Going Deeper than Supervised Discretisation in Processing of Stylometric Features

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

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

Grzegorz Baron
Silesian University of Technology
Department of Graphics Computer Vision and Digital Systems
Gliwice, Poland *grzegorz.baron@polsl.pl*

Abstract

Rough set theory is employed in cases where data are incomplete and inconsistent and an approximation of concepts is needed. The classical approach works for discrete data and allows only nominal classification. To induce the best rules, access to all available information is advantageous, which can be endangered if discretisation is a necessary step in the data preparation stage. Discretisation, even executed with taking into account class labels of instances, brings some information loss. The research methodology illustrated in this paper is dedicated to extended transformations of continuous input features into categorical, with the goal of enhancing the performance of rule-based classifiers, constructed with rough set data mining. The experiments were carried out in the stylometry domain, with its key task of authorship attribution. The obtained results indicate that supporting supervised discretisation with elements of unsupervised transformations can lead to enhanced predictions, which shows the merits of the proposed research framework.

Keywords: Discretisation, Stylometry, Decision Rules, Rough Set Theory

1. Introduction

Rough set theory (RST) is employed in data mining when knowledge is uncertain and incomplete, when approximations of concepts are needed [18]. The classical version allows only nominal classification, and the universe of discourse is expected to occupy a discrete domain [17]. Decision rules, induced by one of many available algorithms, can be used as a form of representation of the discovered knowledge [12]. On the other hand, conditions included in rules describe patterns detected and can be applied to unknown samples to label them. A set of rules can be used as a classifier, with the support of some conflict resolution strategy for cases where several rules with conflicting decisions match a sample [15], [22].

When an application domain brings with it real-valued attributes, discretisation can be executed inside the data preparation step, to transform continuous values into categorical [13]. Here, supervised methods lead the way, as they take into account information about class labels when cut-points are established to divide the range of attribute values into bins. Algorithms of-

ten rely on the Minimum Description Length Principle [11] and measures such as entropy to find the best solution. However, in top-down processing, the transformation begins with assigning a single interval to represent the entire range of values for attributes. Depending on the chosen criteria, the recursive discretisation procedure can stop right after the beginning, leaving such 1-bin representation as final for some variables. As a consequence, these attributes are effectively removed from the considerations and their informative content is completely ignored.

Instead of blindly and unconditionally relying on supervised discretisation, it is possible to employ extended processing of problematic features [21]. Then, discretisation can be treated as a two-step process. The first stage is dedicated to a supervised algorithm, at the end of which the available attributes are divided into two groups: the one characterised by multiple intervals established to represent their values, and the second with single bins. The former features are accepted as transformed, while the latter are passed on to the second stage, where the domains are discretised by unsupervised algorithms. As a result, all variables have more than one categorical representation and even this information, which was found to be irrelevant and negligible by supervised procedures, is kept to some extent.

The described deeper transformations of some features, going beyond what can be obtained by the supervised discretisation approaches, were employed to prepare the input space for mining with the RST inside the stylometry domain, with its key task of authorship attribution [14]. Style-markers used in such tasks often refer to lexical and syntactic properties of writing styles and return the continuous characteristics of the linguistic profiles of the studied authors [30]. Inferring decision rules from such data would be possible [24], yet problematic and computationally expensive. The translation into a discrete domain not only enables rough set approach, but leads to induced rules with categorical conditions, which provides some insight of stylistic traits and habits that point to authors.

Within the extensive experiments performed, the input datasets were prepared in several discrete variants. From all of them sets of decision rules were inferred with the exhaustive algorithm implemented in the Rough Set Exploration System (RSES) [4]. All induced rule sets were then used as classifiers to predict authors of unknown text samples from test sets, with simple majority and standard weighted voting applied as strategies to resolve the occurring conflicts. The results of the experiments show that the extended two-step processing of the datasets in many cases brought improved predictions, which illustrates the advantages of the proposed research framework.

