Expert Systems With Applications 62 (2016) 333–346

Contents lists available at ScienceDirect

Expert Systems With Applications

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

What makes classification trees comprehensible?

Rok Piltaver[a,b,*], Mitja Luštrek[a,b], Matjaž Gams[a,b], Sanda Martinčić-Ipšić[c]

[a] Department of Intelligent Systems – Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
[b] Jožef Stefan International Postgraduate School, Jamova cesta 39, Ljubljana 1000, Slovenia
[c] Department of Informatics – University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia

ARTICLE INFO

Article history:
Received 13 January 2016
Revised 7 June 2016
Accepted 7 June 2016
Available online 16 June 2016

Keywords:
Classification tree
Comprehensibility
Understandability
Interpretability
End-user survey

ABSTRACT

Classification trees are attractive for practical applications because of their comprehensibility. However, the literature on the parameters that influence their comprehensibility and usability is scarce. This paper systematically investigates how tree structure parameters (the number of leaves, branching factor, tree depth) and visualisation properties influence the tree comprehensibility. In addition, we analyse the influence of the question depth (the depth of the deepest leaf that is required when answering a question about a classification tree), which turns out to be the most important parameter, even though it is usually overlooked. The analysis is based on empirical data that is obtained using a carefully designed survey with 98 questions answered by 69 respondents. The paper evaluates several tree-comprehensibility metrics and proposes two new metrics (the weighted sum of the depths of leaves and the weighted sum of the branching factors on the paths from the root to the leaves) that are supported by the survey results. The main advantage of the new comprehensibility metrics is that they consider the semantics of the tree in addition to the tree structure itself.

1. Introduction

Classifier comprehensibility, which is sometimes referred to as interpretability (Freitas, 2014; Huysmans, Dejaeger, Mues, Vanthienen & Baesens, 2011; Jin & Sendhoff, 2008; Jin, Sendhoff, & Körner, 2005; Maimon & Rokach, 2005b) or understandability (Allahyari & Lavesson, 2011; Pazzani, 2000; Sommer, 1995), is defined as "the ability to understand the logic behind a prediction of a model" (Martens, Vanthienen, Verbeke, & Baesens, 2011) or "how well humans grasp the induced classifier" (Maimon & Rokach, 2005a). It has been recognised as an important classifier property since the 1980 s (Michalski, 1983) and is continuously emphasised (Allahyari & Lavesson, 2011; Freitas, 2014; Huysmans et al., 2011; Martens et al., 2011; Sommer, 1995; Zhou, 2005; Quinlan, 1999). For example, one of the main features of ID3-like algorithms is their ability to generate easy-to-understand decision trees (Michie, 1987). Similarly, Kodratoff (1994) recognises the comprehensibility as a decisive factor when machine learning models are applied in the industry. Comprehensible classifiers are especially important in domains such as credit scoring, medicine, churn prediction and bioinformatics (Freitas, 2014) because they enable domain experts to classify instances without using a computer, explain classifications of individual instances, validate the classifier, confirm hypotheses, discover new knowledge and improve or refine classifiers in collaboration with data-mining experts (Craven & Shavlik, 1995; Zhou, 2005).

Classifier comprehensibility depends on the type of knowledge representation that is employed (Freitas, 2014; Huysmans et al., 2011; Johansson, Niklasson, & König, 2004; Martens et al., 2011; Zhou, 2005). For example, classification trees and rules are considered to be the most comprehensible (Freitas, 2014; Johansson et al., 2004; Martens et al., 2011; Zhou, 2005), while support vector machines, artificial neural networks and ensembles of classifiers are considered to be the least comprehensible (Chorowski, 2012) and, hence, are termed black-box classifiers (Freitas, 2014; Huysmans et al., 2011; Johansson et al., 2004; Zhou, 2005). There are differences in the comprehensibility of the classifiers based on the same type of knowledge representation as well (Martens et al., 2011): the complexity of a specific classifier (measured as the number of leaves in a tree (Maimon & Rokach, 2005a), conditions in a classification rule set (Sommer, 1995), or connections in a neural network (Jin & Sendhoff, 2008; Liu & Kadirkamanathan, 1995)) is often used as a surrogate metric for classifier comprehensibility (Allahyari & Lavesson, 2011; Freitas, 2003; Freitas, 2004; Jin & Sendhoff, 2008; Jin et al., 2005; Johansson et al., 2004; Martens et al., 2011); a lower complexity corresponds to a higher comprehensibility. However, other properties, such as the structure of the model and its visualisation, affect the comprehensibility as well

* Corresponding author.
E-mail addresses: rok.piltaver@ijs.si (R. Piltaver), mitja.lustrek@ijs.si (M. Luštrek), matjaz.gams@ijs.si (M. Gams), smarti@inf.uniri.hr (S. Martinčić-Ipšić).

http://dx.doi.org/10.1016/j.eswa.2016.06.009
0957-4174/

(Göpferich, 2009; Huysmans et al., 2011), but it is not clear how and to what extent. Therefore, the main problem with regard to most classification algorithms is that they do not explicitly consider the comprehensibility (Huysmans et al., 2011, Johansson et al., 2004), while the ones that do usually simplify it to the classifier complexity (Pazzani, 2000). This approach has several drawbacks (Freitas, 2014) and could lead to over-simplistic models (Elomaa, 1994) that are neither accurate nor comprehensible. This consideration is the motivation for our systematic empirical study of tree properties that potentially influence the comprehensibility of classifiers, in which we tackle classification trees, which are probably the most commonly used type of comprehensible classifiers.

We analyse the comprehensibility through the lens of classifier usability, which is actually the property that is important in practice: the easier a classifier is to comprehend, the easier it is to use. Therefore, the classifier comprehensibility and usability can be interchanged in this paper, although in general, the terms are not exact synonyms (Göpferich, 2009). This study is based on data about the performance of users while solving four types of tasks that involve classification trees and their opinions on the task difficulty and the tree comprehensibility, which was obtained using a carefully designed survey. We collected the answers to 98 questions from 69 respondents and analysed them with statistically sound methodology; we provide the interpretation of the results as well as several empirically supported guidelines on how to construct more comprehensible classification trees. We focus mainly on the influence of the tree structure properties (the number of leaves, branching factor, tree depth) on the comprehensibility, but we also analyse the influence of several tree visualisation properties. One of the most important contributions of this study is the investigation of the influence of the question depth, which is equal to the depth of the deepest leaf that is required to answer a question about a classification tree. Another improvement over the related work is the comparison of the performance and opinions about the tree comprehensibility, from novice versus expert data-miners. Finally, we propose two new classification-tree comprehensibility metrics (the weighted sum of the depths of the leaves and the weighted sum of the branching factors on the paths from the root to the leaves). Comprehensibility metrics are required to act as heuristic functions in learning algorithms (Giraud-Carrier, 1998; Piltaver, Luštrek, Zupančič, Džeroski, & Gams, 2014) and to compare the comprehensibility of the classifiers obtained from various algorithms (Piltaver, Luštrek, Zupančič, et al., 2014; Zhou, 2005).

The paper begins with a review of related work. Section 3 explains the survey design and implementation by listing the general design choices, survey bias prevention strategies, analysed classification tree properties, methods for generating the classification trees used in the survey and survey question examples. Section 4 presents and discusses the survey results. First, the survey and survey respondents are described, followed by a discussion on the performance of different survey respondent groups, the influence of the classification tree parameters on the comprehensibility for each of the survey tasks, and the influence of the classification tree visualisation. The paper closes with a summary of the most interesting findings and suggested directions for further research.

2. Related work

Although many papers emphasise the importance of classifier comprehensibility (Freitas, 2014; Kodratoff, 1994; Martens et al., 2011; Michalski, 1983; Michie, 1987; Quinlan, 1999; Sommer, 1995; Zhou, 2005), related work on classifier comprehensibility is relatively scarce (Allahyari & Lavesson, 2011; Pazzani, 2000). The most general related work comes from the field of cognitive science. Cognitive load theory (Sweller, 1988) divides the total amount of mental effort that is used in working memory into three types: the intrinsic cognitive load is inherent to the specific topic and cannot be altered (the complexity of the classification domain); the extraneous cognitive load depends on the way that information or tasks are presented (the classifier representation); and the germane cognitive load is devoted to the processing and construction of mental structures that organise the categories of information and their relationships. Research in this field has developed a way of measuring the perceived mental effort (Paas & Van Merriënboer, 1993), which motivated us to approach the analysis of the classifier comprehensibility with objective measures. Furthermore, it was shown that experience with a specific task reduces the cognitive load, while the lack of it increases the load (Murphy & Wright, 1984). This concept is addressed in our study by comparing the performance of two groups: data-mining experts (as suggested in Freitas (2014)) and novice data-miners.

More specific studies come from the field of text comprehensibility, where numerous methods for determining comprehensibility have been devised (Göpferich, 2009). Schriver (1989) divides them into three groups and concludes that reader-focused approaches provide advantages over text-focused and expert-judgment-focused approaches. In line with this result, we perform an empirical study that is based on a user survey instead of simply measuring the model complexity, as in Allahyari and Lavesson (2011), Freitas (2003), Freitas (2004), Jin and Sendhoff (2008), Jin et al. (2005) and Martens et al. (2011) or using expert-judgements, as in Freitas (2014).

In the IT field, there are a considerable number of empirical studies that investigate the understandability of conceptual models. Our survey design builds upon the following design issues, which are summarised in a review of experiments from this field (Houy, Fettke, & Loos, 2012): the research design, the number of experiment participants, and the observed dependent variables. In addition, our study follows the framework for empirical evaluation of model comprehensibility (Aranda, Ernst, Horkoff, & Easterbrook, 2007) in all of the recommendations, which can be applied to the classifier comprehensibility.

Finally, a few studies address specifically the classification-tree comprehensibility. Freitas (2014) reviews the case for comprehensible classifier models and discusses the advantages and drawbacks of five types of classification knowledge representations, including classification trees. This work motivated us to study the influence of the question depth on the tree comprehensibility, which has not been empirically evaluated before.

The work by Allahyari and Lavesson (2011) is probably the first empirical study of classification-model comprehensibility. This study compares the comprehensibility of classifiers that are learned by three classification-tree and three rule-learning algorithms, based on subjective comparisons of classifier pairs by 50 students. We focus on classification trees because they are more comprehensible than classification rules (Allahyari & Lavesson, 2011). Furthermore, one of our survey tasks follows the design of their study, comparing classifier pairs. They also report that the classifier complexity has a negative correlation with the understandability, and therefore, we extend their work by analysing the influence of several other tree structure parameters. We also improve on their work by using objective measures and additional survey tasks, including data-mining experts as survey respondents, analysing larger trees and using a domain that is familiar to the respondents.

