree or rule set to make a prediction. However, because finding the matching rule or leaf node for specific conditions is also essential for answering more complex questions, we considered it worthwhile to include symbolic tasks in the experiment.

Until now, we have not discussed the fourth representation that was presented in Section 2: a textual description of oblique rules. Although a direct comparison with the other representations is not feasible, as oblique rules can usually not be transformed into equivalent propositional rules, we expect that even for simple models, this representation will be perceived as very difficult to use. The following proposition is therefore stated:

Proposition 4. Oblique rules.

Oblique rules are considered as difficult to use. Even for simple models, the observed task performance will be low and it will decrease as representation size increases.

3.4. Boundaries of the theory

The cognitive fit model proposed by Vessey [57] only focuses on task-representation fit, thereby not giving sufficient attention to the impact of complexity on task performance. The importance of complexity for task performance has previously been noted by Frownfelter-Lohrke [20] and Speier [46], whose studies failed to empirically confirm the concept of cognitive fit for more complex tasks. Our proposed task-representation fit model investigates a key aspect of their presentation complexity, i.e. representation size. However, it is believed that the proposed model could further benefit from a more comprehensive definition of complexity, and its impact on performance as cognitive burden is dependent on multiple aspects [48] including domain complexity, representation size, information processing activities, and the interplay between these elements.

4. Research method

The effectiveness of the representation formats discussed in Section 2 was empirically tested and compared by means of a computerized experiment. Next, we discuss the setup of this experiment.

4.1. Experimental design

We tested Propositions 1 to 3 using a $2\times3\times7$ within-subjects experimental design. The first factor is representation size consisting of seven different levels, dependent on the number of variables in the model and the number of rules, nodes or columns (Table 1). For example, the decision tree in Fig. 5 corresponds to size level 5. The two other factors, retained from the cognitive fit theory, are representation format and task type. The representation format factor has three levels (textual representation, decision trees, and decision tables) while the task type factor has two levels, i.e. classification (cf. symbolic) and logical (cf. spatial) tasks (see Table 2). Also, a third task type is considered in which respondents must determine whether two representations are equivalent. However, as the representation size is not varied for this task type, it is considered separately. During the analysis, a blocking factor is introduced to avoid the effect of unrecorded nuisance factors such as age, prior experience, and education by taking the individual respondents as blocking factor. An additional representation format is considered, i.e. oblique rules, but since this representation type typically cannot be transformed into one of the other representation formats, it was again taken separately in the analysis.

4.2. Participants

Forty-two graduate business students enrolled in a management informatics course and nine doctoral researchers not involved in the project participated in the study. The graduate students received a small financial compensation for their collaboration. As previous studies have indicated, past experience, age, and education can influence task performance [3,30]. Therefore, we opted to have a fairly homogenous group of respondents which had no prior experience with any of the representation formats or the domain of credit scoring. This was done by taking their study curriculum into account. Moreover, a participant blocking factor was introduced to account for other individual differences.

4.3. Measures

The task performance measures consisted of three elements: accuracy, answer time, and answer confidence. All tasks performed were intellective tasks and as such, had optimal answers [34]. Therefore, accuracy can be objectively determined by measuring it as the Percentage Correctly Classified (PCC) which is the number of correct answers divided by the total number of questions. The respondent accuracy score thus ranges from 0% up to 100%.

The answer time variable is measured in seconds and is the time required for a respondent to answer a question. As indicated earlier, it was stressed at the beginning of the experiment that both accuracy and answer time are equally important. In accordance with other studies (e.g. [59]), only the answer times of the respondents with a correct answer are included and used in the rest of the analysis. Additionally, outlier observations were removed by excluding those observations for which the answer time of the respondent is considerably different from the mean answer time for the particular question.[2]

A third element measured for each question is the confidence a respondent attached to his or her answer. The confidence is measured on a scale of 1 to 5 (Totally Not Confident = 1,…,Very Confident = 5).

4.4. Materials

The respondents participated in a fully computerized experiment which is presented via a web interface. The experimental material included an introduction, three parts with each different tasks and a post-experiment questionnaire. The different tasks served to assess the impact of the previously stated factors on performance while the post-experiment questionnaire evaluated the ease of use for each representation type.

4.4.1. Introduction

The experiment starts with a short introduction presenting the context in which the experiment takes place. The introductory text stresses that a credit manager must be able to work both fast and accurately. Furthermore, an explanation of the different representation formats is given and three example questions are asked to check whether the respondents correctly understand the different representation formats. The example questions are similar to the questions of part 1 in which the respondent will be asked to classify an application based on a given model. The models in the three example questions are very simple models, similar in representation size to the models in Figs. 1 and 2. A large majority of the respondents were able to answer all three example questions correctly.[3]

4.4.2. Experiment

In a first part, the respondent is asked to perform a classification of a new loan application. A credit model is shown to the respondent in one of the different representation formats together with a new loan application. An example of such a question is shown in Fig. 5. The respondent is then asked to assign the correct classification decision (ACCEPT or REJECT) to the new application based on the given model. In addition to making this classification decision, the respondents also indicate how confident they are that they made the correct classification (see Fig. 5, bottom-right corner). Additionally, the computer program records the answer time for each question. A total number of 24 classification problems are presented to the respondent. 3 of the 24 questions are on oblique rules. The 21 other questions are evenly divided across the 3 other representation types for propositional rules (textual description, decision tree and decision table). For the latter representations, it is ensured that each of the seven models has a different representation size and that each model has a semantically equivalent counterpart in the other representation formats. The classification task can be considered 'symbolic' as only a single path or rule is considered each time.

The second part of the experiment takes place in the same context as the first part, but instead of classification applications, a number of logical YES–NO questions about a given credit model are asked. Examples of logical questions are the following.

"Does the model accept all people with an AGE above 60?"

Table 1
Overview representation size (part 1 and part 2).

Size level	1	2	3	4	5	6	7
Variables	2	5	5	6	6	6	6
Decision tree (# leaf nodes)	5	9	11	13	16	19	21
Decision table (# columns)	5	9	11	13	17	19	22
Textual description (# rules)	3	4	5	6	7	8	9

[2] Assume $\bar{x}$ and s are respectively the sample mean time and corresponding sample standard deviation, measured over all correct answers for a particular question. An observation is removed if it is larger than $\bar{x}+5s$ or smaller than $min\{10; \bar{x}/3\}$. The latter condition ensures that no answers of respondents that guessed correctly are included.

[3] 7 respondents made 1 error and only 1 respondent made 2 errors. The results of these respondents were not removed as their error rate was normal when measured over the entire experiment.

