Feature Evaluation Through Decision Trees Structure

Beata Zielosko
University of Silesia in Katowice
Faculty of Science and Technology/Institute of Computer Science
Sosnowiec, Poland *beata.zielosko@us.edu.pl*

Urszula Stańczyk
Silesian University of Technology
Department of Computer Graphics, Vision and Digital Systems
Gliwice, Poland *urszula.stanczyk@polsl.pl*

Kamil Jabłoński
University of Silesia in Katowice
Faculty of Science and Technology/Institute of Computer Science
Sosnowiec, Poland *kamil.jablonski@us.edu.pl*

Mikhail Moshkov
King Abdullah University of Science and Technology (KAUST)
Computer Electrical and Mathematical Sciences & Engineering Division
Thuwal, Saudi Arabia *mikhail.moshkov@kaust.edu.sa*

Abstract

Feature selection plays a significant role in the development of categories of information systems related to decision support, such as diagnostic or recommendation systems. Such systems should ensure the possibility of identifying the most important features as well as analysing data from different locations, taking into account the specificity and characteristics of the local data sources. In the process of data analysis, the stage of data preparation, including the transformation of the attribute domain from continuous form to intervals, plays an important role, as the outcome of this process influences the subsequent stages of the analysis. In the paper, an approach to creating a global feature ranking that takes into account the specifics and characteristics of different discretisation algorithms was proposed. A new weight for the estimation of attribute importance was defined and compared with a measure that is implemented in the Python programming language library. Both types of weights were used to create a hierarchical structure of the global ranking of features. The experiments were carried out on datasets from the stylometry domain dedicated to the task of authorship attribution.

Keywords: Ranking, discretisation, decision tree, authorship attribution.

1. Introduction

Technological progress and constantly increasing amounts of data require the development of advanced tools dedicated to data analysis with a view to interpretation. Feature selection and the construction of attribute rankings are important methods applied in decision-making systems used in many domains [12]. The ranking process aims to identify the most relevant attributes from the input feature set while preserving the descriptive and representative properties of the original feature space [4]. It plays an important role because as the number of available attributes increases, the computational complexity and the risk of overfitting the machine learning models also increase, and the interpretation of the results becomes more difficult. Thus, methods to identify the most relevant features that describe the issue under analysis are of great importance in the data preparation stage.

It can be carried out by identifying the minimum set of features that contribute to a certain satisfactory level of classifier performance, or by ranking the entire set of attributes. In the latter case, features are assigned weights according to a defined criterion and ranked taking into account the values of this weight [1]. Reducing the number of variables used makes it easier to understand the decision-making mechanisms, which is particularly important in the context of interpreting the decisions proposed by the system.

From the point of view of the capabilities of the algorithms used to create machine learning models, the type of data available should be taken into account. Some methods, for example, the approach based on the classical rough set theory, require a discrete (categorical) form of the data to create rule-based models [10]. Therefore, in many cases, it becomes necessary to carry out a discretisation process, i.e., the transformation of continuous variables into categorical ones. There are different approaches to discretisation and different algorithms, and the choice of a particular discretisation method directly affects the subsequent quality of the model and the structure of the dataset. There is no one-size-fits-all approach in this regard.

In the paper, the authors propose a method for ranking attributes that takes into account the specificities of different discretisation algorithms. The aim is to design an approach to attribute evaluation that not only measures the relevance of attributes in the context of the model, but also takes into account how they are discretised, which can significantly affect their informativeness and usefulness in the subsequent modelling process. To evaluate the importance of a feature, a new weight based on decision trees is also proposed.

Decision trees belong to the intuitive and widely used machine learning methods for regression and classification tasks. Their great advantage is the transparency and interpretability resulting from the tree structure, which clearly reflects the path leading to the decision proposed by the model [9]. The hierarchical structure of the tree naturally allows for the evaluation of the importance of attributes. The proposed method takes into account the level at which the node labelled by the attribute occurs on the path leading from the root to the leaf in the decision tree.

The hierarchical attribute ranking approach based on decision trees was investigated and evaluated for the two most common approaches to discretisation: supervised and unsupervised. For these categories, four algorithms were studied: Fayyad and Irani [3] and Kononenko [7], and equal frequency and equal width binning methods [2]. For unsupervised methods, nine variants of each algorithm were analysed. At the lowest level of hierarchy, a total of 50 rankings were created based on the proposed weight, which were then combined to create two main global rankings.

At each level of the hierarchy, starting at the local level, then at the level of variants of unsupervised methods, further at the level of supervised and unsupervised discretisation approaches, and at the global level, the rankings were verified based on the XGBoost model. The attribute sets were reduced backward while following the constructed rankings, starting with the one in the lowest position. For each subset of features, the data were explored and the performance of the classifiers evaluated. The proposed method for estimating the importance of attributes by calculated weights and generated rankings was compared with the importance evaluation method implemented in the DecisionTreeClassifier in the Python library, which is based on measuring the decrease in impurity [14].

The experiments were conducted using two datasets from the field of stylometry. The classifiers were applied to the task of binary authorship attribution. The authorship of the text samples was determined based on writing styles, characterised by a set of quantitative linguistic features. The results obtained confirmed the merit and validity of the proposed methodology and the proposed weight of attributes because feature reduction exploiting generated rankings led to improved performance.

The paper consists of five sections. The introduction is followed by Section 2, which presents background information. The proposed methodology and approach for assigning local and global weights are presented in Section 3. Section 4 describes the obtained rankings and their verification using the performance of constructed classifiers. Section 5 concludes the paper.

2. Background

This section presents the background and tools involved in the research. In particular, methods used in the ranking construction and data modelling process are described.

2.1. Feature Selection and Ranking Construction

In the process of building machine learning models and analysing data, an important step is the selection of attributes and their ranking [11]. The removal of redundant and irrelevant attributes can be performed by finding the minimum set of features that meets the selected criterion or by ranking attributes. The latter involves selecting the most relevant variables that have the greatest impact on the model output. Based on the defined weights assigned to the attributes, the features are ordered according to their values, and the k most relevant attributes can be selected. The criteria for determining the importance of features can rely on various measures [2]. These are often statistic-oriented, for example, entropy, information gain, or measures built into the algorithm, such as Relief or OneR.

In the paper, the authors propose a ranking method that relies on the depth of the decision tree and the node at which features appear. For comparison purposes, a ranking based on the feature importance scores computed by the DecisionTreeClassifier from the scikit-learn library is also considered. This feature weight is assessed by measuring the total decrease in impurity, in this case, the Gini index, contributed by each attribute across all splits of the tree.