The study by Huysmans et al. (2011) empirically evaluates the comprehensibility of decision tables, trees and rules using subjective opinions (answer confidence) and objective measures of the respondent performance (time to answer and answer accuracy) for three tasks: classify an instance, verify whether a classifier agrees with a statement, and compare two classifiers for their equivalence. The results of this study are in favour of the single-hit de-

cision tables over the other representation formats, but they are obtained exclusively from business students who have no prior experience with any of the representation formats and a domain with binary class. Our survey includes both objective measures as well as classification and verification tasks. In contrast with their work, we analyse the influence of the tree properties (and not the type of classifier knowledge representation) on the comprehensibility. Moreover, we extend the survey with additional tasks, with questions about the perceived comprehensibility, and by analysing the influence of classifier visualisation and the user's background in data mining.

In our previous work, we designed a survey (Piltaver, Luštrek, Gams, & Martinčić – Ipšić, 2014a) according to experience from related work (Allahyari and Lavesson, 2011; Aranda et al., 2007; Freitas, 2014; Houy et al., 2012; Huysmans et al., 2011); we validated it with a group of data-mining experts and psychologists, implemented it as an online survey (Piltaver et al., 2014a; Piltaver, Luštrek, Gams, & Martinčić – Ipšić, 2014b) and tested it with respondents (Piltaver et al., 2014b). This paper improves the initial survey design based on the results of the validation and thoroughly discusses the design choices and bias prevention strategies that are crucial for the validity of the results. Finally, we conduct a statistical analysis of the obtained data, propose the question depth as the novel parameter that influences the comprehensibility and introduce two new tree-comprehensibility metrics.

3. Survey design and implementation

The following four subsections describe the survey design and implementation. The first subsection presents the general design choices, and the second is dedicated to the prevention of the bias (i.e., avoiding influencing the results of the survey by its design). The last two subsections present the range of the classification trees that are used in the survey according to the tree structure and visualisation properties.

3.1. General design

Comprehensibility is inherently subjective (Huysmans et al., 2011; Martens et al., 2011) and hence impossible to measure directly (Allahyari & Lavesson, 2011). Therefore, we measure it indirectly by objective measures of the survey respondents' performance. The decision is based on research that argues that the complexity impacts the task performance (Campbell, 1988). We also collect the respondents' subjective opinions about the tree comprehensibility and the question difficulty.

The first design choice is about the objective measures: the respondents' performance is measured in terms of the time that is required to answer a survey question (hereinafter **time-to-answer**) and the probability of the correct answer (hereinafter **answer-correctness**), which is consistent with the methodology that is used in the related work (Aranda et al., 2007; Houy et al., 2012; Huysmans et al., 2011). In general, the shorter time-to-answer and the higher answer-correctness correspond to easier questions and, therefore, to more comprehensible trees. The survey includes questions about a range of trees that have various structure and visualisation properties to objectively quantify the influence of each property on the tree comprehensibility. In addition, the respondents are asked for their subjective opinion about the tree comprehensibility and the difficulty of each question (hereinafter **question-difficulty**). Subjective opinions are used as the gold standard for the evaluation of the tree comprehensibility metrics, to estimate the variability of the respondents' subjective opinions, and to verify the correlations between the subjective opinions and the objective performance measures.

The second choice is about the evaluation: paired statistical tests are used to compare the performance and subjective opinions of each respondent on different trees. Paired tests are required because of the limited number of respondents and the high variability of their performance (see Fig. 8) and subjective opinions. Hence, each respondent is required to answer all of the questions, and the number of questions must be limited (we set the upper bound to 100) because it is difficult to motivate the respondents to answer a lengthy survey and because answering too many questions could influence respondent's performance and expressed subjective opinions.

The third design choice involves the set of analysed tree parameters: we consider tree-structure parameters, tree-visualisation properties and the question depth. The analysed **tree structure parameters** can be computed algorithmically and in turn used for estimation of the tree comprehensibility. We analyse the following parameters, for the following reasons:

- The number of leaves (i.e., nodes without child vertices) – often used as a tree complexity metric that approximates the tree comprehensibility (Freitas, 2003; Martens et al., 2011);
- The branching factor (i.e., the number of child nodes in the inner vertices) – many tree-learning algorithms enable learning binary trees (Demšar et al., 2013) and several heuristic functions for choosing the node splitting attribute take into account the branching factor (Harris, 2001; Quinlan, 1993).
- The tree depth (the length of the longest path from the root to a leaf) – a well-defined tree complexity metric, which depends on the number of leaves and the branching factor.

We analysed the **tree visualisation properties** because they could influence the comprehensibility (Göpferich, 2009) and to obtain data from which we empirically derived recommendations on visualising classification trees in a comprehensible way. We considered the following properties:

- the visualisation style (plain-text, Weka (Hall et al., 2009) and Orange (Demšar et al., 2013) visualisation styles);
- explicitly visualised information about the tree (text and colour-coded information);
- the tree layout.

Finally, motivated by our previous research (Piltaver, Luštrek, & Gams, 2014; Piltaver, Luštrek, Zupančič, et al., 2014) and related work (Freitas, 2014; Sommer, 1995), we investigate the influence of the **question depth**, which is equal to the depth of the deepest leaf that is required to answer a question about a tree.

The survey is divided into two parts. The first part includes four tasks (*classify, explain, validate*, and *discover*), which are dedicated to quantifying the influence of the tree-structure parameters and the question depth on the tree comprehensibility based on the objective measurements and respondents' subjective opinions (Table 1). Different tasks are used because they can generate different evaluation perspectives (Aranda et al., 2007). A uniform tree visualisation style is used throughout the first part of the survey (as suggested by Allahyari and Lavesson (2011)), while the tree structure parameters and question depth are varied systematically. The second part includes two tasks (*rate* and *compare*) that are dedicated to the respondents' subjective opinions about the influence of both the tree structure and the visualisation properties (Table 1) on the comprehensibility. A thorough description of the survey tasks (their design and implementation) is in Piltaver et al. (2014a) and Piltaver et al. (2014b) and is here briefly recapitulated (see supplementary material Section 3 for question examples):

- in the *classify tasks*, the respondents classify an instance that is represented with an attribute-value table using a classification tree shown on the screen;

Table 1
The investigated tree parameters and used measures for each survey task.

Task	Measures (objective, *subjective*)	Varied tree parameters (tree structure, *tree visualisation*)	Values
Classify, *explain*, *validate*, *discover*	Time-to-answer, Answer-correctness, *Question-difficulty*	Number of leaves Branching factor Question depth	[3–11] [2–4] [1–7]
Rate	*Absolute comprehensibility*	Number of leaves Branching factor	[3–10] [2–4]
Compare	*Relative comprehensibility*	Number of leaves Branching factor Tree depth *Visualisation style* *Additional visualised information* *Tree layout*	(4 vs. 10) (2 vs. 3) (2 vs. 4) (Plain-text, Weka and Orange default) (Pie charts, meaningful attribute names) (Ordered vs. random)

- in the *explain task*, the respondents are asked which attribute values would a) have to stay the same, b) have to be changed or c) are irrelevant in order to classify an instance into a given class that is different from the class to which the instance currently belongs (according to a classification tree shown on the screen);
- in the *validate task*, the respondents check whether a classification tree agrees with a statement about the classification domain (e.g., "Does the classification tree agree with the following statement: for animals from class *amphibian* it holds that *backbone* = yes and *breathes* = no?");
- in the *discover task*, the respondents discover which property is unusual for a given class (e.g., flightless bird) based on a classification tree that divides the majority of instances that belong to the class into one leaf and the minority with the unusual property into another leaf (e.g., "Which is the rare property for the animals from class mammal?", to which the respondents should reply with an attribute name and its value);
- in the *rate* task, the respondents rate how comprehensible a tree is on a five-level scale defined in Piltaver et al. (2014b). They are asked to rate trees that have a uniform visualisation style and systematically varied tree-structure parameters (note that the question depth cannot be varied in this task because the comprehensibility of the entire tree is rated);
- in the *compare task* designed as in Allahyari and Lavesson (2011), the respondents are asked to compare the comprehensibility of two trees shown side by side using a scale defined in Piltaver et al. (2014b): we keep the visualisation style fixed and alter a single tree-structure property or keep the tree structure fixed and alter the visualisation style.

To conduct the survey, we implemented a custom on-line survey engine (Piltaver et al., 2014a; Piltaver et al., 2014b), which allows remote participation at a time that is convenient for the respondents. This approach enables accurate measurements of time-to-answer, the recording of all of the respondents' actions (e.g., correcting a wrong answer), displaying data about the respondent's performance (used as a motivation tool), simple translation of the survey into different languages and real-time progress monitoring, which was especially useful in the testing phase. More details on the survey implementation are available in the supplementary material, and the survey itself is available at http://dis.ijs.si/classifier-comprehensibility-survey.

3.2. Preventing survey bias

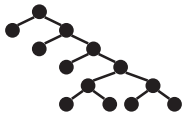
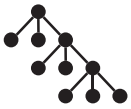
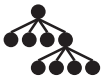
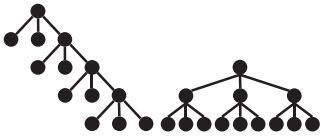
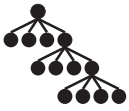
The survey results could be biased by the **classification domain** because the respondents perform better when answering questions about familiar domains compared to unfamiliar domains (Aranda et al., 2007; Martens et al., 2011; Sweller, 1988). This concern was prevented by posing questions about a single domain that is familiar to all of the respondents. The Zoo dataset from the UCI machine-learning repository (Bache & Lichman, 2013) was used because it fits the requirement (i.e., requires only elementary biology knowledge) and enables the construction of a large range of trees (see Table 2). This dataset consists of 101 instances, each representing an animal (e.g., antelope, crow, frog) that belongs to one of 7 classes (*mammal, bird, fish, amphibian, reptile, mollusc* or *insect*) and is described with 15 binary (*hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic, cat-sized*) and 1 numeric (*legs*) attribute.