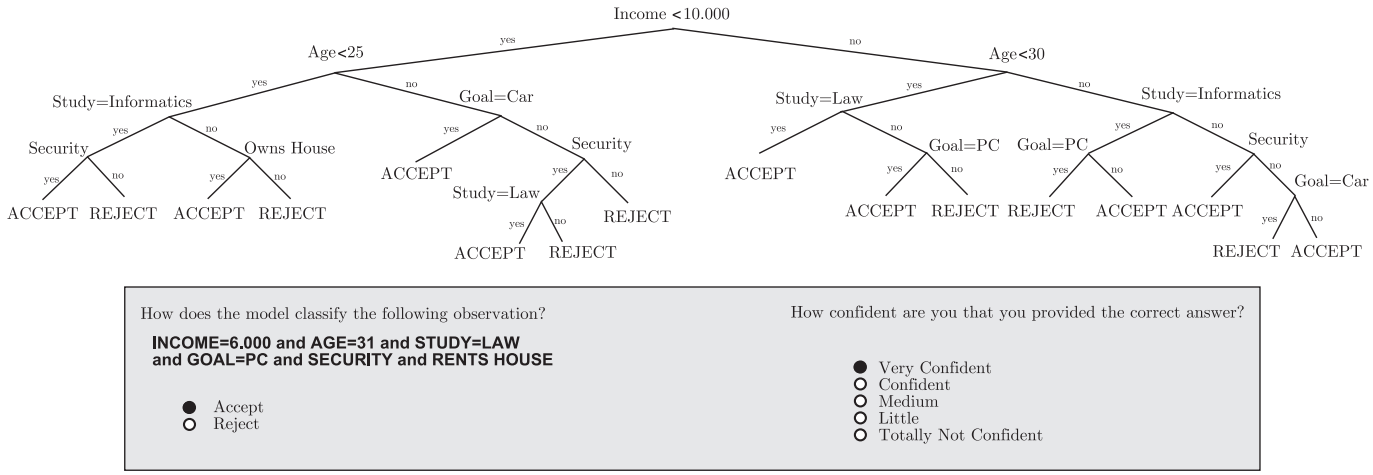


Fig. 5. Example question taken from part 1.

"You are 26 years old. Does it matter for the credit decision whether you have studied INFORMATICS or not?"

The main difference with the first part of the experiment is that the respondents must now combine information from multiple paths. Again, 24 different questions are presented to the respondent. There are 3 questions on oblique rules and 7 for each of the other representation types. Due to the combinatory focus of the questions, questions from this second part are considered as 'spatial' in nature and we believe this task is considerably more demanding.

In the third and last part of the experiment, again in the same context as above, a black-and-white figure is presented and the respondents must decide whether the model depicted next to it is equivalent to this figure. An example question is shown in Fig. 6. The third part consists of 11 questions, 2 questions on oblique rules and 3 on each of the representation formats for propositional rules. To allow a fair comparison, it was ensured that similar questions were selected for all representations of propositional rules. Whereas this part of the experiment often requires the respondents to consider all possible paths, we still consider the task 'symbolic' rather than 'spatial', because the respondent can solve the questions by considering each path in turn and comparing the class label with the figure. No relationships between the different paths must be considered.

4.4.3. Post-experiment questionnaire

After performing the experiment, participants were invited to indicate their preference about which of the four representations was easiest to work with and also to indicate which representation they found the most difficult to work with. Upon completion of this questionnaire, the participants had completed the experiment.

4.5. Procedures

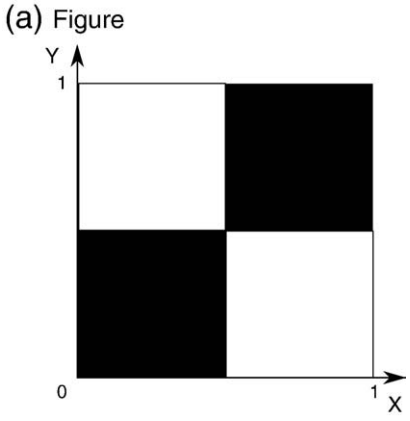
Experimental sessions were held in a university computer lab in order to provide a common computing infrastructure for each session. As the complete experiment takes place in a single computer session, including the introductory text and post-experiment questionnaire, no contact between respondents is required. The different respondents did not receive any feedback on their performance during the experiment as this could bias their results for the rest of the experiment. A single facilitator was present during the experiment, welcoming the respondents. Since the whole experiment was computerized, this facilitator played no other role during the experiment. Throughout the experiment, to avoid bias in favor of any of the representations, the different questions are randomly ordered in a different way for each respondent. Also, to avoid learning effects we randomly switched some of the ACCEPT–REJECT labels. Typically, the experiment was completed in less than 45 min, but no time limits were imposed.

Table 2
Layout of the experiment.

Presentation of the representation formats		
Example questions		
Part 1:	Classification questions	(24 questions in random order)
Part 2:	Logical problem questions	(24 questions in random order)
Part 3:	Visual comparison tasks	(11 questions in random order)
Evaluation of the ease of use of representations		

5. Results

This section compares the results of the three representation types for propositional rules, i.e. decision trees, decision tables, and the textual description. Since it is usually impossible to transform oblique rules into logically equivalent propositional rules, the results for the



X	<0.5		≥0.5	
Y	<0.5	≥0.5	<0.5	≥0.5
BLACK	X			X
WHITE		X	X	

Fig. 6. Example question from part 3: Does the model correctly describe the figure?

textual description of oblique rules are discussed in a separate section afterwards.

In order to test the statistical significance of the obtained results, a number of non-parametric tests and parametric tests are applied in accordance with the statistical literature. Each of the different tests are assessed at a significance level of 95%. If groups of statistical tests are performed, a Bonferroni correction is applied to more correctly assess the statistical significance of the individual tests. The appropriate p-value for the individual tests is indicated each time. Non-parametric tests are chosen as they do not make assumptions about the distribution of the underlying data whereas ANOVA (ANalysis Of VAriance) typically assumes the data is normally distributed.

Firstly, to test for differences in *accuracy*, a Cochran's Q test is performed [44]. This non-parametric test is suitable for a dichotomous outcome variable and can be used to test differences in proportions across multiple treatments. In this context, the dichotomous outcome variable is whether an individual answer provided by the user is correct or not, and the various size levels constitute the different 'treatments'. Thus, Cochran's Q test is used to test whether accuracy (the percentage of correct answers) is significantly different for the various representation sizes considered. Subsequently, a Friedman test is applied; this non-parametric test can be used for analyzing randomized complete block designs. A Friedman test is used to detect if there are differences across multiple treatments by ranking them in each of the blocks according to some criterion of interest and looking at the average and variance of ranks to decide whether the treatments are ranked similarly or not. The Friedman test is used to determine whether there is a difference in accuracy between the representation formats (treatments) over respondents (blocks). Whereas the Friedman test verifies whether there any differences among the various representations, we are also interested in comparing pairs of two representations to find out which representation is significantly better or worse. Hence, we also use a Wilcoxon matched pairs test. This test can be regarded as the non-parametric equivalent of a matched pairs t-test and is used to assess the difference between two groups of paired data. In our study, the test is applied to determine whether the accuracy of two representations is statistically different.