2.2. Decision Trees

Decision trees belong to popular forms of knowledge representation and are often used in classification and regression tasks. A decision tree consists of internal nodes where features are tested, branches that represent the outcomes of these tests, and leaf nodes that assign a class label or a numerical value. The beauty of decision trees is that they are interpretable, as each path leading from the root to the leaf represents a decision-making process.

There are many algorithms for decision tree induction [6]. Popular ones include CART (Classification and Regression Trees), ID3 (Iterative Dichotomiser 3), or C4.5. Each of these algorithms builds a tree structure by recursively splitting the data based on feature values and an adapted criterion, such as information gain (ID3, C4.5) or the Gini index (CART). Among the ensemble classifiers based on decision trees, Random Forest, XGBoost, and AdaBoost should be mentioned. They build models based on multiple decision trees, differing in learning strategy, speed of operation, and resistance to overfitting. XGBoost builds trees sequentially, where each new tree corrects the errors of previous ones.

In this work, decision trees are used to build feature ranking. Attributes existing in nodes that appear closer to the root are considered more influential, providing an intuitive measure of feature importance derived directly from the model structure.

The weights of attributes, the ones proposed and these implemented in DecisionTreeClassifier, are based on the Gini index. It is an impurity measure applied during the decision tree construction process [8]. It selects those divisions in the decision tree that most reduce the impurity in the node. Therefore, the smaller the value, the better, because the data are more homogeneous in relation to the decision class. For a set with class labels $d_1, \dots d_k$ and probabilities $p(d_i)$, for each class value, the Gini impurity is given by: $Gini = 1 - \sum_{i=1}^{k} p(d_i)^2$, where $p(d_i)$ is the proportion of samples that belong to the class d_i within the subset.

2.3. Discretisation

Discretisation is part of data reduction and plays an important role in the data preparation stage. It transforms numerical attributes into discrete or nominal ones with a finite number of intervals (bins). There are many discretisation methods and approaches that can be divided based on various criteria. The two popular approaches are supervised and unsupervised approaches. The latter do not consider class information during the process of transformation of attributes' values, and the number of bins is provided as an input parameter. In the case of supervised methods, class information is taken into account to find proper intervals among ranges of attribute values. Often, some heuristic measures, e.g., entropy, are used to determine the best cut-points.

The most popular unsupervised discretisation algorithms are equal width and equal frequency binning [2]. The equal width algorithm sorts the values of a continuous attribute, designates the minimum and maximum values of the discretised attribute, and then divides the range into k equal width discrete intervals, where k is a parameter defined by a user. In the case of the equal frequency algorithm, each bin contains the same number of attribute's values. The two methods are simple and sensitive to a number of bins defined by a user. The disadvantage is that in cases where the values of the continuous attribute are not distributed evenly, some information can be lost after the discretisation process.

Fayyad and Irani [3], and Kononenko [7] belong to supervised discretisation approaches. They are based on the class entropy of the considered intervals for evaluating cut-points and the Minimum Description Length (MDL) principle as a stopping criterion. The processing starts from one interval containing all values of the discretised attribute, which is then partitioned recursively, until a stopping criterion is met.

2.4. Datasets from Stylometry Domain

The attribute ranking approach proposed in the paper was applied and tested on datasets from the stylometry domain. This is a field of research dedicated to the quantitative analysis of the style of written texts. The research aims to identify the authorship of a text based on measurable language properties such as the frequency of function words or punctuation marks [5].

To better observe style changes, the selected novels were divided into smaller parts. In each excerpt, the frequency of 24 style features was measured: 22 common function words and 2 punctuation marks. After data preparation, sets with real-valued attributes were obtained. Each set consisted of the training set and two test sets, which were used to check the effectiveness of the models. All sets were prepared for the binary classification task with balanced decision classes. The recognised classes corresponded to pairs of writers, the female writer dataset (F-writers) reflected works of Edith Wharton and Mary Johnston, and the male writer (M-writers) dataset compared works of Jack London and James Oliver Curwood.

The datasets were discretised using algorithms implemented in the WEKA environment [13]. In the supervised approach, two variants of data were created using the Fayyad and Irani algorithm (denoted dsF) and the Kononenko method (denoted dsK). In the unsupervised approach, 18 variants of the datasets were created using equal width and equal frequency binning algorithms, with the number of bins ranging from 2 to 10 (denoted duf2÷duf10 and duw2÷duw10, respectively). All sets were discretised independently.

3. Methodology for Construction of Attribute Ranking

The proposed methodology consists of the following steps:

- preparation of input datasets;
- discretisation;
- induction of the decision tree using the Gini index, for each discretised variant of datasets;

- ranking construction using two different methods based on decision trees;
- hierarchical approach for global ranking construction based on discretisation categories and algorithms;
- rankings verification at each level of hierarchy using XGBoost model and strategy for backward elimination of attributes.

The proposed ranking method refers to the depth of the decision tree, which is the maximum length of the path from the root to the leaf. For each attribute labelling a node in a decision tree, its depth is determined as the path length from the root to the node it labels. This value is denoted as $h(node)$, the smaller the value, the greater the importance of the attribute. The weight of each attribute $w(a)$ is determined by:

$$w(a) = \frac{1}{2^{h(node)}}. \tag{1}$$

In the case of binary decision trees, the number of nodes at each level increases exponentially with depth, following the growth pattern 2^h. Therefore, the denominator $2^{h(node)}$ naturally reflects the structural properties of the tree: attributes appearing closer to the root are given exponentially greater importance compared to those appearing deeper. This approach emphasises that features used at early splits, when the number of instances is still large, typically provide more significant information about the data.

In order to create a universal global ranking, the proposed approach takes into account the characteristics of the various discretisation algorithms and their categories. It is based on a hierarchical structure by sequential summation of attribute weights at lower levels of the hierarchy, incorporating the specificity of the rankings at lower hierarchies. In general, the weight of an attribute at the global level can be expressed as the sum of its weights across all levels of the hierarchy, appropriately adjusted to reflect the importance of each local ranking.