The second potential source of bias is the **learning effect** (Vessey & Galletta, 1991): the respondents learn while answering the survey, and therefore, they perform better at questions that are posed later in the survey compared with the initial questions. For example, novice data-miners learn how to use classification trees efficiently (e.g., use the mouse pointer to mark the current node while looking up the value of the node attribute). On the other hand, data-mining experts might need a few initial questions to become accustomed to the classification-tree visualisation style and the survey answering form. The learning effect was noticed while testing the initial version of the survey, and therefore, several precautions were taken:

- The questions are posed in six groups; each corresponds to one of the six tasks. All of the questions in a task are posed in exactly the same way: the phrasing of the question, alphabetical order of the attributes in an attribute-value table that contains exactly 10 attributes, visualisation style and layout of the classification tree, and the answer forms are the same. These precautions shorten the learning period.
- Instructions on how to answer an example question are provided before each task, and a test question must be answered correctly before the respondent is allowed to proceed to the first question (similar to Vessey and Galletta (1991)). This approach verifies that the respondent understands the question and knows how to answer it.
- The tasks are ordered from the simplest to the most difficult, to diminish the learning effect.
- The Latin-square design (Box, Hunter, & Hunter, 2005; Vessey & Galletta, 1991) is used for the question order within each task. This arrangement means that the respondents answer the questions in different orders and that the frequency of posing each question at any given position in the question ordering is constant. This approach distributes the remaining learning effect equally over all of the questions, and therefore, the average results over all of the respondents are not biased by the learning effect.

The respondents' subjective opinions are biased by their mental abilities (Sweller, 1988) and **data-mining experience** (Murphy & Wright, 1984) and could have a high variability due to the subjective interpretations of the comprehensibility rating scales. To

Table 2
The structure of the trees used in the survey.

Number of leaves	Branching factor 2	3	4
3 or 4			
7			
9 or 10			

prevent this type of bias, the two tasks that ask for subjective opinions are placed at the end of the survey. In this way, novice data-miners obtain experience on which they can soundly base their subjective opinions. Second, the offered answers to the multiple choice questions are textual (instead of numeric), and their meaning is explained with at least one sentence and one example (Piltaver et al., 2014b). This approach mitigates subjective interpretations of the answers' meaning. For example, the explanation that says: "*I can use the knowledge represented by the classification tree as soon as I read it for the first time; I can easily remember it; I can explain it to another person without looking at the figure.*" leaves fewer interpretation possibilities than a mark 2 on a 1 to 5 scale. Finally, the respondents freely select the survey language (English, Slovenian or Croatian), which prevents decreased performance due to the use of a foreign language.

Finally, the survey results can be biased by the **respondent's motivation**: well-motivated respondents perform better and rate questions as easier and trees as more comprehensible compared to non-motivated respondents. Instead of offering monetary awards (Huysmans et al., 2011; Vessey & Galletta, 1991), we used a gamification approach (Zichermann & Cunningham, 2011): a table that compares the performance of the respondent to the best, average and worst respondent is shown after finishing each task. According to the post-survey interviews, the comparison table turned the survey into a competition and motivated the respondents to answer quickly and correctly.

3.3. Varying the tree structure

We systematically varied the investigated tree structure parameters, as illustrated in Table 2, while keeping the other factors that might influence the respondents' answers unchanged. This approach enables analysing the influence of a single tree structure parameter on the comprehensibility. The size of the trees is limited to assure text readability. Note that this sizing limitation is not a drawback because large trees are rarely used by humans in practice. The tree construction process is described briefly below, while the details and figures of all of the trees used in the survey are given in Piltaver et al. (2014b).

The basic classification tree (Fig. 1) was learned using the default Orange (Demšar et al., 2013) parameters, and larger trees were obtained by changing the pruning parameters. The smaller trees (whose leaves correspond to instance clusters) were constructed manually because pruning resulted in unnatural trees that could cause survey bias (Göpferich, 2009; Martens et al., 2011).

To construct trees with a branching factor that is above 2, additional attributes were computed as Cartesian products of the original attributes. For example, combining the attributes *aquatic* and *breathes* produces a ternary attribute with the values: *aquatic-does not breathe* (e.g., *fish*), *aquatic-breathes* (e.g., *amphibian*) and *terrestrial*. As a result, the same or at least very similar structure as in the original binary tree was obtained (e.g., Fig. 2).

For the *discover* task, special trees were constructed: they split the majority of instances that belong to a class (e.g., *bird)* into one leaf and the minority with an unusual property (e.g., *not airborne)* into another leaf (Fig. 3). The limited number of outliers in the dataset resulted in 8 questions about 4 binary trees with 3, 5, 7 and 11 leaves.

3.4. Varying the tree visualisation

The influence of the tree visualisation is analysed in the *compare* task. The goal is to determine which tree visualisation parameters influence the comprehensibility and to assess their relative importance; therefore, only a single question per visualisation parameter is posed.

First, the plain-text tree visualisation (Fig. 4 left) and Weka (Hall et al., 2009) default visualisation styles (Fig. 4 right) were compared with the Orange default visualisation style (Figs. 1–3).

Two options with regard to the information shown in the tree visualisation were considered: classification trees with (Fig. 5a) and without pie charts (Fig. 5c) and trees with (Fig. 5a) and without meaningful attribute names, attribute values and class names (Fig. 5b).

Finally, the importance of the tree layout was evaluated by comparing the comprehensibility of a tree with a random layout (Fig. 6b) and a tree with a left-to-right layout (Fig. 6a and Table 2). The layout is obtained by placing the shallower subtrees on the left and the deeper subtrees on the right of the parent node. The advantage is that this layout positions the simpler instances on the left side of the tree figure, which is commonly read first (in all three survey languages) and makes the tree structure systematic.

4. Results and discussion

This section presents the survey results and discusses the influence of the classification trees' properties on their comprehensibility. General information about the survey questions, tasks, the

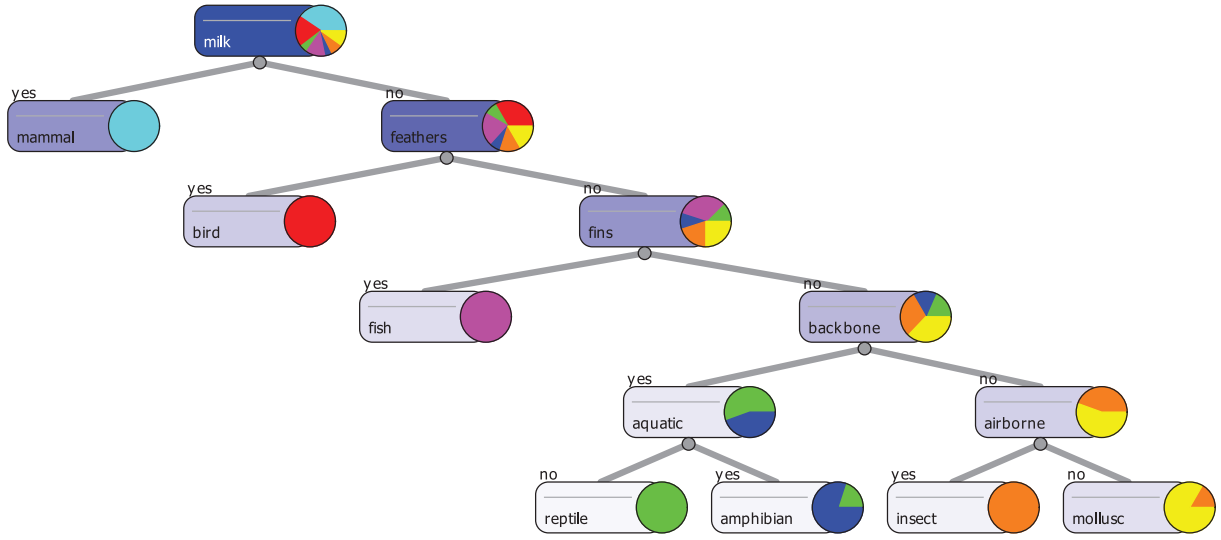


Fig. 1. The basic classification tree.

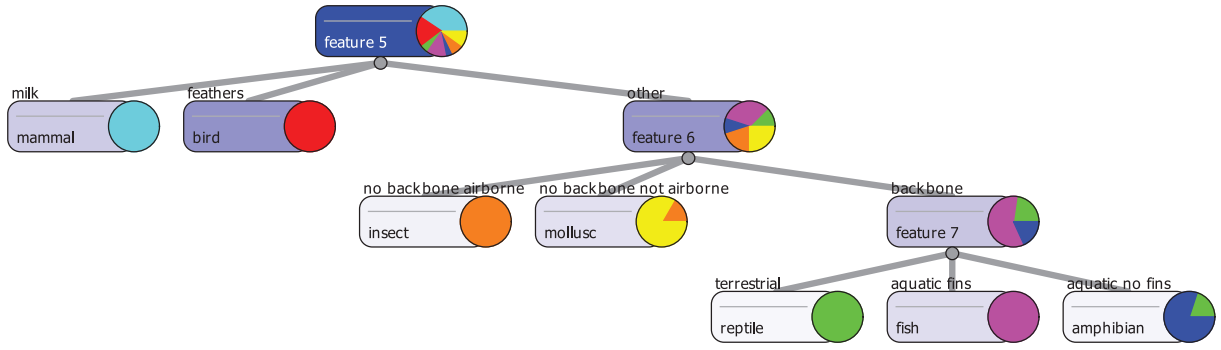


Fig. 2. Tree with branching factor 3 corresponds to the tree shown in Fig. 1.

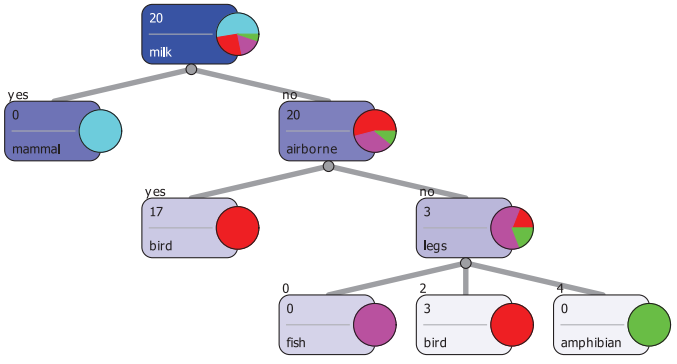


Fig. 3. The tree shows an unusual property for a *bird: not airborne*. The numbers in the nodes correspond to the number of training instances from the class *bird* that belong to each node.

respondents and their performances are given in the first two subsections, followed by subsections with the results about the influence of the tree structure parameters and the question depth, the influence of the visualisation properties, and the comprehensibility metrics.