A 3-way ANOVA is used to assess the results for *answer time*. ANOVA is a parametric test in which the mean across different groups is compared, and it is chosen here to analyze differences in answer time considering its suitability for a repeated measures design and its ability to cope with the number of observations differing across the groups. The first two factors consist of the representation type and the representation size, and the respondents are considered as the third (blocking) factor.

In the section concerning *confidence*, two separate Friedman tests are used; first a Friedman test is applied to assess differences in confidence across the size levels. Furthermore, a Friedman test is also used to assess whether there exists differences between the various representation formats. Finally, a Wilcoxon matched pairs test to pairwise compare the individual representation formats across the three parts of the experiment is performed.

5.1. Accuracy

In Table 3, an overview of the percentage correctly answered questions per size level and representation type is given.[4] It can be observed that for the first part of the experiment, the accuracy levels are close or equal to 100%. This indicates that the respondents are able to work correctly with each of the representation formats. The bold-faced values in Table 3 indicate for each size level the representation format with the highest performance. Note that, on average, decision tables performed best in all three parts of the experiment.

Table 3
Percent correct answers.

Size level/question	1	2	3	4	5	6	7	Avg.
Part 1: Classification questions								
Decision tree	96.1	**100**	96.1	96.1	86.3	**98.0**	86.3	94.1
Decision table	**100**	**100**	96.1	**100**	**100**	94.1	**96.1**	**98.0**
Textual description	**100**	**100**	**100**	94.1	94.1	90.2	90.2	95.5
Avg.	98.7	100	97.4	96.7	93.5	94.1	90.9	
Part 2: Logical problem questions								
Decision Tree	94.1	**94.1**	82.4	92.2	70.6	**62.7**	72.5	81.2
Decision Table	**100**	92.2	82.4	92.2	**92.2**	**62.7**	**76.5**	**85.5**
Textual Description	94.1	82.4	**90.2**	**100**	82.4	58.8	54.9	80.4
Avg.	96.1	89.6	85.0	94.8	81.7	61.4	68.0	

	1	2	3	Avg.
Part 3: Visual comparison tasks				
Decision Tree	**82.4**	84.3	76.5	81.1
Decision Table	76.5	**90.2**	**96.1**	**87.6**
Textual Description	58.8	78.4	80.4	72.5
Avg.	72.6	84.3	84.3	

For the second part and irrespective of the representation format used, the accuracy levels are shown to drop slightly for the models with low size, and rather drastically for those with a higher size level. However, it should be noted that the drop in accuracy for levels 6–7 is most likely not just caused by the increased representation size, but also by the type of logical question. Whereas an example question for the lower size levels 1–5 is '*A person with INCOME = 60.000 wants to buy a car (GOAL = CAR). Will he always be ACCEPTED?*', a typical question at level 6 and 7 is as follows: '*A person is ACCEPTED. Will a person with the same characteristics, but a higher INCOME, then also be ACCEPTED?*'. While these questions might seem similar at first, the second question is more difficult, as can be experienced by solving both questions for the tree of Fig. 5. For the first question, one only has to find the leafs whose paths satisfy the conditions 'INCOME = 60.000 and GOAL = PC', and check the corresponding class label. If the label is 'ACCEPT', then one has to continue checking the next leaf that satisfies the conditions, until either all leafs are checked or until a leaf is encountered with a 'REJECT' class. To answer the second question, a more complex reasoning is required. For each leaf with the 'ACCEPT' class and a 'smaller than' condition on INCOME, one has to check whether all paths with the same conditions on other variables, but a 'higher than' condition on INCOME, also lead to leafs with the 'ACCEPT' label.

Nonetheless, both in the results of part 1 and those of levels 1–5 of part 2, a negative correlation between accuracy and representation size can be observed. This seems to support Proposition 1: for more elaborate models, one can expect a decrease in answer performance, measured as the percentage of correct answers. To verify this statement more formally, Cochran's Q-test was performed to evaluate:

$$H_0 : Acc_{level1} = Acc_{level2} = \ldots = Acc_{level7}.$$

Most of the six tests (separate tests were performed for each representation and separately for parts 1 and 2 of the experiment) were significant ($p < 0.0083$) after Bonferroni multi-comparison correction, providing support for Proposition 1. Only for decision tables and the textual description in part 1 of the experiment, the result was not statistically significant ($p = 0.125$ and $p = 0.029$). The test results are summarized in Table 4.

To assess whether there is a difference in the proportion of correct answers between the various representations, a Friedman test was

[4] As discussed above, size is only varied in part 1 and part 2 of the experiment. For the third part, the columns must only be considered as 'question number', because each of the three questions was created with a more or less equal difficulty level.

Table 4
Cochran's Q test: results for accuracy differences across size levels.

	Part 1	Part 2
Decision tree	17.4 (p = 0.0080)	36.3 (p < 0.001)
Decision table	10 (p = 0.1247)	40.3 (p < 0.001)
Textual description	14 (p = 0.0292)	59.6 (p < 0.001)

performed to compare the percentage of correct answers of each respondent for each representation. The null hypothesis is that each representation format provides the same accuracy.

$$H_0 : Acc_{Tree} = Acc_{Table} = Acc_{Textual}$$

The test is first performed for each part of the experiment separately, and afterwards for the combined results over the three parts. Although the test is only significant for part 3 of the experiment using an $\alpha = 0.0167$, the combined result is still significant ($p = 0.019$) because a Bonferroni correction is not needed for the (single) overall test. The test details are given in Table 5.

After the Friedman tests, Wilcoxon matched pairs tests were performed between each pair of representation formats. For the entire experiment, we can conclude that overall decision tables performed better than both decision trees and textual description ($p = 0.018$ and $p < 0.001$), whereas the difference between decision trees and a textual description was not found to be significant ($p = 0.45$).

Table 5
Friedman test: results for accuracy Differences across representation formats.

	Q	p (χ^2 > Q)
Part 1	6.8	0.033
Part 2	2.46	0.29
Part 3	12.29	<0.01
Overall	7.79	0.019

5.2. Answer time

In Table 6 and Fig. 7, an overview of the mean time required to answer each question is given. Additionally, the standard deviation for each question is provided (in italic). For each size level, the boldface value indicates the representation that required the least time. As mentioned in Section 4.3, some respondents are excluded as they needed disproportionately much or little time. The values between brackets in Table 6 refer to the number of observations from which the corresponding average time was calculated. Note that for the global averages these counts were not taken into account.