Formally, with $w_i(a)$ denoting the weight of the attribute a at the hierarchy level i, the global weight $W(a)$ can be defined as:

$$W(a) = \sum_i \alpha_i \cdot w_i(a), \tag{2}$$

where α_i is a coefficient reflecting the importance of the ranking at level i. In this work, all local rankings are treated as equally important. Therefore, the coefficient α_i is set to 1 for each local ranking. As a result, the global weight of an attribute is obtained as a simple sum of its weights across all hierarchy levels, without any additional scaling or prioritisation:

$$W(a) = \sum_i w_i(a). \tag{3}$$

4. Experimental Results

This section describes the results obtained. These include two types of rankings, the one proposed and the one implemented in DecisionTreeClassifier, at the considered level of hierarchy, and the performance of classifiers recorded for different variants of discretised datasets.

4.1. Rankings

The two types of rankings presented in this section were first obtained at the lower level of the hierarchy, and then they were calculated at higher levels. Table 1 shows the attribute rankings at the global level, denoted with G1 and G2, the supervised level denoted with S1 and S2, and the unsupervised level denoted with U1 and U2. Values in the row with the name of ranking and index 1 are related to the proposed method for feature weight calculation, values in the row with the ranking name and index 2 are related to the method included in the Python library.

Table 1. Rankings of attributes obtained at the global (G1 and G2), supervised (S1 and S2), and unsupervised (U1 and U2) levels, for the two types of rankings, for the female (F-writers) and male (M-writers) writer datasets.

F-writers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
G1	a23	a17	a2	a16	a6	a1	a3	a9	a7	a19	a18	a11	a22	a15	a8	a0	a21	a13	a4	a20	a5	a12	a14	a10
G2	a23	a17	a16	a2	a3	a6	a1	a18	a9	a7	a8	a19	a13	a0	a11	a15	a22	a20	a5	a4	a21	a10	a14	a12
S1	a23	a16	a7	a2	a6	a3	a1	a9																
S2	a23	a16	a7	a6	a2	a3	a10	a13	a5															
U1	a23	a17	a2	a16	a1	a6	a3	a9	a19	a18	a11	a22	a15	a8	a0	a21	a13	a4	a20	a5	a7	a12	a14	a10
U2	a23	a17	a16	a2	a3	a1	a6	a18	a9	a8	a7	a19	a0	a13	a11	a15	a22	a20	a5	a4	a21	a10	a14	a12
M-writers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
G1	a23	a17	a2	a16	a6	a3	a1	a7	a9	a18	a8	a11	a15	a19	a22	a13	a0	a20	a10	a5	a14	a21	a4	a12
G2	a23	a17	a16	a2	a6	a3	a1	a18	a9	a11	a13	a8	a0	a7	a19	a20	a22	a4	a5	a21	a15	a14	a12	a10
S1	a23	a16	a7	a2	a6	a3	a10	a13																
S2	a23	a16	a7	a6	a2	a8	a3	a5																
U1	a23	a17	a2	a16	a6	a3	a1	a9	a18	a8	a11	a15	a19	a22	a0	a13	a7	a20	a5	a14	a10	a21	a4	a12
U2	a23	a17	a16	a2	a3	a6	a1	a18	a9	a11	a13	a0	a8	a19	a20	a22	a7	a4	a21	a15	a5	a14	a12	a10

Compared to unsupervised and global rankings, supervised rankings contain fewer attributes, i.e. 8 and 9, respectively for the female and male writers datasets. The first 3 positions in these rankings are occupied by the same attributes, given the two different ways of determining their weights. It can be noted that the attributes appearing in the first two ranking positions, for both methods determining the attribute weights, are the same, and they differ more in the lower ranking positions.

Table 2 presents information on the structure of the decision trees, based on which the rankings were created at the lower level of hierarchy. For all discretised variants of the datasets, the depth of the tree (row D) that corresponds to the longest path from the root to the leaf node, and the number of nodes (row N) are given.

Table 2. Structure of the decision trees, for the female (F-writers) and male (M-writers) writer datasets.

			Equal frequency binning									Equal width binning								
Nr	dsK	dsF	2	3	4	5	6	7	8	9	10	2	3	4	5	6	7	8	9	10
F-writers																				
D	6	6	7	7	6	6	5	5	5	6	6	7	7	6	6	6	7	6	7	6
N	11	11	21	10	17	12	12	12	13	13	10	25	14	15	13	15	16	15	16	15
M-writers																				
D	6	6	7	7	6	6	5	5	5	6	6	7	7	6	6	6	7	6	7	6
N	11	11	21	10	17	12	12	12	13	13	10	25	14	15	13	15	16	15	16	15

For supervised discretisation methods, the decision trees had the same structure, i.e. the same number of nodes and depth. Differences in tree structure are more apparent for unsupervised methods. Decision trees with the shortest depth were obtained for equal frequency binning with 6, 7, and 8 bins. In general, the tree depth and the number of nodes were comparable for the female and male writer datasets.

4.2. Performance of Classifiers

The rankings obtained were validated using the XGBoost model because of its efficiency and popularity. For each attribute in a ranking, the backward elimination technique was applied by sequentially removing the attributes from the lower ranking position and checking the accuracy of the classification for the model with the reduced number of features. The accuracy was calculated as the number of properly recognised samples relative to the cardinality of the test set. Due to the fact that two test sets were used in the study, the values presented in the tables are averages. The performance of classifiers obtained for sets with 24 features was considered as the reference level for each variant of the discretised datasets.

Tables 3 and 4 present the classification accuracy obtained at the lower level of hierarchy for the two types of rankings constructed for both datasets. For each attribute position (column Nr), the first row indicates the classification accuracy obtained for the proposed method of weight calculation, and the second row indicates the classification accuracy for the method implemented in the Python library. The values in bold denote an accuracy equal to or greater than the reference one obtained with the minimum number of features. The coloured cells denote the maximum accuracy detected.

Table 3. Average classification accuracy of inducers obtained for all variants of discretised data for the female writer dataset.