4.1. General data about the survey

The survey includes 98 questions: 80% of the questions are about the influence of the tree structure parameters on their comprehensibility, and 20% are on subjective opinions about the tree comprehensibility. The data about the size of the survey are presented in Table 3. The *classify* task included one question for each possible combination of the tree size, branching factor and question depth that was available, using the trees in Table 2. The number of questions in the subsequent tasks was reduced to optimize the trade-off between the number of questions and the number of possible comparisons for the analysis. The *compare* task included one question for each analysed tree property listed in Table 1. The *rate* task included one question for each classification tree illustrated in Table 2. On average, the respondents needed ~29 minutes to answer the questions plus ~26 minutes to rate the question difficulty and tree comprehensibility.

Fig. 7 shows the median time-to-answer and the mean question-difficulty (mapped to integers that range from 1 to 5) for each question (represented with a circle). The questions in the *classify* task were rated as the easiest (the mean question-difficulty over all of the questions 1.6) and the answer times are short compared to the other tasks. The most difficult was the *discover* task: the mean question-difficulty over all of the questions is 2.4, and the median time-to-answer for the easiest question is 14.5, which is the highest among the four tasks. The *explain* task required more time to answer because the number of clicks needed to answer a question was equal to the depth of the question, while a single mouse click was needed in the other tasks. Fig. 7 shows a high correlation between the question-difficulty and time-to-answer, which confirms that the time-to-answer is a suitable objective measure and that the difficulty scale was well-designed.

Table 3
The statistical data about the size of the survey.

Task	*Classify*	*Explain*	*Verify*	*Discover*	*Compare*	*Rate*	Total
Number of questions	30	18	23	8	8	11	98
Number of respondents	69	52	52	52	52	52	
Number of answers	2070	936	1196	416	416	572	5606
Mean time-to-answer per respondent [s]	17.8	27.3	15.9	25.9	/	/	
Total time-to-answer for all respondents [h]	10.2	7.1	5.3	3.0	/	/	25.6+~24
Answer-correctness [%]	97.6	87.7	96.0	72.8	/	/	93.0

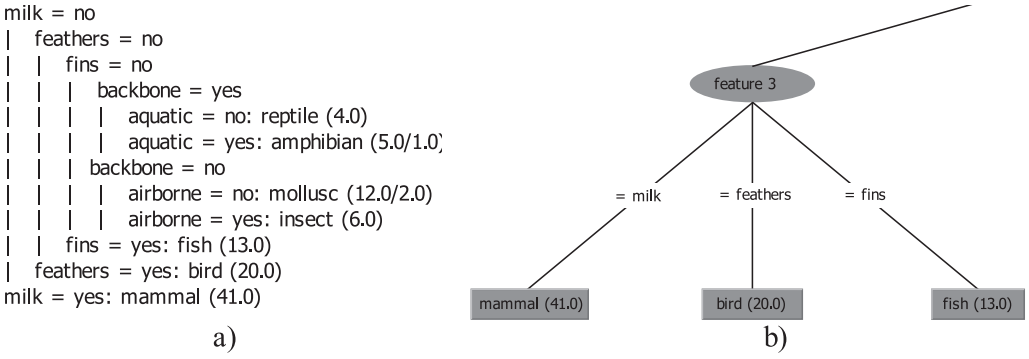


Fig. 4. A tree plotted using the plain-text visualisation (a) and a branch of a tree with the Weka visualisation style (b).

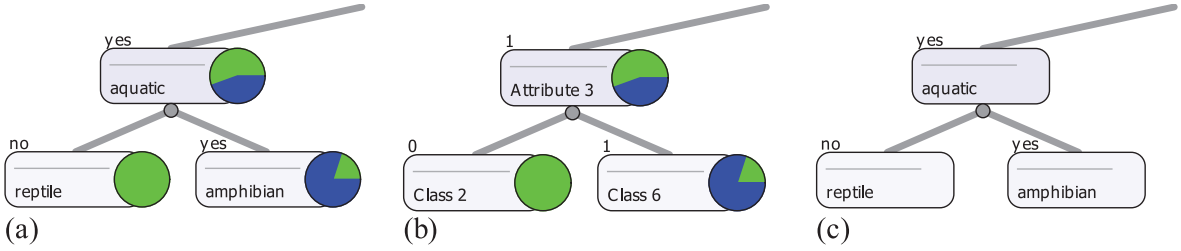


Fig. 5. Three subtrees that illustrate the information given in each node: the default (a), with meaningless names of attributes/class (b) and without pie charts (c).

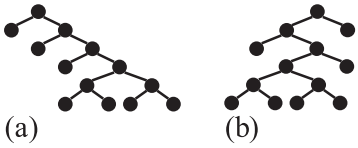


Fig. 6. Two trees with the same structure but different layouts: left-to-right (a) and random (b) order.

High correlations between the answer-correctness and time-to-answer (or question-difficulty) were also observed. Graphs (see supplementary material Section 4) show that the answer-correctness is constant for the questions that have a low question-difficulty or time-to-answer. The answer-correctness starts decreasing when the time-to-answer increases above 10 s or when the question difficulty increases above easy. In the *classify* task, the answer-correctness remains constant because the questions are so easy that the incorrect answers are due to random errors and not due to higher question-difficulty or lower tree comprehensibility.

4.2. Survey respondents and the performance of the respondent groups

The survey respondents are data-mining and artificial-intelligence researchers from the Jožef Stefan Institute, Slovenia and faculty staff and students from the Department of Informatics – University of Rijeka, Croatia. Because previous studies have indicated that past experience, education and individual cognitive abilities can influence the task performance (Aranda et al., 2007; Benbasat & Taylor, 1982; Freitas, 2014; Huysmans et al., 2011; Lee, Cheng, & Cheng, 2007; Murphy & Wright, 1984; Vessey & Galletta, 1991), the respondents are divided into three groups according to their experience with classification trees: data-mining *experts* (PhD in data mining), computer science and informatics *students* (BsC in informatics) who took at least an introductory data-mining course, and other *IT specialists* (MsC or PhD in computer science) with very limited or no experience with classification trees. In total, 52 respondents answered all of the questions: 26 *students*, 19 *experts*, and 7 *IT specialists*, of whom 62% were male. An additional 17 respondents answered only the *classify* task: 15 *students*, one *expert* and one *IT specialist*. The numbers are comparable with most of the studies that are reviewed in Houy et al. (2012). More details on the respondents' demographic data are available in the supplementary material.

The respondent sample enables a comparison of the performance of the *experts* and *students*. In the *classify* task, the *experts* achieve the shortest median time-to-answer (13 s), the lowest average question-difficulty (1.4) and the highest answer-correctness (99%) among the three respondent groups. The *students* perform the worst: the median time-to-answer is approximately 2 s longer and the probability of an incorrect answer is 3 times higher compared to the *experts*. Nevertheless, the best performing *students* are on a level with some of the best *experts*.

The comparison of the performance between the respondent groups in the *explain* task is the same as in the *classify* task, with a few differences. The average answer-correctness of the *experts* is 96% and of the *students* is 83% (this task is more difficult than

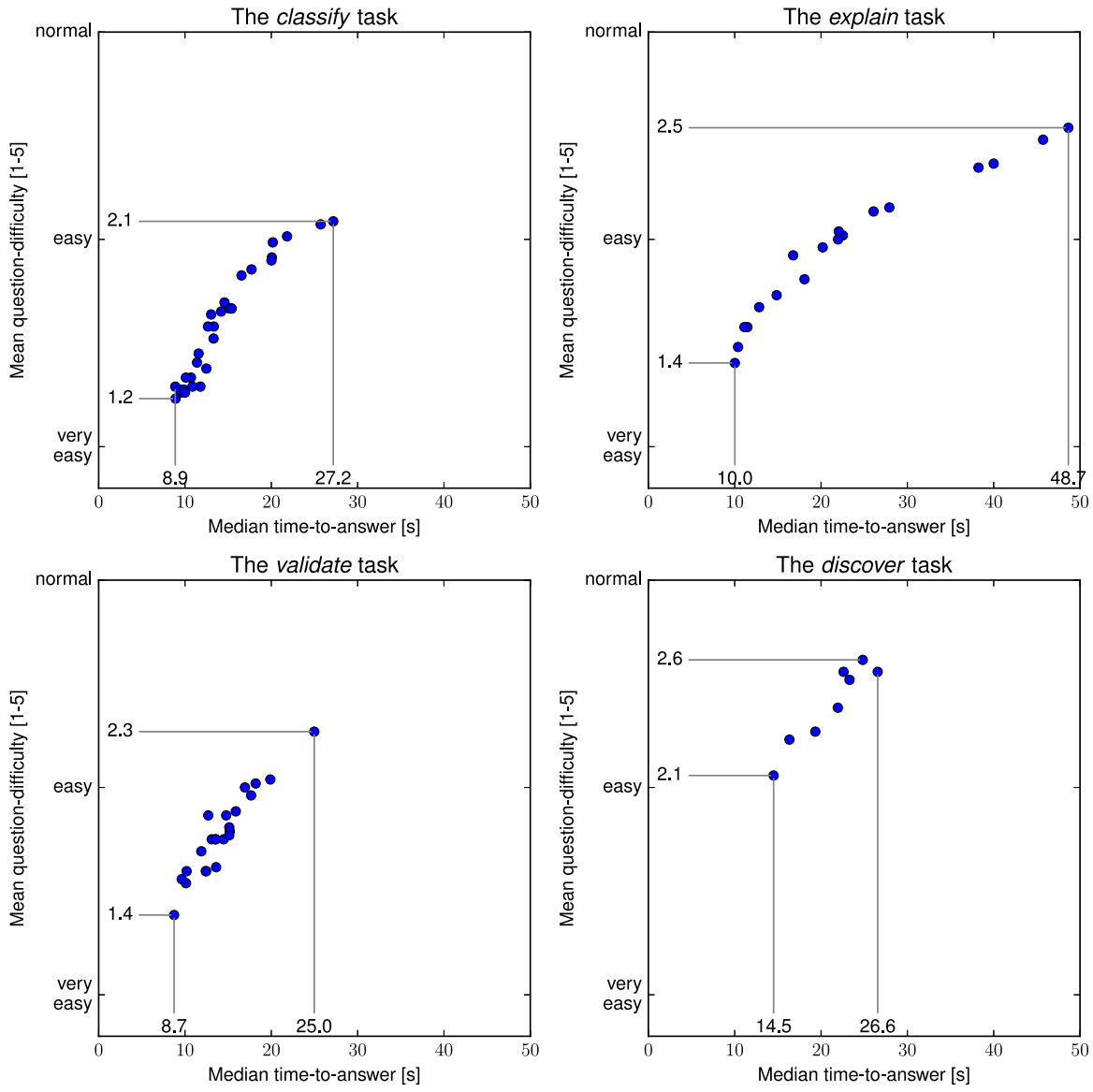


Fig. 7. Median time-to-answer and the mean question-difficulty for the questions in the first part of the survey.

the *classify* task). The difference between the two groups is much greater in the *explain* task in both relative and absolute terms. In the *verify* task, *experts* and *students* perform equally well according to the time-to-answer, answer-correctness and question-difficulty.