To assess Proposition 1, relating to the impact of representation size on required time to answer, a 3-way ANOVA is performed for which the results can be observed in Table 7. The tests show that there are significant differences in the mean answer time for the different representations and size levels. Pairwise comparisons between the representations show that for part 1 decision tables score significantly better than both decision trees and textual description. For part 2 of the experiment both decision tables and textual description perform better than decision trees. In part 3, decision tables score significantly better than textual descriptions, which in turn require significantly less time than decision trees.

5.3. Confidence

For each question, the respondents were requested to indicate how confident they were that they had provided the correct answer. In Table 8, we provide an overview of the mean confidence per question and representation format. As indicated earlier, confidence was measured on a 5-point scale with 1 = Totally Not Confident and 5 = Very Confident. The bold-face values indicate the representation format for which the confidence is highest. For 12 of the 17 questions, the respondents expressed most confidence in their answers when the representation format was a decision table. Also, the results of Table 8 correlate heavily with the results for accuracy in Table 3. The respondents are clearly able to assess the difficulty of the questions

Table 6
Mean time spent in seconds (# observations included in mean).

Size level/question								
	1	2	3	4	5	6	7	Avg.
Part 1: Classification questions								
Decision tree	11.8 (48)	**14.0** (51)	19.8 (49)	22.7 (48)	35.0 (44)	25.0 (50)	36.9 (44)	23.6
St. dev.	*4.8*	*7.5*	*10*	*10.1*	*18.7*	*12*	*20.3*	
Decision table	**10.4** (51)	16.2 (51)	22.3 (49)	**22.0** (50)	**19.3** (51)	**23.5** (48)	**21.3** (48)	**19.3**
St. dev.	*3.2*	*9.4*	*11.7*	*8.9*	*6.1*	*16.8*	*9.4*	
Textual description	14.0 (51)	19.1 (50)	**19.5** (48)	24.0 (48)	21.2 (47)	25.5 (46)	31.2 (46)	22.1
St. dev.	*5.8*	*11*	*10.2*	*10.8*	*10.3*	*14.5*	*14.3*	
Avg.	12.0	16.4	20.5	22.9	25.2	24.7	29.8	
Part 2: Logical problem questions								
Decision tree	25.9 (48)	24.5 (48)	40.4 (42)	24.0 (47)	39.6 (36)	62.5 (31)	51.5 (35)	38.3
St. dev.	*11.8*	*9.7*	*17.7*	*11*	*14.6*	*45.6*	*28.8*	
Decision table	**15.9** (51)	**21.4** (46)	**22.6** (42)	40.8 (47)	**31.3** (47)	52.1 (32)	55.3 (38)	**34.2**
St. dev.	*5.7*	*9.9*	*10.3*	*19.3*	*10.5*	*32.6*	*30.1*	
Textual description	25.0 (48)	29.4 (40)	24.5 (44)	**21.6** (50)	40.3 (39)	**48.8** (29)	**50.2** (27)	34.3
St. dev.	*11.8*	*14.5*	*10.9*	*8.9*	*21.3*	*33.6*	*21.4*	
Avg.	22.2	25.1	29.2	28.8	37.1	54.5	52.3	
		1		2		3		Avg.
Part 3: Visual comparison tasks								
Decision tree		68.8 (40)		64.1 (40)		66.9 (36)		66.6
St. dev.		*36.3*		*28.6*		*33.9*		
Decision table		**39.6** (36)		**44.5** (45)		40.0 (47)		**41.4**
St. dev.		*13.9*		*24.8*		*14.4*		
Textual description		48.4 (28)		58.9 (37)		**38.3** (40)		48.5
St. dev.		*24.3*		*31.6*		*13.3*		
Avg.		52.3		55.8		48.4		

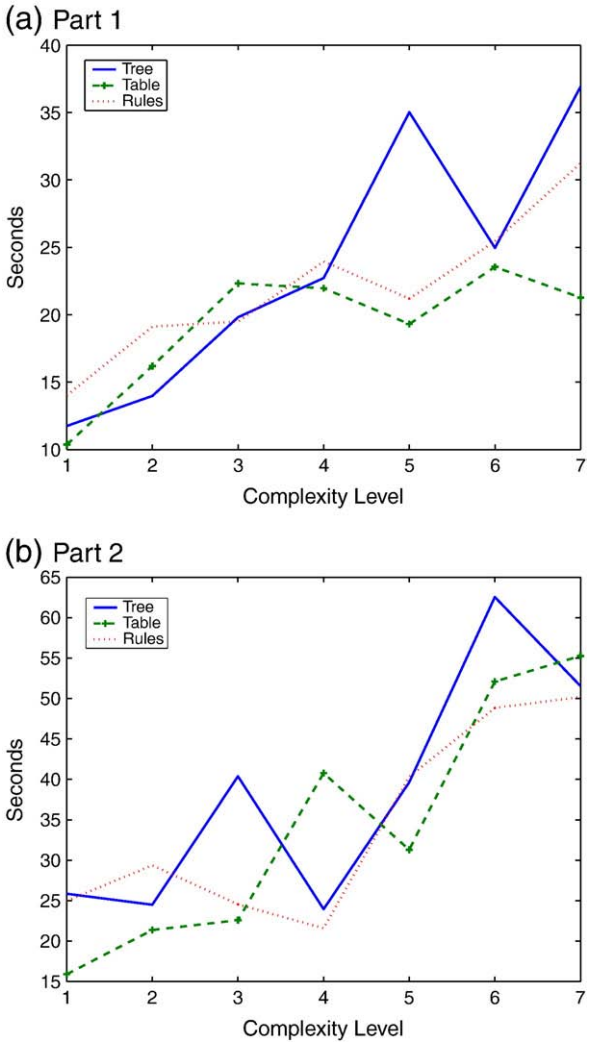


Fig. 7. Mean time spent.

correctly. To further assess Proposition 1, a Friedman test was performed to evaluate whether the reported confidence is similar over the different size levels.

$$H_0 : Conf_{level1} = \ldots = Conf_{level7}$$

Table 7
ANOVA: results for answer time.

Source	DF	Type III SS	F	Pr>F
Part 1				
Representation	2	2919.1	12.39	<0.001
Size	6	30240.5	42.78	<0.001
Block	50	27537.7	4.67	<0.001
Total	1017	173214.5		
Part 2				
Representation	2	2691.0	4.09	0.017
Size	6	107004.4	54.27	<0.001
Block	50	71299.3	4.34	<0.001
Total	866	453066.6		
Part 3				
Representation	2	40824.1	42.23	<0.001
Question Number	2	2555.6	2.64	0.073
Block	48	88716.1	3.82	<0.001
Total	348	275439		

Table 8
Confidence.