			Equal frequency binning									Equal width binning								
Nr	dsK	dsF	2	3	4	5	6	7	8	9	10	2	3	4	5	6	7	8	9	10
1	0.89	0.89	0.88	0.83	0.88	0.88	0.88	0.88	0.89	0.90	0.88	**0.87**	0.89	0.87	0.89	0.87	0.89	0.87	0.88	0.87
	0.89	0.89	0.88	0.83	0.88	0.88	0.88	0.88	0.89	0.90	0.88	**0.87**	0.89	0.87	0.89	0.87	0.89	0.87	0.88	0.87
2	0.91	0.89	0.87	0.83	0.87	0.84	0.89	0.87	**0.94**	0.88	0.83	**0.87**	0.89	0.88	0.89	0.91	0.84	0.88	**0.93**	0.86
	0.89	0.89	0.87	0.83	0.87	0.93	0.84	0.86	0.91	0.88	0.89	**0.87**	0.89	0.87	0.88	0.86	**0.93**	0.88	**0.93**	0.91
3	0.91	0.91	0.87	**0.92**	0.88	0.89	0.88	0.91	**0.94**	0.91	0.90	**0.87**	**0.91**	0.89	0.87	0.90	**0.92**	0.88	**0.93**	0.89
	0.91	0.91	0.87	**0.92**	0.89	0.94	0.88	0.89	0.89	0.90	0.90	**0.87**	**0.91**	0.87	0.81	0.82	**0.92**	0.86	**0.94**	**0.93**
4	0.91	0.91	0.85	**0.92**	0.91	0.91	0.89	0.91	0.92	0.96	0.91	**0.87**	0.82	0.90	0.87	0.89	**0.93**	0.89	**0.92**	0.89
	0.92	0.91	0.84	**0.93**	0.90	**0.97**	0.88	0.91	0.92	0.91	0.92	**0.87**	0.82	0.87	0.86	0.89	**0.93**	0.87	**0.92**	**0.93**
5	0.91	0.91	0.85	**0.91**	0.91	0.91	0.88	0.91	0.91	0.96	0.92	**0.88**	0.84	0.88	0.87	0.90	**0.95**	0.90	**0.92**	0.89
	0.91	0.91	0.85	**0.95**	0.91	**0.98**	0.89	0.91	0.91	0.91	0.89	**0.81**	0.85	0.90	0.84	0.91	**0.95**	0.91	**0.92**	0.91
6	0.94	0.84	0.85	**0.89**	0.90	0.95	0.88	0.89	0.91	0.96	0.92	**0.88**	0.86	0.88	0.84	0.90	**0.97**	0.91	**0.93**	**0.93**
	0.91	0.91	0.87	**0.93**	0.85	**0.98**	0.88	0.88	0.91	0.91	0.89	**0.87**	**0.94**	0.88	0.87	0.90	**0.94**	**0.92**	**0.92**	**0.93**
7	0.83	0.84	0.87	0.86	0.93	0.95	0.88	0.89	0.91	**0.97**	0.92	**0.88**	0.86	0.87	0.86	0.88	**0.97**	0.91	**0.96**	**0.92**
	0.93	0.89	0.87	**0.91**	0.87	**0.98**	0.91	0.91	0.91	0.90	0.89	**0.87**	**0.95**	0.90	0.86	0.89	**0.94**	0.89	**0.92**	**0.92**
8			0.88	0.85	0.90	**0.97**	0.89	0.94	0.91	0.96	0.89	**0.88**	**0.96**	0.87	0.88	0.89	**0.96**	**0.92**	**0.97**	**0.92**
	0.89		0.88	**0.90**	0.89	**0.99**	0.89	0.94	0.91	0.91		**0.87**	**0.95**	0.88	0.84	0.89	**0.94**	0.90	**0.92**	**0.92**
9			0.88	**0.89**	0.90	**0.98**	0.89	0.93	0.91			**0.88**	**0.96**	0.85	0.88	0.90	**0.97**	0.90	**0.97**	**0.93**
			0.89	**0.88**	0.91	**0.98**	0.89	0.94		0.96		**0.87**	**0.96**	0.88	0.86	0.93	**0.98**	0.90	**0.93**	**0.92**
10			0.88		0.91	**0.99**	0.88					**0.88**	**0.95**	0.86	0.88	0.91	**0.97**	0.90	**0.96**	**0.93**
			0.89		0.92			0.93		0.96		**0.88**	**0.96**	0.88	0.88	0.93	**0.96**	0.91	**0.92**	**0.93**
11			0.89		0.92							**0.88**	**0.95**	0.86		0.91	**0.97**	**0.92**	**0.96**	**0.93**
			0.89		0.92							**0.88**	**0.96**	0.87	0.88	**0.94**	**0.97**	0.91	**0.92**	**0.94**
12			0.89		0.92							**0.88**		0.84		0.92	**0.96**		**0.97**	
			0.89		0.92							**0.88**	**0.96**	0.88			**0.97**	**0.93**		**0.94**
13			0.89		0.93							**0.88**							**0.97**	
			0.91		0.93							**0.88**		0.87			**0.97**			**0.94**
14			0.89		0.92							**0.87**								
			0.91		0.93							**0.88**								
15					0.93							**0.87**								
16					0.92							**0.88**								
24	**0.96**	**0.96**	**0.96**	**0.88**	**0.97**	**0.97**	**0.96**	**0.98**	**0.94**	**0.97**	**0.95**	**0.81**	**0.91**	**0.94**	**0.92**	**0.94**	**0.92**	**0.92**	**0.92**	**0.92**
	0.96	**0.96**	**0.96**	**0.88**	**0.97**	**0.97**	**0.96**	**0.98**	**0.94**	**0.97**	**0.95**	**0.81**	**0.91**	**0.94**	**0.92**	**0.94**	**0.92**	**0.92**	**0.92**	**0.92**

It can be observed that for both methods of weight calculation, the reference values were comparable. For the female writers, the proposed weight, and equal width binning method with 2 bins, all attributes in the rankings led to greater accuracy than the reference values. For the equal frequency binning with 8 bins used as the discretisation method, instead of 24 attributes, it was enough to use only attributes from the three highest positions in the ranking to obtain the accuracy equal to 94%, which is greater than the reference value. The highest classification accuracy of 99% was obtained for both methods of attribute weighting, for the equal frequency binning with five bins, taking into account, respectively, 8 and 10 attributes out of 24, and exceeding the reference value by 2%.

For the male writer dataset, the classification accuracy equal to 92% was obtained for the proposed weight taking into account only 4 attributes from the ranking obtained for the equal width binning with 7 bins. It should also be noted that in the case of equal width binning with 2 and 7 bins it was enough to use the attribute only from the first position in the ranking and obtain accuracy greater than the reference value, for both methods of weight calculation.

For both datasets and for the supervised methods, it was not possible to achieve the classification accuracy at the reference level for fewer than 24 attributes. For the female writer dataset, the smallest difference for the Kononenko algorithm was 2% and for Fayyad and Irani it was 5%. For the male writer dataset, these differences were greater: for the Kononenko algorithm it was 11% and for Fayyad and Irani 2%. It is also worth noting that, in the case of unsupervised methods, the rankings at the lowest level contained fewer than 24 attributes. The maximum

Table 4. Average classification accuracy of inducers obtained for all variants of discretised data for the male writer dataset.