In the *discover* task, the *experts* again perform best: the median time-to-answer (~20 s) and its variability are the smallest, the question-difficulty (2.2) the lowest (but comparable to *students*), and the answer-correctness the highest (89%). The *students* achieve a slightly longer median time-to-answer (~23 s) but a much lower answer-correctness (64%). The basic knowledge about classification trees does not suffice in the *discover* task; therefore, the *experts* have a larger advantage than in the previous three tasks. The difference in the performance of the respondent groups indicates that the results of the surveys performed only with students could be biased.

4.3. The influence of the tree structure parameters and the question depth

The influence of the tree structure parameters on the respondents' performance was analysed using the Wilcoxon signed-rank test for time-to-answer and question-difficulty and McNemar's test for answer-correctness as well as graphs that show summary statistics (e.g., Fig. 8). Paired tests were used to account for the differences in the performances between the respondents. We tested whether increasing a single parameter by one step influences a performance metric while the two remaining parameters were fixed. The Holm-Bonferroni correction for multiple comparisons was applied to check for statistical significance.

Fig. 8 shows the influence of the question depth on the time-to-answer. Each subgraph shows a Tukey boxplot for questions with respect to a given branching factor and number of leaves. It is not possible to obtain trees that have some combinations of tree structure parameter values (e.g., a tree with 3 leaves and a branching factor of 4), and therefore, the trees that have the most similar number of leaves were used instead (4 or 10 leaves instead of 3 or 9, respectively). The time-to-answer has no upper bound, which makes it sensitive to outliers (not shown in the figure), and hence, the median and quartiles are observed instead of the mean. The annotations represent the raw p-values of the Wilcoxon signed-rank tests (a non-parametric version of the paired t-test). Fig. 8 shows that increasing the question depth increases the time-to-answer.

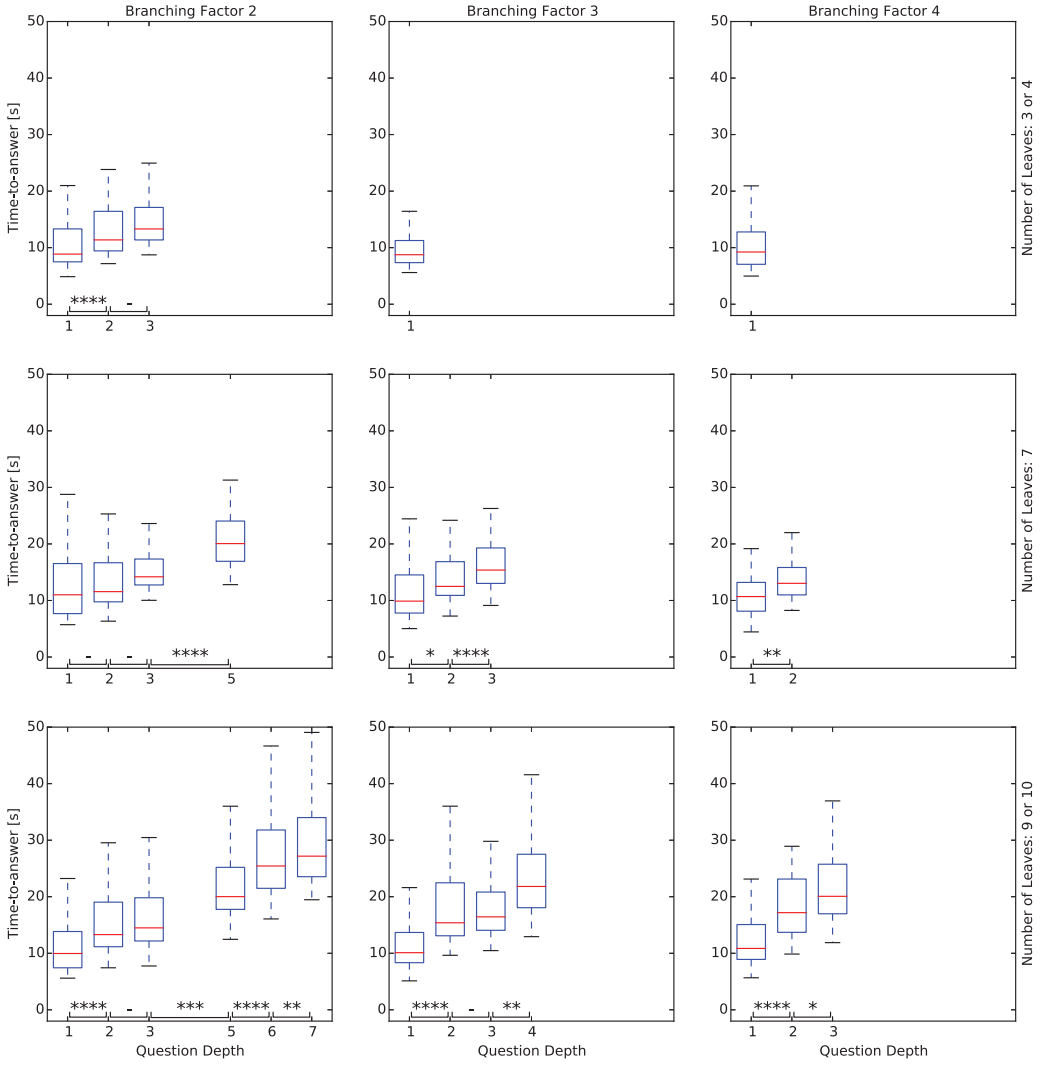


Fig. 8. The influence of the question depth on the time-to-answer a classification question about a tree that has a given branching factor (column) and number of leaves (row). The annotations represent the raw *p*-values of the Wilcoxon signed-rank tests: a minus sign for a *p*-value above 0.05 and *, **, *** or **** for *p*-values below 0.05, 0.01, 0.001 or 0.0001, respectively.

Similar graphs were drawn for each combination of the three question/tree parameters, the three respondents' performance metrics and the four tasks in the first part of the survey. The results of their analysis are summarised in Table 4, while the details are explained in the supplementary materials in Section 6. Because there is dense information summarised in the table, an example on how to read the cell in the *classify* row and question depth column follows. Increasing the question depth (d) in the *classify* task significantly (with very low *p*-values) increases the time-to-answer in 13 out of 18 tests (each test corresponds to one star or minus sign in Fig. 8) and significantly (with very low *p*-values) increases the question-difficulty in 16 out of 18 tests. There is no influence on the answer-correctness.

Table 4 shows that increasing the question depth most strongly increases the time-to-answer and question-difficulty; it also consistently decreases the answer-correctness (except in the easiest task – *classify*), albeit not at a statistically significant level. Increasing the number of leaves increases the time-to-answer and question-difficulty as well, but only if the questions are not trivial (i.e., a low question depth and/or branching factor). Furthermore, the *p*-values are generally higher compared to the corresponding tests for the question depth. The gap between the influence of the number of leaves and the question depth is smaller in the *validate* and *discover* tasks than in the *classify* and *explain* tasks because they require analysing a larger part of the tree and not only a single path from the root to a leaf. The branching factor does not influence the comprehensibility if a question is trivial, but it does have an effect for more complex questions and trees. The *p*-values for the influence of the branching factor are lower than those for the number of leaves in the *classify* and *explain* task, but comparable in the *validate* task. The analysis for the discover task is limited due to the number of available trees and questions, but the observed trends are in line with the other tasks.

4.4. The influence of the visualisation properties

In the *compare* task, we analyse the influence of the tree visualisation properties on the comprehensibility. The respondents' ratings on the seven-point scale (Piltaver et al., 2014b) are mapped to integer values between −3 and 3. The rating 3 means that the visualisation B is much more comprehensible than the visualisation A, 2 that B is more comprehensible, 1 that B is slightly more comprehensible, and 0 that they are equally comprehensible. The negative marks represent the inverse relations: A is more comprehensible than B.

Table 4
Each cell lists ***x*/*y***: the number of significant (x) vs. the total number of performed tests (y) for three respondent performance parameters: time-to-answer (⏱), question-difficulty (💡) or answer-correctness (%). Each column shows the results for the studied question parameters: the question depth (d), the number of leaves (l) and the branching factor (b). Each ratio is accompanied with a comment on the following: the significant tests (i); or the influence of the increased parameter values (j); or the data availability (k).