Size level/question								
	1	2	3	4	5	6	7	Avg.
Part 1: Classification questions								
Decision tree	**4.9**	**4.7**	**4.5**	4.5	4.0	4.4	3.7	4.4
St. dev.	*0.45*	*0.55*	*0.78*	*0.67*	*0.98*	*0.87*	*1.23*	
Decision table	4.8	4.6	4.5	**4.6**	**4.5**	**4.5**	**4.5**	**4.6**
St. dev.	*0.51*	*0.67*	*0.78*	*0.76*	*0.88*	*0.76*	*0.7*	
Textual description	4.6	4.5	4.4	4.2	4.4	4.4	4.1	4.4
St. dev.	*0.6*	*0.78*	*0.94*	*0.86*	*0.82*	*0.85*	*0.74*	
Avg.	4.8	4.6	4.5	4.4	4.3	4.4	4.1	
Part 2: Logical problem questions								
Decision tree	4.4	**4.5**	4.2	**4.5**	3.9	3.5	3.6	4.1
St. dev.	*0.78*	*0.64*	*0.91*	*0.7*	*1.1*	*1.1*	*1.2*	
Decision table	**4.6**	4.4	**4.4**	4.2	**4.4**	**3.7**	**3.7**	**4.2**
St. dev.	*0.57*	*0.67*	*0.67*	*0.89*	*0.69*	*1.03*	*1.06*	
Textual description	4.4	4.2	4.2	4.3	3.9	2.9	2.9	3.8
St. dev.	*0.63*	*0.84*	*0.78*	*0.82*	*1*	*1.3*	*1.26*	
Avg.	4.5	4.4	4.3	4.3	4.1	3.3	3.4	

	1	2	3	Avg.
Part 3: Visual comparison tasks				
Decision tree	4.1	4.0	4.0	4.0
St. dev.	*1.21*	*1.22*	*1.2*	
Decision table	**4.2**	**4.2**	**4.3**	**4.2**
St. dev.	*1.13*	*0.92*	*0.86*	
Textual description	3.9	3.9	4.1	4.0
St. dev.	*1.08*	*1.16*	*1.1*	
Avg.	4.0	4.1	4.2	

Separate tests were performed for part 1 and part 2 of the experiment and for each representation totaling 6 tests. Applying the Bonferroni correction, the required significance level of the separate tests is $\alpha = 0.0083$. In Table 9, the results of the different tests are given. Most of the 6 tests are found to be highly significant. Only for the decision tables in part 1 is the confidence between the size levels not significantly different ($p = 0.025$ while the significance threshold is $\alpha = 0.0083$). Thus support for Proposition 1 is found: respondents showed more confidence in their responses if the representation size is small.

To determine whether there exists a difference in the mean confidence over the different representations, a Friedman test was applied to each part of the experiment separately. Thus three different tests are performed (significance level is $\alpha = 0.0167$ after Bonferroni correction). The following null hypothesis is tested.

$$H_0 : Conf_{Table} = Conf_{Tree} = Conf_{Textual}$$

The results of the test are presented in Table 10. In two of the three cases, the null hypotheses are rejected. Only for part 3 is the difference between the representation types not significant. The overall test has a p-value <0.001 indicating that the differences in the confidence levels over the three representations are highly significant.

Finally, a Wilcoxon matched pairs test is performed to compare individual representations for each of the 3 parts of the experiment. Table 11 shows that overall, the respondents are more confident using decision tables than decision trees or a textual representation. The difference between a textual representation and a decision tree is not statistically significant with an overall p-value of 0.036 (higher than

Table 9
Friedman test: results for confidence differences across size levels.

	Part 1	Part 2
Decision Tree	100.21 (p<0.001)	81.68 (p<0.001)
Decision Table	12.45 (p=0.025)	78.57 (p<0.001)
Textual Description	40.05 (p<0.001)	128.88 (p<0.001)

Table 10
Friedman test: results for confidence differences across representation formats.

	Q	p (χ^2>Q)
Part 1	18.57	<0.001
Part 2	19.41	<0.001
Part 3	7.36	0.025
Overall	27.45	<0.001

the required $\alpha = 0.016$). Between brackets it is indicated which representation format of each pair has the highest confidence.

5.4. Ease of use

After conducting the test, the respondents were requested to indicate their preference on the various representation formats. From Fig. 8, we can conclude that a large majority voted for decision tables as the preferred representation format, whereas oblique rules were considered to be the least comprehensible. These results add to the evidence that decision tables are the preferred representation format if the aspect of interpretability is a key requirement.

5.5. Oblique rules

In this section, it is investigated whether the perceived complexity of oblique rules is also reflected in a low accuracy or confidence and a longer required time. In the design of the experiment, it was taken into consideration that oblique rule sets usually require fewer rules than propositional rule sets since oblique rules can create decision boundaries that are non-parallel to the input axes. Thus, oblique rule sets allow for decision boundaries that better capture the input space. It is therefore highly recommended to perform experiments with real-life data to correctly take into consideration these differences in rule set sizes. For oblique rules, we only performed some tests with very small representations consisting of just a few rules. The goal is to test whether the performance is acceptable for such very small models. If this is not the case, i.e. when the respondents are not even able to answer questions correctly for these small representations, then larger representations should certainly be avoided.

For parts 1 and 2, the three questions consist of two, three, and five rules, respectively, whereas for the third part the two questions consist of three and five rules, respectively. These numbers correspond approximately to size levels 1–3 for the propositional rules. In Table 12, an overview is given of the results for oblique rules.

If these results are compared informally with those for the representations discussed before, one can observe that the performance on part 1 and 2 is slightly worse on the aspects of accuracy and time, and more or less similar for the aspect of confidence. Only for the third part of the experiment, the performance of the oblique rules seems drastically lower than that of the other representations. Although a more elaborate study is certainly required, it seems that from this experiment with very small models, one can conclude that a textual description of oblique rules could be useful for those data sets that can be described with far fewer oblique rules than propositional rules. One should however also take into account the reports on ease of use: a majority of the respondents consider oblique rules as very difficult to use.

Table 11
Wilcoxon matched pairs: results for confidence pairwise comparison of representation formats.