			Equal frequency binning									Equal width binning								
Nr	dsK	dsF	2	3	4	5	6	7	8	9	10	2	3	4	5	6	7	8	9	10
1	0.50	0.66	0.88	0.80	0.88	0.84	0.88	0.86	0.87	0.86	0.88	**0.82**	0.83	0.82	0.89	0.82	**0.88**	0.82	0.87	0.89
	0.50	0.66	0.88	0.80	0.88	0.84	0.88	0.86	0.87	0.86	0.88	**0.82**	0.83	0.82	0.89	0.82	**0.88**	0.82	0.87	0.89
2	0.50	0.66	0.88	0.86	0.86	0.87	**0.89**	0.83	**0.91**	0.86	0.84	**0.83**	0.88	0.83	0.77	0.84	**0.86**	0.89	**0.89**	0.88
	0.50	0.66	0.88	0.86	**0.91**	0.89	0.88	0.88	0.88	0.86	0.88	0.76	0.88	0.82	0.86	0.84	**0.91**	0.83	**0.89**	0.88
3	0.50	0.66	0.87	0.87	0.86	0.86	0.88	0.87	**0.91**	0.89	0.88	0.80	0.89	0.86	0.76	0.85	0.84	0.87	**0.92**	0.87
	0.50	0.66	0.87	0.87	0.86	0.91	0.88	0.87	**0.91**	0.88	0.88	0.80	0.89	0.79	0.87	0.83	0.84	0.81	**0.92**	0.87
4	0.50	0.66	**0.90**	0.88	0.86	0.90	**0.90**	0.88	0.89	0.89	0.89	**0.81**	0.87	0.81	0.82	0.86	**0.92**	0.88	**0.91**	0.89
	0.50	0.77	0.89	0.87	**0.89**	0.91	**0.90**	0.85	**0.93**	0.88	0.88	0.80	0.87	0.81	0.82	0.86	**0.88**	0.82	**0.91**	0.89
5	0.50	0.77	0.88	0.90	0.88	**0.92**	**0.92**	0.88	**0.91**	0.90	0.89	**0.83**	0.88	0.83	0.84	0.86	**0.89**	0.86	**0.91**	0.89
	0.50	0.77	**0.90**	0.86	**0.89**	**0.93**	**0.91**	0.84	**0.92**	0.88	0.89	**0.81**	0.86	0.82	0.85	**0.90**	**0.89**	0.86	**0.92**	0.90
6	0.50	0.79	0.89	**0.91**	**0.89**	**0.93**	**0.91**	0.91	**0.91**	**0.92**	0.87	**0.87**	0.89	0.84	0.90	0.87	**0.91**	0.84	**0.92**	0.91
	0.84	0.84	**0.91**	0.88	**0.89**	**0.92**	**0.89**	0.88	**0.93**	0.91	0.92	**0.83**	0.88	0.85	0.89	**0.89**	**0.91**	0.90	**0.94**	0.89
7	0.84	0.92	0.89	0.89	**0.89**	**0.93**	**0.92**	0.92	**0.91**	**0.92**	0.92	**0.83**	0.88	0.86	0.91	**0.90**	**0.90**	0.86	**0.92**	**0.92**
	0.84	0.84	**0.91**	0.87	**0.89**	**0.94**	**0.91**	0.88	**0.93**	0.91	0.92	**0.83**	**0.90**	0.84	0.87	**0.90**	**0.92**	**0.91**	**0.94**	0.87
8			**0.90**	**0.91**	**0.92**	**0.92**	**0.93**	0.91	**0.92**	0.91		**0.84**	0.89	0.85	0.91	**0.90**	**0.90**	0.87	**0.93**	**0.92**
		0.92	**0.91**	0.88	**0.89**	**0.93**	**0.92**	0.89	**0.94**		0.92	**0.82**	**0.90**	0.83	0.89	**0.90**	**0.93**	**0.92**	**0.93**	0.89
9			0.89		**0.94**	**0.93**	**0.91**		**0.92**	**0.92**		**0.86**	**0.90**	0.83	0.91	**0.91**	**0.89**	0.87	**0.93**	**0.94**
			0.89	0.90	**0.90**	**0.92**	**0.92**	0.91	**0.93**		**0.94**	**0.83**	0.89	0.84	0.92	**0.91**	**0.93**	**0.91**	**0.93**	**0.92**
10			**0.91**		**0.93**	**0.92**						**0.82**	**0.91**	0.84	0.90	**0.92**	**0.89**	**0.92**		**0.95**
			0.90		**0.91**			0.91				**0.82**	**0.90**	0.84	0.92	**0.91**	**0.91**	**0.91**	**0.93**	0.90
11			0.89		**0.92**							**0.83**		0.86		**0.92**	**0.86**	**0.91**		**0.93**
			0.92		**0.94**							**0.84**	**0.90**	0.85		**0.92**	**0.93**	**0.91**	**0.94**	0.90
12			0.89		**0.92**							**0.86**		**0.88**		**0.92**	**0.88**	0.89		**0.93**
			0.93		**0.93**							**0.84**		0.83			**0.91**		**0.94**	0.91
13			**0.92**		**0.93**							**0.86**				**0.92**	**0.91**			**0.94**
			0.93		**0.92**							**0.84**					**0.92**			
14			**0.91**		**0.92**							**0.86**					**0.91**			
												0.84								
15			**0.91**		**0.92**												**0.92**			
												0.84								
16																	**0.92**			
24	**0.95**	**0.94**	**0.90**	**0.91**	**0.89**	**0.92**	**0.89**	**0.93**	**0.91**	**0.92**	**0.93**	**0.81**	**0.90**	**0.88**	**0.94**	**0.88**	**0.86**	**0.91**	**0.88**	**0.92**
	0.95	**0.94**	**0.90**	**0.91**	**0.89**	**0.92**	**0.89**	**0.93**	**0.91**	**0.92**	**0.93**	**0.81**	**0.90**	**0.88**	**0.94**	**0.88**	**0.86**	**0.91**	**0.88**	**0.92**

length of the ranking is 16 in both the female and male writer groups, but the rankings at the higher level of the hierarchy contained 24 attributes, demonstrating the diversity of attributes in these rankings, ensuring that their ordering at the higher level produced a full set of attributes.