Task	Question/tree parameter		
	Question depth (d)	**Number of leaves (l)**	**Branching factor (b)**
Classify	⏱ **13/18** (72%) ⏱ most with a very low p-value [i] 💡 **16/18** (89%) 💡 some with a very low p-value [i] % no influence [j]	⏱ **4/15** (27%) ⏱ low p-value for $d>1$ & $b>2$ [i] 💡 **2/15** (13%) 💡 only for $d>1$ and $b>2$ [i] % no influence [j]	⏱ **2/12** (17%) ⏱ only for $d>1$ [i] 💡 **4/12** (33%) 💡 only for $d>1$ [i] % no influence [j]
Explain	⏱ **10/10** (100%) ⏱ most with a very low p-value [i] 💡 **8/10** (80%) 💡 most with a very low p-value [i] % decreases (8/10) but test are not significant [j]	⏱ **1/5** (20%) ⏱ only tested for $b=2$ or $d=1$ [k] 💡 **2/5** (40%) 💡 only tested for $b=2$ or $d=1$ [k] % no influence [j]	⏱ **2/4** (50%) ⏱ tested for $l=10$ or $l=4$ & $d=1$ [k] 💡 **0/4** (0%) 💡 tested for $l=10$ or $l=4$ & $d=1$ [k] % no influence [j]
Validate	⏱ **6/11** (54%) ⏱ most with a very low p-value [i] 💡 **6/11** (54%) 💡 all with moderate p-values [i] % decreases (9/11), but the tests are not significant [j]	⏱ **3/6** (50%) ⏱ significant for $b>2$ [i] 💡 **1/6** (17%) 💡 with moderate p-value [i] % no influence [j]	⏱ **3/6** (50%) ⏱ for increase of b from 2 to 3 [i] 💡 **1/6** (17%) 💡 tested only for $l=10$ [k] % no influence [j]
Discover	⏱💡 % trends observed but not tested for significance [j]		⏱💡 % no data [k]

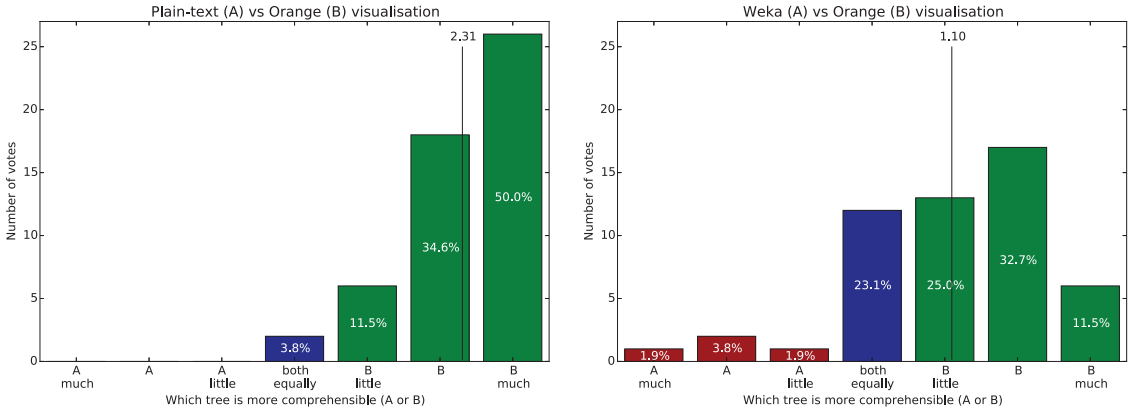


Fig. 9. The influence of the tree visualisation style on the tree comprehensibility.

The respondents' opinions about the influence of the tree visualisation style on the comprehensibility are shown with histograms in Fig. 9. The left histogram compares the plain-text visualisation (Fig. 4 left), and the right histogram compares the Weka visualisation (Fig. 4 right) with the Orange visualisation of the same trees. The vertical lines show the average ratings. The respondents find the plain-text visualisation to be considerably less comprehensible than the Orange visualisation: the average rating is 2.31, which is the highest among all of the questions in the *compare* task. Therefore, the plan-text visualisation should be avoided when analysing the classification trees. The majority of respondents prefer the Orange over the Weka visualisation (Fig. 9 on the right), but the average rating of 1.1 is lower than in the comparison with the plain-text visualisation. This finding shows that the additional colour-coded information is beneficial. On the other hand, 3 *students* and 1 *IT specialist* prefer Weka visualisation; they may find colour-coded information confusing and therefore prefer the simpler visualisation.

Second, the influence of the pie charts (which represent class distributions, see Fig. 5a and c) in the Orange visualisation is compared using the histogram shown on the left of Fig. 10. The result is similar to the result above, but the average rating is lower, which suggests that pie charts are the main but not the only difference between the Weka and Orange visualisations. The remainder can be attributed to minor differences such as colour-coding the number of examples in the leaves, and the respondents' bias toward the Orange visualisation introduced by the survey design.

Third, the importance of the attribute names, attribute values and class names is analysed by comparing simple trees with meaningful (Fig. 5a) and random (Fig. 5b) values and names. The results are presented in Fig. 10: the average rating (1.54) indicates that meaningful names are more important for comprehensibility than the class-distribution pie charts. This result empirically confirms the fact (used in a related survey design (Allahyari and Lavesson, 2011)) that meaningful concept names make information easier to remember and support faster processing and the triggering of associations with the user's domain knowledge, which facilitates

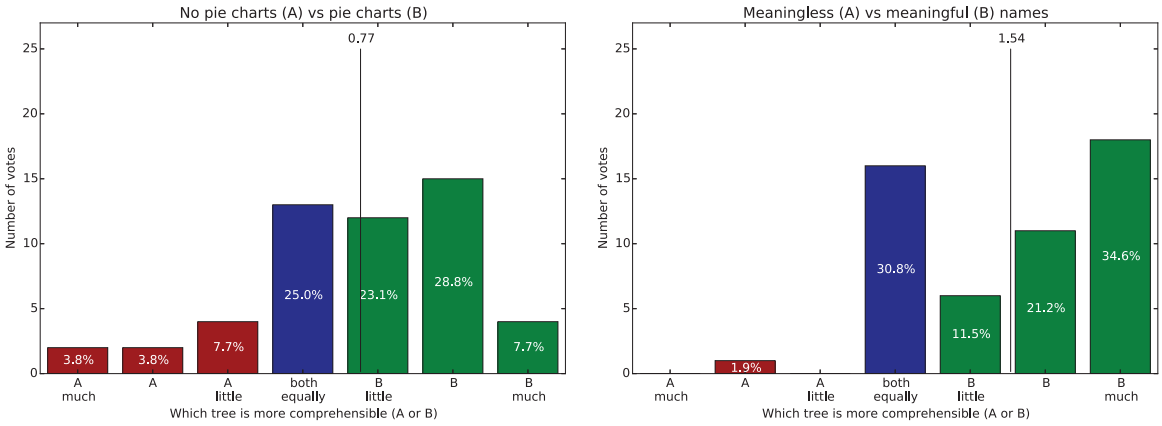


Fig. 10. The influence of visualised information on the tree comprehensibility.

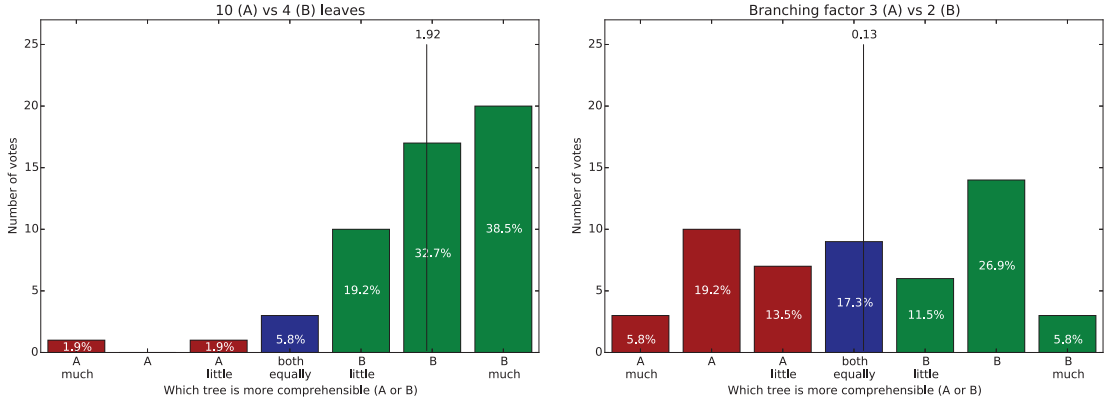


Fig. 11. The influence of the tree structure on the tree comprehensibility.

comprehension. Therefore, we suggest using meaningful names in tree visualisations whenever possible, and we warn against using misleading or ambiguous names.

Fourth, we analyse the influence of the tree layout on its comprehensibility. A tree shown in Fig. 1 is compared with the same tree when drawn using the same visualisation style but a random layout – with random relative horizontal positions for the sibling nodes. The layouts of both trees are illustrated in Fig. 6. Although almost half of the respondents rated the trees as equally comprehensible, 42% preferred the tree with the left-to-right layout (suggested in Section 3.4), and less than 10% preferred the random layout. The average rating (0.56) is the lowest among all of the results presented in this section; nonetheless, we believe that the importance of the tree layout should be investigated further, especially in larger trees and for users whose native reading direction is not from left to right.

Finally, we analyse the influence of the tree structure parameters on the tree comprehensibility. Over 90% of the respondents found a tree with 4 leaves to be more comprehensible that a tree with 10 leaves (Fig. 11 on the left), which is in line with the results from the previous tasks and related work (Allahyari and Lavesson, 2011; Huysmans et al., 2011). The average rating was 1.92, which is more than the influence of meaningful names but less than the difference between plain-text and Orange visualisation.

Comparing a tree with 9 leaves and a branching factor of 3 with its a binary version (Table 2, bottom row left and middle) exhibits an inconclusive but interesting result shown in Fig. 11 on the right. We cannot say which tree is more comprehensible: while 42% of the respondents think that there is no difference or only a slight difference, the majority think that one or the other tree is more or much more comprehensible. However, the preferences of these respondents are opposing: one half prefers the deeper tree with binary splits, while the other half prefers the shallower tree with a branching factor of 3. This finding could explain why a limited influence of the branching factor is observed in the previous four tasks. The influence of the branching factor on the comprehensibility should be investigated further in domains in which there are multiple values of nominal attributes, which naturally yield trees that have higher branching factors.

Finally, the influence of the tree depth is analysed by comparing trees with 9 leaves and a branching factor of 3: one with a depth of 2 and the other with a depth of 4 (Table 2, bottom middle cell). The average rating is 0.15. No difference in the comprehensibility is noticed by 35% of the respondents, and only 29% noticed more than a slight difference: approximately half of them prefer the shallower tree, and the other half prefer the deeper tree. We conclude that the tree depth itself does not have an important influence on the comprehensibility, which is the reason why we decided to omit an in-depth analysis of the tree depth for the first four survey tasks in this paper. The other reason is that the number of tree pairs with different tree depths but no difference in the other tree structure parameters is very limited. Hence, similar trees with different depths would have to be compared, but the small influence of the tree depth on the comprehensibility would then drown in the influence of the other parameters.

4.5. The comprehensibility metrics

The comprehensibility ratings of the 11 trees (illustrated in Table 2) obtained in the *rate* task are used to analyse the match between various tree comprehensibility metrics and the respondents' subjective opinions and to suggest new comprehensibility metrics.