	Part 1	Part 2	Part 3	Overall
Table-Textual	<0.01 (1)	<0.01 (1)	<0.01 (1)	<0.01 (1)
Table-Tree	<0.01 (1)	0.090 (1)	0.078 (1)	<0.01 (1)
Textual-Tree	0.93 (=)	<0.01 (2)	0.66 (=)	0.036 (2)

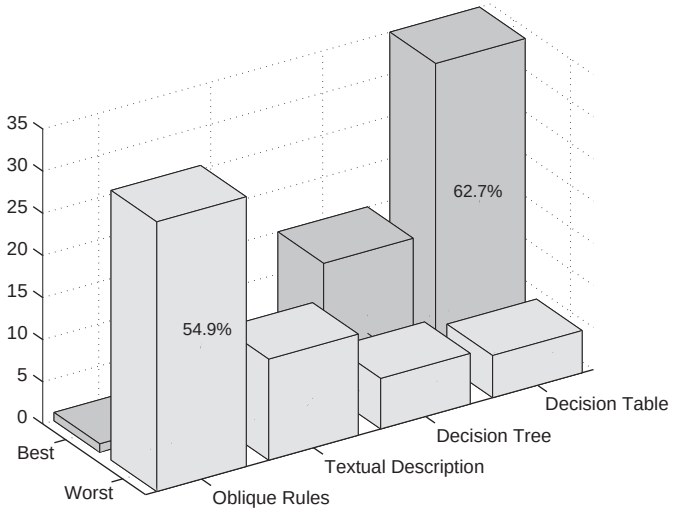


Fig. 8. Ease of use: Percent of respondents evaluating each representation as best/worst.

5.6. Relation representation size/comprehensibility

Another item of interest in this study was to quantify the impact of representation size on comprehensibility. Whereas the previous tests indicated that larger models are less comprehensible than smaller ones, the exact relation between both measures was not yet specified. Based on the percentage of correct answers on part 2 (Table 3), a drop of approximately 4% to 6% can be observed per size level increment. This indicates that there exists a relation between representation size and comprehensibility, and thus the existence of some threshold value for representation size. If a model is larger than this threshold, the model should be regarded as opaque and not suited for practical use.

In part 1, the negative impact of representation size on comprehensibility was less pronounced, although it needs to be taken into account that all models in this study, even those of size level 7, are still small in comparison with some of those reported frequently in the machine learning literature.

6. Discussion

In this section, the results of the experiment are discussed in light of the propositions stated in Section 3. Afterwards, the implications for future research are presented and the limitations of this study are identified.

6.1. Evaluation of propositions

The results of the study found support for Proposition 1 as discussed in Section 3. Larger representations result in a decrease in answer accuracy, an increase in answer time, and a decrease in confidence. These findings are consistent with a generally accepted assumption by the data mining community that people find smaller models more comprehensible. However, this assumption was until now not tested empirically in this context [35].

Table 12
Results for oblique rules.

Question									
	1	2	3	1	2	3	1	2	3
	Accuracy			Time (# obs.)			Confidence		
Part 1	98.0	86.3	92.2	15.0(50)	26.9(44)	37.5(47)	4.8	4.6	4.4
Part 2	96.1	68.6	94.1	15.9(49)	24.3(35)	32.1(46)	4.6	4.5	4.3
Part 3	66.6	45.1	–	76.1(32)	92.7(18)	–	2.7	2.8	–

There is limited support for Propositions 2 and 3, mainly due to the excellent results observed for decision tables. Beforehand, it was assumed that the graphical representations, i.e. trees and tables, either have equivalent or better performance than the textual description on both symbolic and spatial tasks. Whereas this assumption proved correct for decision tables, the results are more mixed for decision trees. Although both are graphical representations, we believe that the better results for the decision tables can mainly be explained by the physical conciseness of this representation format. While a decision tree of moderate representation size already occupies most of the computer's screen, the logically equivalent decision table can often be represented on a fraction of it. This conciseness enhances the respondent's ability of establishing the relationships between different columns.

Besides conciseness, we believe that decision tables have the added advantage that the order in which variables must be considered is the same for each column, whereas for a decision tree the order in which conditions are encountered depends upon the path followed. For example, in order to reach the 6th leaf in the decision tree of Fig. 5, one has to evaluate the variable 'goal' before encountering the variable 'study', whereas in the path leading to the 10th leaf, the testing order of both variables is reversed. This path-dependent ordering of the variables seems inefficient for a fast search within the tree as the user has to switch often between testing a condition in the tree and retrieving a value in the description of the observation. For decision tables, because it is known in advance what the next variable is on which a test can occur, we suspect the user might be able to memorize multiple values at once, which facilitates a more efficient search process.

Proposition 4 concerning oblique rules was also only partially supported. Whereas the textual description of oblique rules was indeed considered by a majority of the users as difficult to use, the actual performance was not as bad as we had expected. However, only experiments with very small rule sets were performed and the above result might therefore not be valid for larger rule sets. We expect that oblique rules will only prove useful for those data sets that can be described with far fewer oblique rules than propositional rules.

6.2. Implications and limitations

In spite of the excellent performance of decision tables in this study, there seems to be a lack of learning algorithms for this representation format. Whereas there are many algorithms capable of learning rules [10,11] or trees [5,37] directly from a set of observations, only very few algorithms are able to learn decision tables directly [28]. While one can always convert the obtained rule set or tree into an equivalent table, we expect that direct learning of decision tables might result in more compact tables. For example, one could use specifically adapted pruning techniques in the learning process. Especially in the context of rule extraction, where comprehensibility is the main motivation, algorithms that are able to convert a black box directly into an equivalent decision table might offer significant advantages.

It is important to consider the above conclusions in light of the study's limitations. Firstly, none of the respondents can be considered as an expert user in any of the representation formats used in the study. One should also be careful not to generalize these results beyond the representation formats that were included in the study. For example, the excellent comprehensibility of single-hit decision tables is not necessarily valid for other types of decision tables.

Because all models in this study were designed for a binary classification problem, i.e., to accept or reject a credit application, the default rule in the textual descriptions of the models could account for approximately half the leafs or columns in the equivalent decision tree or table. The number of rules in the textual description is therefore only half the number of columns or leafs in the corresponding decision tree or table (see Table 1). If the data would correspond to a multi-class problem (e.g. classify as either strong accept, weak accept, weak reject, strong reject), the default rule would probably account for only a few leafs, and the ratio of rules versus leafs or columns would be larger than 1/2. Performance of the textual description is therefore expected to be even worse for multi-class problems.

Furthermore, the study confirms the negative impact of representation size on comprehensibility. The question can therefore be raised to what extent the representations discussed in this study continue to remain useful once they exceed a certain size as we can imagine that the expected number of errors made by users becomes too large for many practical applications. Similar observations can be made for the relationships between representation size and answer time, and between representation size and answer confidence.

7. Conclusion and future research

In this paper, an empirical study was presented which investigates the comprehensibility of a number of alternative predictive model representations, i.e. decision tables, decision trees, propositional rules, and oblique rules. An end-user experiment was designed to test the accuracy, response time, and answer confidence for a set of problem-solving tasks involving the former representations. The results showed that, on the aspect of comprehensibility, decision tables provide significant advantages. For each part of the experiment, the respondents were able to answer the questions faster, more accurately and more confidently using decision tables than using any of the other representation formats. Additionally, a majority of the users found decision tables the easiest representation format to work with.