Table 5 presents the classification accuracy obtained for the supervised (rows S1 and S2), unsupervised (U1 and U2), and global (G1 and G2) rankings, for two types of weights, for the female and male writer datasets. Index 1 assigned to the ranking name refers to the proposed weight, and index 2 to the weight implemented in the Python library. The first row of the table lists the position of the attribute in the ranking. Values in bold denote the accuracy equal to or greater than the reference one. The coloured cells denote the maximum accuracy.

Table 5. Average classification accuracy of inducers obtained for supervised, unsupervised and global rankings, for two types of weights, for the female and male writer datasets.

F-writers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
G1	0.88	0.9	0.9	0.89	0.9	0.92	0.91	0.91	0.91	0.91	0.91	0.92	**0.95**	**0.94**	**0.94**	**0.94**	**0.94**	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**0.93**
G2	0.88	0.9	0.9	0.89	0.87	0.89	0.91	0.91	0.91	0.91	0.92	0.92	0.92	0.92	**0.93**	**0.93**	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**0.93**
S1	0.89	0.89	0.91	0.91	0.91	0.89	0.84	0.84																**0.93**
S2	0.89	0.89	0.91	0.91	0.91	0.89	0.89	0.89	0.89															**0.93**
U1	0.88	0.9	0.9	0.89	0.92	0.92	0.92	0.92	0.92	0.92	0.92	**0.95**	**0.94**	**0.94**	**0.94**	**0.94**	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**0.93**
U2	0.88	0.9	0.9	0.89	0.88	0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92	**0.93**	**0.94**	**0.94**	**0.95**	**0.95**	**0.95**	**0.95**	**0.95**	**0.94**	**0.94**	**0.93**
M-writers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
G1	0.83	0.89	0.86	0.86	0.88	**0.90**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.90**
G2	0.83	0.89	0.87	0.86	0.88	**0.90**	**0.92**	**0.92**	**0.92**	**0.93**	**0.93**	**0.92**	**0.93**	**0.92**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.90**
S1	0.58	0.58	0.58	0.58	0.63	0.71	0.88	0.88																**0.90**
S2	0.58	0.58	0.58	0.63	0.63	0.84	0.88	0.88																**0.90**
U1	0.85	0.88	0.85	0.86	0.88	**0.90**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.93**	**0.92**	**0.93**	**0.93**	**0.93**	**0.93**	**0.90**
U2	0.85	0.88	0.87	0.86	0.89	**0.90**	**0.92**	**0.92**	**0.92**	**0.93**	**0.93**	**0.93**	**0.92**	**0.92**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.90**

Considering the quality of the classification for the female writer dataset, it can be observed in the global rankings that there is little difference between the values of G1 and G2. It can also be seen that the maximum classification accuracy of 95% can be obtained by taking into account only 13 attributes from the ranking, using the weight proposed by the authors. In the case of

G2, the same level of predictions was obtained with 17 attributes from the ranking. A similar trend can be observed for the unsupervised rankings U1 and U2. For the male writer dataset, for both G1 and G2, and U1 and U2, the advantage of Python-implemented weight can be detected. However, it should be noted that the accuracy equal to 90%, the same as the reference point, was obtained with only 6 out of the 24 attributes.

The graphical representation of the G1 and G2 rankings for the male and female writer datasets is presented in Figures 1 and 2. X axis is labelled with the names of the test sets corresponding to the discretisation algorithms. Y axis shows the number of attributes considered. Presented average classification accuracy values greater than or equal to the reference values for the female writers are marked in green and in blue for the male writers. Considering the colour intensity, for the female writer dataset, the green colour is more visible with G1. For the male writer dataset, the colour intensity between G1 and G2 is comparable.

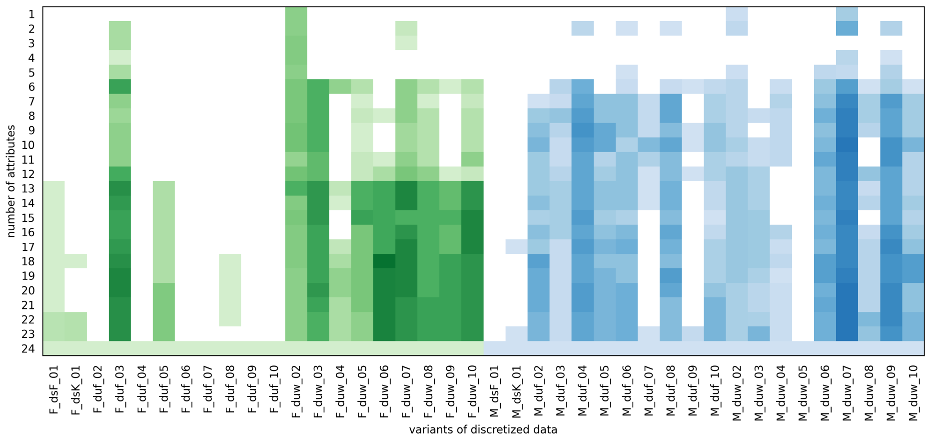

Fig. 1. Average accuracy based on ranking G1

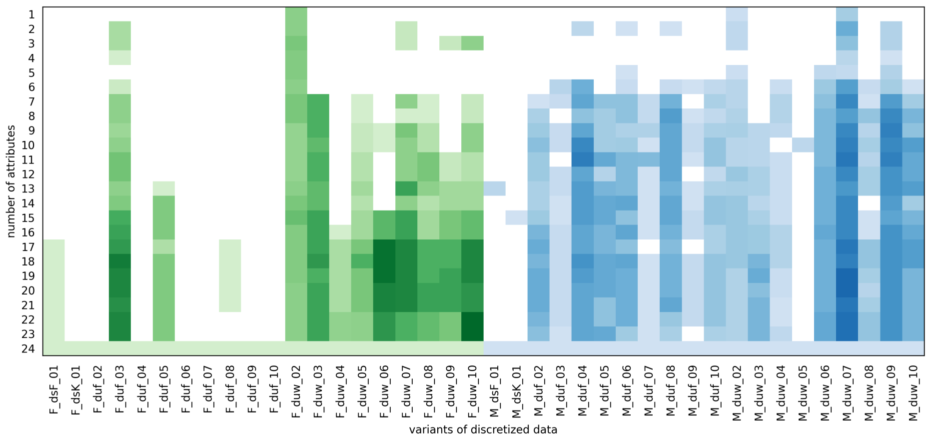

Fig. 2. Average accuracy based on ranking G2

4.3. Summary and Discussion of Obtained Results

The summary of the experiments is presented in Table 6. It lists the maximum classification accuracy obtained for the highest possible position of the attribute in the ranking. The results are related to two methods of weight calculation for the hierarchy of rankings at the supervised (columns S1 and S2), unsupervised (columns U1 and U2), and the global level (columns G1 and

G2). Index 1 assigned to the ranking name refers to the proposed weight, and index 2 refers to the weight implemented in the Python library. The last row of the table includes average values for each column considered. Colored values indicate the best average classification accuracy achieved among different rankings, for a given discretisation method.