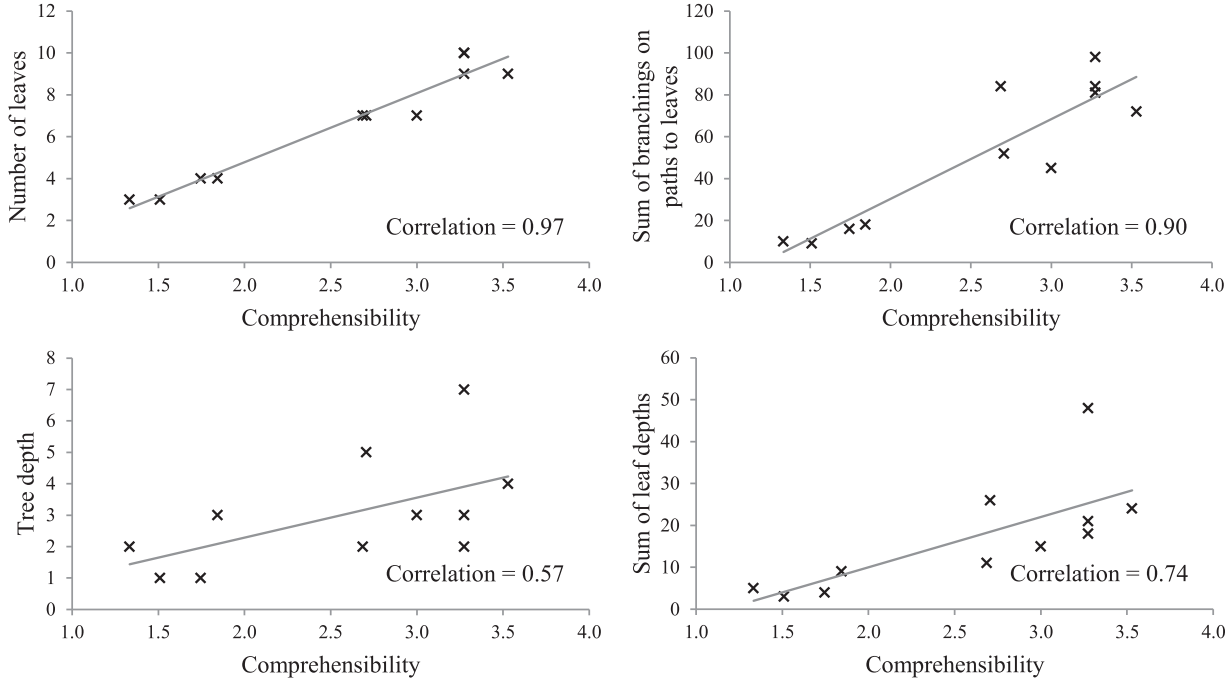


Fig. 12. Correlations between comprehensibility metrics and the average respondent ratings (1: very easy to comprehend, 5: very difficult to comprehend).

The respondents' opinions about the tree comprehensibility differ but are generally within ± 1 of the average answer (see supplementary material, Section 7). This finding suggests that the used rating scale is valid. Furthermore, the opinions show that trees with more leaves are more difficult to comprehend. More leaves in turn causes a higher tree depth, and therefore, increasing the tree depth appears to decrease the tree comprehensibility. However, the tree depth and branching factor do not have an obvious (or consistent) effect on the comprehensibility if the number of leaves is fixed. These results agree with the results obtained in the first part of the survey.

Note that the complexity/comprehensibility of the analysed trees is limited: the average ratings of all of the trees are below *difficult to comprehend*. The results are therefore valid for trees that have modest complexity and should not be overgeneralised to more complex trees.

Finally, we evaluate several tree complexity metrics that are being used as surrogates for the comprehensibility metrics. They are evaluated based on the Pearson correlation coefficients with the average comprehensibility ratings (Fig. 12). The correlations are as follows: 0.97 for the number of leaves; 0.88 for the number of nodes; 0.57 for the tree depth; and 0.17 for the branching factor. This finding confirms that the number of leaves is a valid metric for the tree comprehensibility in trees that have modest complexity.

Finally, we defined two new comprehensibility metrics that can be weighted by the parameter w_l, which is defined in each leaf l. The weighting is important because it enables specifying user preferences that are related to the semantics of a classification tree, rather than relying solely on the tree structure to estimate its comprehensibility. For example, weighting by the number of instances in a leaf (mentioned in Sommer (1995)) corresponds to preferring the trees that classify the most instances with shallow leaves (i.e., simple classification rules). Another example is weighting by the importance of a class, which corresponds to preferring the trees that classify instances that belong to the important class(es) with short rules. Both comprehensibility metrics account for the question depth, which is defined per leaf and has the highest influence on the question-difficulty and time-to-answer among all of the analysed tree structure parameters. The first metric c_1 (Eq. 1) is the weighted sum of the depths d_l over all n leaves in a tree and has a correlation coefficient of 0.74 (p-value below 0.01). The second metric c_2 (Eq. 2) is the weighted sum of the branching factors b_i on the paths from the root of the tree to a leaf l over all n leaves and has a correlation coefficient of 0.90 (p-value below 0.001). The Pearson correlation coefficients were obtained using the default weights $w_l = 1$.

$$c_1 = \sum_{l=1}^{n} w_l d_l \tag{1}$$

$$c_2 = \sum_{l=1}^{n} w_l \sum_{i \in path(root,l)} b_i \tag{2}$$

The correlation coefficient of the metric c_2 is comparable to the best correlation coefficient among the previously mentioned simple comprehensibility metrics. Furthermore, its correlation increases to 0.94 if the weighting is set by the number of instances in a leaf n_l divided by the number of all instances n ($w_l = n_l/n$).

5. Discussion and conclusions

This paper analyses how tree structure parameters and visualisation style influence the comprehensibility of classification trees. The results are based on empirical data obtained using a carefully designed survey with 98 questions answered by 69 respondents. The survey design allows detecting relatively small differences in the comprehensibility even with a limited number of respondents. Furthermore, the survey implementation enables analysing the effect of a single tree-structure parameter independently of the other parameters. The 11 trees used in the survey have 3 to 11 leaves, 4 to 20 nodes, branching factors between 2 and 4, and tree depths between 2 and 7. Therefore, the conclusions are valid for trees that have modest complexity (commonly used in practice) and should be generalised to large trees with caution. The respondents' performance is measured as the time-to-answer and answer-correctness

in four tasks: *classify* an instance, *explain* the classification of an instance, *validate* that the tree agrees with a domain-knowledge rule, and *discover* new knowledge about unusual instances. The survey measures the tree usability directly and objectively, while the tree comprehensibility is measured objectively but indirectly (through its usability). In addition, the respondents' subjective opinions about comprehensibility are collected by asking how difficult each question is in the first four tasks (question-difficulty), how comprehensible the tree is (the *rate* task) and by comparing the comprehensibility of pairs of trees (the *compare* task) that differ in their visualisation properties.

The main contributions of this paper are the quantification and analysis of the influence of (i) the tree structure parameters, (ii) the question depth, (iii) the tree visualisation properties, (iv) data-mining experience on the tree comprehensibility, and (v) definition of two new classification-tree comprehensibility metrics that consider the structure and semantics of the tree.

The time-to-answer is highly correlated to the question-difficulty, and therefore, the two metrics are well suited for measuring the classification tree comprehensibility in the absence of direct and objective measures. This finding also confirms that the question-difficulty scale is well-designed. The variability of subjective opinions is considerable, which emphasises the subjective nature of comprehensibility. Data mining experts perform better than students, which shows that more experienced classification tree users comprehend the trees faster and better. The *classify* task is the easiest, and the *discover* task is the hardest (of the four tasks), according to the question-difficulty, answer-correctness and time-to-answer.

Tree-structure properties clearly influence the time-to-answer and question-difficulty, but they have a limited effect on the answers-correctness. Increasing the **number of leaves** (which in turn increases the number of nodes and the tree depth) increases the time-to-answer and question-difficulty. The result is statistically significant in the *classify* and *explain* tasks for non-trivial questions. The influence on the time-to-answer is also significant in the *validate* task. The same trend is observed in all of the other cases, but the results are not statistically significant. Increasing the number of leaves clearly makes a tree less comprehensible, as confirmed by the *rate* and *compare* tasks. The result empirically supports that the number of leaves is a usable tree comprehensibility metric, but only if leaves with sufficiently high depths and trees with branching factors of greater than 2 are considered. Although the number of leaves and the tree depth corresponds well to the comprehensibility of the entire tree, they are not relevant if the needed information is close to the root of the tree. Finally, the results show that the number of leaves has a greater influence on the tree comprehensibility in the *discover* task compared to the other three tasks because discovering new knowledge from the tree requires at least scanning through if not reading and understanding the entire tree.

Increasing the **branching factor** increases the time-to-answer for non-trivial questions in the *classify, explain* and *validate* tasks (not analysed in the *discover* task). Increasing the branching factor also increases the question-difficulty; however, the result is statistically significant only in the *classify* task. Higher branching factors should be investigated further because there are two types of respondents that have opposite preferences: half of them prefer a binary tree, while the other half prefer a tree with a branching factor of 3 (according to the *compare* task). This finding could explain why a limited influence of the branching factor is observed in the first part of the survey. In addition, further work should focus on trees that have non-uniform branching factors.

The **tree depth** itself does not have an important influence on the tree comprehensibility and strongly depends on other tree structure parameters (the number of leaves and branching factor); therefore, its influence on the comprehensibility cannot be systematically analysed.

Increasing the **question depth**, which corresponds to the depth of the deepest leaf that is required to answer a question, increases the time-to-answer and question-difficulty. The influence is statistically significant in almost all of the comparisons. Although this finding is intuitively expected, the contribution of our study is that it empirically proves the result even for a simple task such as classification. In addition, increasing the question depth decreases the answer-correctness in all of the tasks except for the easiest task (*classify*), although this result is not statistically significant. Increasing the number of respondents would clarify the observed trend. The question depth has the highest influence on the comprehensibility among all of the observed parameters and is overlooked in the related work.

The results on the question depth suggest that users prefer trees that follow the following principle: the most important information and information about the most common instances should be given at the top of the tree, i.e., with short rules, while the remaining instances can be classified with deeper leaves. The relative importance and distribution of the instances according to the depth of the leaves to which they correspond should be accounted for by the tree comprehensibility metrics and tree learning heuristics.