Furthermore, evidence supporting the thesis that answering logical questions is considerably more difficult than classifying new observations was found. Whatever the representation format, the proportion of correct answers drops sharply for more elaborate models and it can be questioned whether such degree of interpretability is acceptable in practical applications where these models must be validated or where explanatory power is deemed important. Therefore, further collaboration between the data mining and cognitive science communities is recommended to create algorithms and representations that are both predictive as well as comprehensible.

Since this study only considered a selection of representation formats presented to an inexperienced audience, an interesting topic for further research would be to investigate whether the obtained conclusions also hold for experienced users. Also the comprehensibility of oblique rules should be further investigated. Although the experimental results indicated that end-users consider this type of rules hard to use, it should be noted that oblique rules have some attractive features. The nature of oblique rules allows for more complex decision boundaries and therefore, the number of rules is often smaller than in the case of propositional if–then rules. Finally, our results suggest the existence of a relation between representation size and comprehensibility. Hence, further research could be undertaken to quantify the precise relationship in order to determine the size threshold for a model to be comprehensible.

Acknowledgements

This research was supported by the Odysseus program (Flemish Government, FWO) under grant G.0915.09.

The authors also extend their gratitude to the editor and the anonymous reviewers, as their insightful feedback and suggestions certainly contributed much to the quality of the paper.

References

[1] R. Andrews, J. Diederich, A. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge Based Systems 8 (6) (1995) 373–389.

[2] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, J. Vanthienen, Benchmarking state of the art classification algorithms for credit scoring, Journal of the Operational Research Society 54 (6) (2003) 627–635.

[3] I. Benbasat, R.N. Taylor, Behavioral aspects of information processing for the design of management information systems, IEEE Transactions on Systems, Man, and Cybernetics 12 (4) (1982) 439–450.
[4] I. Benbasat, A.S. Dexter, P. Todd, An experimental program investigating color-enhanced and graphical information presentation: An integration of the findings, Communications of the ACM 29 (11) (1986) 1094–1105.
[5] L. Breiman, J. Friedman, R. Olsen, C. Stone, Classification and Regression Trees, Wadsworth and Brooks, 1984.
[6] D. Campbell, Task Complexity: A Review and Analysis, Academy of Management Journal 13 (1) (1988) 40–52.
[7] E. Cantú-Paz, C. Kamath, Inducing oblique decision trees with evolutionary algorithms, IEEE Transactions on Evolutionary Computation 7 (1) (2003) 54–68.
[8] A. Chandra, R. Krovi, Representational congruence and information retrieval: Towards an extended model of cognitive fit, Decision Support Systems 25 (1999) 271–288.
[9] F. Chen, Learning accurate and understandable rules from SVM classifiers, Simon Fraser University, Master's thesis, 2004.
[10] P. Clark, T. Niblett, The CN2 induction algorithm, Machine Learning 3 (4) (1989) 261–283.
[11] W. Cohen, Fast effective rule induction, in: A. Prieditis, S. Russell (Eds.), Proc. the 12th International Conference on Machine Learning, Morgan Kaufmann, Tahoe City, 1995, pp. 115–123.
[12] R.A. Coll, J.H. Coll, G. Thakur, Graphs and tables: A four-factor experiment, Communications of the ACM 37 (4) (1994) 76–86.
[13] R. Colomb, Representation of propositional expert systems as partial functions, Artificial Intelligence 109 (1–2) (1999) 187–209.
[14] A. Dennis, T. Carte, Using geographic information systems for decision making: extending cognitive fit theory to map-based presentations, Information Systems Research 9 (2) (1998) 194–203.
[15] G. DeSanctis, Computer grapics as decision aids: directions for research, Desicion Sciences 15 (4) (1984) 463–487.
[16] G.W. Dickson, G. DeSanctis, D.J. McBride, Understanding the effectiveness of computer graphics for decision support: A cumulative experimental approach, Communications of the ACM 29 (1) (1986) 40–47.
[17] Equal Credit Opportunity Act. United States Code, 1974.
[18] A. Feelders, H. Daniels, M. Holsheimer, Methodological and practical aspects of data mining, Information & Management 37 (2000) 271–281.
[19] A.A. Freitas, Are we really discovering "interesting" knowledge from data? expert update, The BCS-SGAI Magazine 9 (1) (2006) 41–47.
[20] C. Frownfelter-Lohrke, The effects of differing information presentations of general purpose financial statements on users' decisions, Journal of Information Systems 12 (4) (1998) 99–107.
[21] D. Gilmore, T. Green, Comprehension and recall of miniature programs, International Journal of Man-Machine Studies 21 (1) (1984) 31–48.
[22] D. Goodhue, R. Thompson, Task-technology fit and individual performance, MIS Quarterly 19 (2) (1995) 213–236.
[23] R. Halverson, An empirical investigation comparing if–then rules and decision tables for programming rule-based expert systems, Proceedings of 26th Hawaii International Conference on System Sciences, 1993, pp. 316–323.
[24] J. Huysmans, B. Baesens, J. Vanthienen, ITER: an algorithm for predictive regression rule extraction, 8*th International Conference on Data Warehousing and Knowledge Discovery*, DaWaK 2006, 4081, Springer Verlag, 2006, pp. 270–279.
[25] J. Huysmans, B. Baesens, J. Vanthienen, Using rule extraction to improve the comprehensibility of predictive models, FETEW research report KBI 0612, Technical report, Katholieke Universiteit Leuven, 2006.
[26] S. Jarvenpaa, G. Dickson, Graphics and managerial decisionmaking: Research based guidelines, Communications of the ACM 31 (6) (1988) 764–774.
[27] U. Johansson, R. König, L. Niklasson, Automatically balancing accuracy and comprehensibility in predictive modeling, Proceedings of the 8th International Conference on Information Fusion, 2005.
[28] R. Kohavi, The power of decision tables, in: N. Lavrac, S. Wrobel (Eds.), Proceedings of the European Conference on Machine Learning, Lecture Notes in Artificial Intelligence, 914, Springer Verlag, Berlin, Heidelberg, New York, 1995, pp. 174–189.
[29] J.H. Larkin, H.A. Simon, Why a diagram is (sometimes) worth ten thousand words, Cognitive Sience 11 (1987) 65–99.
[30] C.-C. Lee, H. Cheng, H.-H. Cheng, An empirical study of mobile commerce in insurance industry: Task-technology fit and individual differences, Decision Support Systems 43 (2007) 95–110.
[31] H. Lucas, An experimental investigation of the use of computer based graphics in decision-making, Management Science 27 (7) (1981) 757–768.
[32] D. Martens, B. Baesens, T. Van Gestel, J. Vanthienen, Comprehensible credit scoring models using rule extraction from support vector machines, European Journal of Operational Research 183 (3) (2007) 1466–1476.
[33] D. Martens, B. Baesens, T. Van Gestel, Decompositional rule extraction from support vector machines by active learning, IEEE Transactions on Knowledge and Data Engineering 21 (2) (2009) 178–191.
[34] J. McGrath, Groups: Interaction and Performance, Prentice Hall, Englewood Cliffs, NY, 1984.
[35] M. Pazzani, Knowledge discovery from data? IEEE Intelligent Systems 15 (2) (2000) 10–13.
[36] J. Quinlan, Simplifying decision trees, International journal of man-machine studies 27 (3) (1987) 221–234.
[37] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 1993.
[38] W. Remus, A study of graphical and tabular displays and their interaction with environmental complexity, Management Science 33 (9) (1987) 1200–1204.
[39] L. Santos-Gomez, M. Darnell, Empirical evaluation of decision tables for constructing and comprehending expert systems, Knowledge Acquisition 4 (4) (1992) 427–444.
[40] R. Setiono, W.K. Leow, Pruned neural networks for regression, in: R. Miziguchi, J. Slaney (Eds.), Proceedings of the 6th Pacific Rim Conference on Artificial Intelligence, PRICAI 2000, Springer, Melbourne, Australia, 2000, pp. 500–509.
[41] R. Setiono, H. Liu, Neurolinear: From neural networks to oblique decision rules, Neural Computing 17 (1) (1997) 1–24.
[42] R. Setiono, J. Thong, An approach to generate rules from neural networks for regression problems, European Journal of Operational Research 155 (1) (2004) 239–250.
[43] T. Shaft, I. Vessey, The role of cognitive fit in the relationship between software comprehension and modification, MIS Quarterly 30 (1) (2006) 29–55.
[44] D.J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, CRC Press, 1997.
[45] A.P. Sinha, H. Zhao, Incorporating domain knowledge into data mining classifiers: An application in indirect lending, Decision Support Systems 46 (2008) 287–299.
[46] C. Speier, The influence of information presentation formats on complex task decision-making performance, International Journal of Human-Computer Studies 64 (11) (2006) 1115–1131.
[47] G. Subramanian, J. Nosek, S. Raghunathan, S. Kanitkar, A comparison of the decision table and tree, Communications of the ACM 34 (1) (1992) 89–94.
[48] J. Swait, W. Adamowicz, The influence of task complexity on consumer choice: a latent class model of decision strategy switching, Journal of Consumer Research 28 (2001) 135–148.
[49] J. Sweller, Cognitive load during problem solving: Effects on learning, Cognitive Science 12 (2) (1988) 257–285.
[50] I. Taha, J. Ghosh, Symbolic interpretation of artificial neural networks, IEEE Transactions on Knowledge and Data Engineering 11 (3) (1999) 448–463.
[51] D. Tegarden, Business information visualisation, Communications of AIS 1 (4) (1999) 1–37.
[52] E. Tufte, The Visual Display of Quantitative Information, Graphics Press, Chesire, CT, 2001.
[53] N. Umanath, I. Vessey, Multiattribute data presentation and human judgement: a cognitive fit perspective, Decision Sciences 25 (5) (1994) 795–825.
[54] J. Vanthienen, A more general comparison of the decision table and tree: A response, Communications of the ACM 37 (2) (1994) 109–113.
[55] J. Vanthienen, G. Wets, From decision tables to expert system shells, Data and Knowledge Engineering 13 (3) (1994) 265–282.
[56] J. Vanthienen, C. Mues, A. Aerts, An illustration of verification and validation in the modelling phase of KBS development, Data and Knowledge Engineering 27 (3) (1998) 337–352.
[57] I. Vessey, Cognitive fit: A theory-based analysis of the graphs versus tables literature, Decision Sciences 22 (2) (1991) 219–240.
[58] I. Vessey, The effect of information presentation on decision making: a cost-benefit analysis, Information & Management 27 (1994) 103–119.
[59] I. Vessey, R. Weber, Structured tools and conditional logic: An empirical investigation, Communications of the ACM 29 (1) (1986) 48–57.
[60] S. Viaene, R. Derrig, B. Baesens, G. Dedene, A comparison of state-of-the-art classification techniques for expert automobile insurance fraud detection, Journal of Risk and Insurance (Special Issue on Fraud Detection) 69 (3) (2002) 433–443.
[61] C. Ware, Information Visualisation: Perception for Design, Academic Press, San Diego, CA, 2000.
[62] R. Wood, Task complexity: definition of the construct, Organizational Behavior and Human Decision Processes 37 (1986) 60–82.
[63] Z.-H. Zhou, Y. Jiang, S.-F. Chen, Extracting symbolic rules from trained neural network ensembles, AI Communications 16 (1) (2003) 3–15.