Table 6. The best performance obtained for supervised, unsupervised, and global rankings.

	F-writers				M-writers				F-writers				M-writers				F-writers				M-writers			
	G1		G2		G1		G2		U1		U2		U1		U2		S1		S2		S1		S2	
	Pos	Acc	Pos	Acc	Pos	Acc	Pos	Acc	Pos	Acc	Pos	Acc	Pos	Acc	Pos	Acc	Pos	Acc	Pos	Acc	Pos	Acc	Pos	Acc
dsF	22	0.96	**17**	**0.96**	24	0.95	**13**	**0.96**									24	0.96	24	0.96	24	0.95	24	0.95
dsK	**22**	**0.97**	24	0.96	17	0.94	**15**	**0.94**									24	0.96	24	0.96	24	0.94	24	0.94
duf2	**24**	**0.96**	**24**	**0.96**	18	0.94	**17**	**0.94**	**24**	**0.96**	**24**	**0.96**	18	0.94	19	0.94								
duf3	19	0.97	**18**	**0.97**	**8**	**0.93**	6	0.92	**18**	**0.97**	**18**	**0.97**	6	0.92	6	0.92								
duf4	**24**	**0.97**	**24**	**0.97**	9	0.95	11	0.96	**24**	**0.97**	**24**	**0.97**	8	0.95	**12**	**0.97**								
duf5	20	0.98	**14**	**0.98**	**9**	**0.95**	11	0.95	19	0.98	**14**	**0.98**	17	0.94	11	0.95								
duf6	**24**	**0.96**	**24**	**0.96**	17	0.93	**14**	**0.94**	**24**	**0.96**	**24**	**0.96**	17	0.93	15	0.94								
duf7	**24**	**0.98**	**24**	**0.98**	10	0.95	11	0.95	**24**	**0.98**	**24**	**0.98**	9	0.94	**11**	**0.95**								
duf8	18	0.94	**17**	**0.94**	19	0.95	**8**	**0.95**	**17**	**0.94**	**17**	**0.94**	16	0.95	**8**	**0.95**								
duf9	**24**	**0.97**	**24**	**0.97**	23	0.93	**9**	**0.93**	**24**	**0.97**	**24**	**0.97**	16	0.93	16	0.93								
duf10	**24**	**0.95**	**24**	**0.95**	**21**	**0.96**	10	0.95	**24**	**0.95**	**24**	**0.95**	**21**	**0.96**	10	0.95								
duw2	13	0.91	15	0.89	17	0.86	**14**	**0.86**	**12**	**0.91**	15	0.89	17	0.86	17	0.86								
duw3	13	0.97	18	0.97	23	0.93	**19**	**0.94**	**12**	**0.97**	18	0.97	23	0.93	**19**	**0.94**								
duw4	6	0.96	22	0.96	**7**	**0.89**	8	0.89	**5**	**0.96**	22	0.96	**7**	**0.89**	8	0.89								
duw5	15	0.97	18	0.96	24	0.94	**10**	**0.95**	**12**	**0.97**	18	0.96	24	0.94	**10**	**0.95**								
duw6	18	0.99	**17**	**0.99**	18	0.93	16	0.93	**17**	**0.99**	**17**	**0.99**	18	0.93	**15**	**0.93**								
duw7	13	0.98	17	0.98	10	0.95	**19**	**0.96**	**12**	**0.98**	17	0.98	20	0.96	22	0.96								
duw8	21	0.97	20	0.97	**22**	**0.94**	8	0.93	**17**	**0.97**	20	0.97	**22**	**0.94**	8	0.93								
duw9	18	0.97	19	0.97	17	0.95	**9**	**0.96**	**14**	**0.97**	19	0.97	**9**	**0.96**	12	0.96								
duw10	15	0.98	22	0.99	18	0.96	**10**	**0.96**	14	0.98	**22**	**0.99**	18	0.96	**10**	**0.96**								
average		0.97		0.96		0.94		0.94		0.97		0.96		0.94		0.94		0.97		0.96		0.95		0.95

It should be noted that global rankings were tested on all test data, that is, 20 sets times two, as each dataset contains two test sets. In the case of unsupervised and supervised rankings, only test sets occurring within the framework of a given approach were examined. Thus, for the ranking involving unsupervised discretisation algorithms, 18 variants of the set times two were tested; for the ranking involving supervised discretisation algorithms, two variants of the set were tested, each with two test sets.

It can be observed that for the proposed weight calculation method, the results obtained were slightly better or comparable with the existing one already implemented in the Python library, taking into account the average classification accuracy. The highest value for the female writer dataset and the global ranking was the same for G1 and G2, and equal to 99%. It was obtained for the same discretisation algorithm, i.e., duw6. In the case of the male writer dataset, the highest classification accuracy for G1 and G2 was equal to 96%. In both types of rankings, it was obtained for different discretisation algorithms.

The differences in the classification accuracy between the proposed weight and the weight implemented in Python, obtained at the global level of ranking hierarchy, are presented in Figure 3. At the bottom of the figure, the number of selected features is indicated and shows the position of the attributes in the ranking, observed in backward feature reduction process. The left side of the figure shows the discretisation methods, which, in the case of unsupervised algorithms, presents an averaged set of results. Green shades indicate higher classification accuracy obtained for the proposed method of weighting attributes, red shades – higher results for scikit-learn libraby. The more intense the color, the greater the difference.

It is visible that the proposed method significantly improves classification accuracy for female writers, especially for supervised dataset variants (F_dsF, F_dsK), with improvements up to 0.056, notably when selecting 5 to 8 features, highlighting its effectiveness with limited attributes. For male datasets, differences were generally small, but in some cases, scikit-learn performed better classification accuracy (e.g., Fayyad and Irani's method (M_dsF) with 8 and 10 features, differences up to –0.083). However, scikit-learn's advantage was isolated, while the proposed method more often yielded superior, consistent results, especially for female datasets.