Verifying **negated statements** increases the time-to-answer and question-difficulty and reduces the answer-correctness compared to verifying positive statements. Therefore, negated statements should be avoided in classifier representations whenever possible. The effect of negating a statement is greater than the effects of increasing the question depth, number of leaves or branching factor.

The tree **visualisation style** has an important influence on the tree comprehensibility and usability. Plain-text visualisation should be avoided while colour-coded information is beneficial. Meaningful attribute names, attribute values and class names should be used in tree visualisation whenever possible, while misleading and ambiguous names should be avoided.

The results also indicate that the **tree layout** is important for many users, and thus, its influence on the comprehensibility should be verified with objective measures. The tree layout algorithm introduced in this paper can be used as a starting point for further research and to improve the visualisation of classification trees.

Finally, this paper defines two new classification-tree **comprehensibility metrics** that are in line with the survey results: the weighted sum of the depths of the leaves and the weighted sum of the branching factors on the paths from the root to the leaves. Both metrics can be weighted by a parameter that is defined in each leaf, which enables specifying user preferences that are related to the semantics of a classification tree rather than relying solely on the tree structure to estimate its comprehensibility.

In **future work**, we plan to study the influence of the number of leaves in larger (binary) trees and the branching factor in trees with higher (and non-uniform) branching factors, which results from nominal attributes with more than two possible values. Studying trees with higher complexity as well as increasing the number of respondents will clarify the results about the influence on the answer-correctness. Another promising direction for future study is analysing the influence of the visualisation properties, especially the tree layout, using the objective measures (time-to-answer, answer-correctness). Including respondents form various cultures (e.g., who differ in the reading direction of their languages) will confirm whether the reported results are culture specific. Finally, the results of this study should be confirmed in real-life domains that involve strong numeric components, such as in medical applications (e.g., stress level and activity monitoring), and

insurance or bank loans, for example. A deeper investigation of the influence of the question depth could ultimately lead to learning algorithms that will produce more comprehensible trees.

Acknowledgements

The authors would like to thank the Slovene Human Resources and Scholarship Fund for co-funding the international research cooperation that made this research possible. The discussions with psychologists Jana Krivec and Amir Tokić as well as with data-mining experts Ljupčo Todorovski and Ana Meštrović helped us to improve the survey design. We would also like to thank our colleagues and students at the University of Rijeka, Department of Informatics, and the Jožef Stefan Institute for answering the lengthy survey and also thank the anonymous reviewers for suggesting improvements in the paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2016.06.009.

References

Allahyari, H., & Lavesson, N. (2011). User-oriented assessment of classification model understandability. *11th scandinavian conference on Artificial intelligence*.

Aranda, J., Ernst, N., Horkoff, J., & Easterbrook, S. (2007). A framework for empirical evaluation of model comprehensibility. In *Proceedings of the international workshop on modeling in software engineering, MISE '07* (p. 7). IEEE Computer Society.

Bache, K. and Lichman, M. (2013). UCI machine learning repository.

Benbasat, I., & Taylor, R. N. (1982). Behavioral aspects of information processing for the design of management information systems. *Systems, Man and Cybernetics, IEEE Transactions on, 12*(4), 439–450.

Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for experimenters: design, innovation, and discovery* (2nd ed.). John Wiley & Sons, Inc..

Campbell, D. J. (1988). Task complexity: A review and analysis. *The Academy of Management Review, 13*(1), 40–52.

Chorowski, J. (2012). *Learning understandable classifier models* PhD thesis. Louisville, Kentucky: Department of Electrical and Computer Engineering – University of Louisville.

Craven, M. W., & Shavlik, J. W. (1995). Extracting comprehensible concept representations from trained neural networks. In *Working notes on the IJCAI'95 workshop on comprehensibility in machine learning* (pp. 61–75).

Demšar, J., Curk, T., Erjavec, A., Crt Gorup, Hocevar, T., Milutinovic, M., et al. (2013). Orange: Data mining toolbox in python. *Journal of Machine Learning Research, 14*, 2349–2353.

Elomaa, T. (1994). In defense of C4. 5: Notes learning one-level decision trees. In W. W. Cohen, & H. Hirsh (Eds.), *Proceedings of the eleventh international conference on machine learning* (pp. 62–69). Morgan Kaufmann.

Freitas, A. A. (2003). A survey of evolutionary algorithms for data mining and knowledge discovery. In *Advances in evolutionary computing* (pp. 819–845). Springer Berlin Heidelberg.

Freitas, A. A. (2004). A critical review of multi-objective optimization in data mining: a position paper. *ACM SIGKDD Explorations Newsletter, 6*(2), 77–86.

Freitas, A. A. (2014). Comprehensible classification models: A position paper. *SIGKDD SIGKDD Explorations Newsletter, 15*(1), 1–10.

Giraud-Carrier, C. (1998). Beyond predictive accuracy: What? In *ECML'98 workshop notes - upgrading learning to the meta-level: Model selection and data transformation* (pp. 78–85). Technical University of Chemnitz. Conference Proceedings/Title of Journal: ECML'98 Workshop Notes - Upgrading Learning to the Meta-Level: Model Selection and Data Transformation.

Göpferich, S. (2009). Comprehensibility assessment using the karlsruhe comprehensibility concept. *The Journal of Specialised Translation, 11*, 31–48.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10–18.

Harris, E. (2001). Information gain versus gain ratio: A study of split method biases. Technical report, The MITRE Corporation.

Houy, C., Fettke, P., & Loos, P. (2012). Understanding understandability of conceptual models- what are we actually talking about?. In P. Atzeni, D. Cheung, & S. Ram (Eds.), *Conceptual modeling* In *Lecture notes in computer science: vol. 7532* (pp. 64–77). Berlin Heidelberg: Springer.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems, 51*(1), 141–154.

Jin, Y., & Sendhoff, B. (2008). Pareto-based multiobjective machine learning: An overview and case studies. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 38*(3), 397–415.

Jin, Y., Sendhoff, B., & Körner, E. (2005). Evolutionary multi-objective optimization for simultaneous generation of signal-type and symbol-type representations. In C. Coello Coello, A. HernÃ˜ndez Aguirre, & E. Zitzler (Eds.), *Evolutionary multi-criterion optimization*. In *Lecture notes in computer science: vol. 3410* (pp. 752–766). Springer Berlin Heidelberg.

Johansson, U., Niklasson, L., & König, R. (2004). Accuracy vs. comprehensibility in data mining models. In *Proceedings of the seventh international conference on information fusion: 1* (pp. 295–300).

Kodratoff, Y. (1994). The comprehensibility manifesto. *KDD Nuggets, 94*(6).

Lee, C.-C., Cheng, H. K., & Cheng, H.-H. (2007). An empirical study of mobile commerce in insurance industry: Task-technology fit and individual differences. *Decision Support Systems, 43*(1), 95–110 Mobile Commerce: Strategies, Technologies, and ApplicationsDSS on M-Commerce.

Liu, G. P., & Kadirkamanathan, V. (1995). Learning with multi-objective criteria. In *Artificial neural networks, 1995., fourth international conference on* (pp. 53–58).

Maimon, O., & Rokach, L. (2005a). *Data mining and knowledge discovery handbook*. Secaucus, NJ, USA: Springer-Verlag New York, Inc..

Maimon, O., & Rokach, L. (2005b). Decomposition methodology for knowledge discovery and data mining. In O. Maimon, & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 981–1003). US: Springer.

Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. *Decision Support Systems, 51*(4), 782–793 Recent Advances in Data, Text, and Media Mining & Information Issues in Supply Chain and in Service System Design.

Michalski, R. S. (1983). A theory and methodology of inductive learning. In R. Michalski, J. Carbonell, & T. Mitchell (Eds.), *Machine learning*. In *Symbolic computation* (pp. 83–134). Springer Berlin Heidelberg.

Michie, D. (1987). *Second European working session on learning*. Slovenia: Bled.

Murphy, G. L., & Wright, J. C. (1984). Changes in conceptual structure with expertise: Differences between real-world experts and novices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 144–155.

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 35*(4), 737–743.

Pazzani, M. J. (2000). Knowledge discovery from data? *IEEE Intelligent Systems, 15*(2), 10–13.

Piltaver, R., Luštrek, M., & Gams, M. (2014). Multi-objective learning of accurate and comprehensible classifiers – a case study. In U. Endriss, & J. Leite (Eds.), *Proceedings of 7th European starting AI researcher symposium* (pp. 220–229). IOS Press.

Piltaver, R., Luštrek, M., Gams, M., & Martinčić – Ipšić, S. (2014a). Comprehensibility of classification trees – survey design. In R. Piltaver, & G. Matjaž (Eds.), *Proceedings of 17th international multiconference information society* (pp. 6–10). Slovenia: Ljubljana.

Piltaver, R., Luštrek, M., Gams, M., & Martinčić – Ipšić, S. (2014b). Comprehensibility of classification trees – survey design validation. In *Proceedings of 6th international conference on information technologies and information society* (pp. 5–7). Slovenia: Šmarješke toplice.

Piltaver, R., Luštrek, M., Zupančič, J., Džeroski, S., & Gams, M. (2014). Multi-objective learning of hybrid classifiers. In T. Schaub, G. Friedrich, & B. O'Sullivan (Eds.). In *Proceedings of European conference on artificial intelligence – ECAI 2014: 263* (pp. 717–722). IOS Press. doi:10.3233/978-1-61499-419-0-717.

Quinlan, J. (1999). Some elements of machine learning. In S. Džeroski, & P. Flach (Eds.), *Inductive logic Programming*. In *Lecture notes in computer science: vol. 1634* (pp. 15–18). Springer Berlin Heidelberg.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..

Schriver, K. A. (1989). Evaluating text quality: The continuum from text-focused to reader-focused methods. *Professional Communication, IEEE Transactions on, 32*(4), 238–255.

Sommer, E. (1995). An approach to quantifying the quality of induced theories. In C. Nedellec (Ed.), *Proceedings of the IJCAI workshop on machine learning and comprehensibility*.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.

Vessey, I., & Galletta, D. (1991). Cognitive fit: An empirical study of information acquisition. *Information Systems Research, 2*(1), 63–84.

Zhou, Z.-H. (2005). Encyclopedia of data warehousing and Mining, *chapter Comprehensibility of data mining algorithms* (pp. 190–195). Hershey: Idea Group Reference.

Zichermann, G., & Cunningham, C. (2011). *Gamification by design: Implementing game mechanics in web and mobile apps* (1st ed.). O'Reilly Media, Inc.