Johan Huysmans received the Ph.D. degree in applied economic sciences from the Katholieke Universiteit Leuven (K.U. Leuven), Leuven, Belgium, in 2007. Currently employed for the Boston Consultancy Group as consultant, his professional interests include data mining, classification, rule extraction, and web mining.

Karel Dejaeger graduated in 2009 as business engineer at the Department of Decision Sciences and Information Management at the Katholieke Universiteit Leuven (K.U. Leuven). Currently employed as a doctoral researcher, his research interests include data mining, fraud detection and software engineering.

Christophe Mues is a lecturer (assistant professor) at the School of Management of the University of Southampton (UK). Prior to his appointment at the University of Southampton, he was employed as a researcher at K.U. Leuven (Belgium), where he obtained the degree of Doctor in Applied Economics in November 2002. He has done extensive research on decision table and diagram techniques in a variety of problem contexts, such as business rule modelling, verification and validation, and knowledge discovery and data mining. His findings have been published in various international journals and conference proceedings.

Jan Vanthienen received the Ph.D. degree in applied economics (information systems) from the Katholieke Universiteit Leuven (K.U. Leuven), Leuven, Belgium. He is a Full Professor of information systems with the Department of Decision Sciences and Information Management, K.U. Leuven. He is also Chairholder of the Pricewaterhouse-Coopers Chair on E-Business at K.U. Leuven. He is the author or coauthor of numerous papers published in international journals and conference proceedings. His current research interests include information and knowledge management, business intelligence and business rules, and information systems analysis and design.

Bart Baesens is an associate professor at K.U. Leuven (Belgium), and a lecturer at the University of Southampton (United Kingdom). He has done extensive research on predictive analytics, data mining, customer relationship management, fraud detection, and credit risk management. His findings have been published in well-known international journals and presented at international top conferences. He is also co-author of the book Credit Risk Management: Basic Concepts, published in 2008.