Although the hierarchical structure of the global feature ranking procedure takes into account the relevant level in the hierarchy, the number of datasets discretised using unsupervised

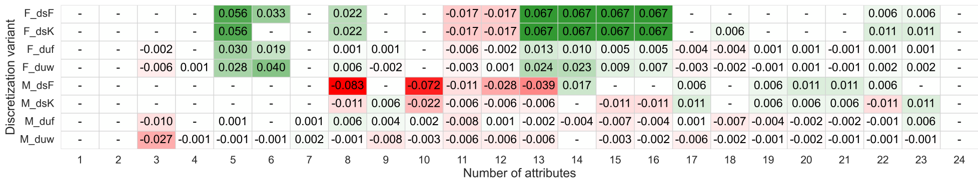

Fig. 3. Mean differences in the classification accuracy between G1 and G2

and supervised methods can affect the resulting global ranking as shown in Table 6. Having a more balanced number of datasets for each discretisation approach (supervised and unsupervised) could potentially yield a more objective and generalizable global ranking. This limitation will be addressed in future work.

5. Conclusions

The lack of explicit guidance on the choice of algorithms for the input data discretisation and the need for feature reduction due to the interpretability requirements of the data models motivated the authors to propose the global ranking construction method. This approach, based on a hierarchical structure, takes into account the characteristics and specificities of sets discretised by different categories of algorithms.

An important element in the data preparation process is the selection of the attributes that are most relevant and have the greatest impact on the decisions proposed by the model. One solution to this is to create a ranking of attributes, where dimensionality reduction can be performed based on their importance. The interpretability of machine learning models, especially in this era of developing various artificial intelligence techniques, plays an important role. With this factor in mind, the authors proposed an assessment of the importance of attributes based on the structure of the decision tree, which is advantageous due to the transparency of the decision-making process. The proposed attribute weights were compared with a measure implemented in the Python library based on an impurity analysis of the decision tree nodes. Both attribute importance measures were used to create attribute rankings that ensure universality with respect to discretisation methods and algorithms.

The experiments were carried out on stylometric datasets prepared for the authorship attribution task. The results obtained show a significant reduction in the number of attributes for many cases of classification accuracy higher than the reference level, considering the entire set of attributes. The method developed and the approach to the hierarchical structure of the global ranking creation were tested at all hierarchy levels for 20 variants of discretised datasets for the female and male writer datasets, respectively, and verified by the process of backward elimination of attributes. These results demonstrate the validity and merit of the proposed approach.

The presented research methodology can be applied in the development of more efficient and interpretable information systems. This approach enhances decision-making processes in applications such as recommendation, diagnostic systems, and automated decision support.

In future research, attribute weights will be developed based on heterogeneous sets of machine learning models. The efficiency of the proposed ranking methods will also be compared with other approaches to attribute evaluation.

Acknowledgements

The research activities were co-financed by the funds granted under the Research Excellence Initiative of the University of Silesia in Katowice. Research reported in this paper was carried out within the statutory project of the Department of Computer Graphics, Vision and Digital

Systems (RAU-6, 2025), at the Silesian University of Technology, Gliwice, Poland, and was supported by the King Abdullah University of Science and Technology (KAUST), Saudi Arabia.

References

[1] Bolón-Canedo, V., Sechidis, K., Sánchez-Maroño, N., Alonso-Betanzos, A., Brown, G.: Insights into distributed feature ranking. Information Sciences 496, pp. 378–398 (2019)

[2] Elhilbawi, H., Eldawlatly, S., Mahdi, H.: A taxonomy of discretization techniques based on class labels and attributes' relationship. In: 2019 14th International Conference on Computer Engineering and Systems (ICCES) (2019)

[3] Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuousvalued attributes for classification learning. In: 13th International Joint Conference on Articial Intelligence. vol. 2, pp. 1022–1027. Morgan Kaufmann Publishers (1993)

[4] Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. (eds.): Feature Extraction: Foundations and Applications, Studies in Fuzziness and Soft Computing, vol. 207. Springer (2006)

[5] He, X., Lashkari, A.H., Vombatkere, N., Sharma, D.P.: Authorship attribution methods, challenges, and future research directions: A comprehensive survey. Information 15(3) (2024)

[6] He, Z., Wang, J., Jiang, M., Hu, L., Zou, Q.: Random subsequence forests. Information Sciences 667, pp. 120478 (2024)

[7] Kononenko, I.: On biases in estimating multi-valued attributes. In: 14th International Joint Conference on Articial Intelligence. pp. 1034–1040 (1995)

[8] Marudi, M., Ben-Gal, I., Singer, G.: A decision tree-based method for ordinal classification problems. IISE Transactions 56(9), pp. 960–974 (2024)

[9] Moshkov, M., Zielosko, B., Tetteh, E.T., Glid, A.: Learning decision rules from sets of decision trees. In: Buchmann, R.A., Silaghi, G.C., Bufnea, D., Niculescu, V., Czibula, G., Barry, C.L., Lang, M., Linger, H., Schneider, C. (eds.) Information Systems Development: Artificial Intelligence for Information Systems Development and Operations (ISD2022 Proceedings), Cluj-Napoca, Romania, 31 August - 2 September 2022. Risoprint / Association for Information Systems (2022)

[10] Stańczyk, U., Zielosko, B.: On combining discretisation parameters and attribute ranking for selection of decision rules. In: Polkowski, L., Yao, Y., Artiemjew, P., Ciucci, D., Liu, D., Slezak, D., Zielosko, B. (eds.) Rough Sets - International Joint Conference, IJCRS 2017, Olsztyn, Poland, July 3-7, 2017, Proceedings, Part I. Lecture Notes in Computer Science, vol. 10313, pp. 329–349. Springer (2017)

[11] Stańczyk, U., Zielosko, B.: Heuristic-based feature selection for rough set approach. International Journal of Approximate Reasoning 125, pp. 187–202 (2020)

[12] Wissuchek, C., Zschech, P.: Prescriptive analytics systems revised: a systematic literature review from an information systems perspective. Information Systems and e-Business Management (2024)

[13] Witten, I., Frank, E., Hall, M.: Data Mining. Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 3rd edn. (2011)

[14] Zollanvari, A.: Decision trees. In: Machine Learning with Python: Theory and Implementation, pp. 187–207. Springer International Publishing, Cham (2023)