The paper is organised as follows. Section 2 brings forward the fundamental notions of the rough set theory. Section 3 contains comments on data transformations by discretisation algorithms. Section 4 provides an explanation of the proposed procedure for the deeper processing of attributes. Experiments are described in Section 5, and Section 6 concludes the paper.

2. Rough Set Approach to Data Mining

The rough set theory was proposed by Z. Pawlak as a method to handle incomplete and inconsistent data, as well as a way to reduce dimensionality by analysing data dependencies [19]. An important notion is knowledge granularity based on partitioning of a dataset into indiscernible classes, which aims to approximate imprecise concepts. An indiscernibility relation allows to obtain granules of knowledge about the universe U, represented as classes of objects that cannot be discerned as they are characterised by the same values of chosen attributes. The relation is defined for a subset of attributes $B \subseteq A$ and the set of objects U:

$$IND(B) = \{(x, y) \in U \times U : \forall_{a \in B} \quad a(x) = a(y)\}. \tag{1}$$

In the rough sets theory, granules of indiscernible objects are considered instead of particular objects. Any imprecise (rough) concept is replaced by a pair of precise concepts known as the lower and upper approximations of the original concept. Imprecision of a concept is indicated

by the use of a boundary region, which represents the difference between the upper and lower approximations of the concept. If the boundary region of a set is not empty, it means that knowledge about the set is limited and does not allow for its precise definition [27].

The main structure employed for data representation is a tabular form defined as a decision table, $S = (U, A \bigcup \{d\})$. U is a non-empty, finite set of objects, $A = \{a_1, \ldots, a_m\}$ is a non-empty, finite set of condition attributes, i.e., $a_i : U \rightarrow V_{a_i}$. For every $a \in A$, V_{a_i} is the set of values of the attribute a_i. $d \notin A$ is a distinguished attribute called a decision or a class label, with values $V_d = \{d_1, \ldots, d_{|V_d|}\}$.

In the rough set theory, reducts and decision rules are popular forms (objects) used in knowledge representation and classification processes. Additionally, there is a relationship between these objects: decision rules can be induced based on the reducts [31]. There are different types of decision rules, however, very often they are presented as formulas:

$$(a_{i_1} = v_1) \wedge \ldots \wedge (a_{i_k} = v_k) \rightarrow d = v_d, \tag{2}$$

where $1 \leq i_1 < \ldots < i_k \leq m, v_i \in V_{a_i}$, and $1 \leq v_d \leq |V_d|$. This form implies nominal or discrete values of conditional attributes and also nominal classification.

The concept of rough sets can be extended to a broader framework, so this theory has applications in many areas such as classification, cluster analysis, probability theory, granular computing, and many others.

3. Transformations of Input Domain

One of the important factors affecting the implementation of data mining is the proper preparation of the input data. This step comprises three key elements: (i) the use of data cleaning techniques that are employed to eliminate inconsistencies and noise present in the data, (ii) the implementation of data transformation methods, and (iii) the adoption of data reduction methods to acquire a condensed representation of a dataset.

Discretisation methods serve as a crucial technique in the framework of data reduction methods. They transform the continuous space of attribute values into discrete or nominal ones with a finite number of intervals. Through it, data becomes simpler, and potential noise can be eliminated. However, it should be noted that during the discretisation process, there may be a loss of certain information. Therefore, this stage of data preparation should be carefully considered and implemented with caution, as data irregularities can cause problems in categorical representation found, in particular, for independent transformations of multiple datasets [23].

The selection of an appropriate method is an important aspect of the discretisation. Supervised algorithms take into account information about class labels during the discretisation process, as opposed to unsupervised ones that focus entirely on attribute values. In the research, as a representative of supervised methods, the Fayyad and Irani algorithm was used [8]. It belongs to top-down approaches where finding cut-points for continuous values of a given attribute starts from one interval containing all values. Its partitioning is repeated in a recursive way until a stopping criterion is met. This algorithm is based on the class entropy of the considered intervals, calculated in the evaluation of cut-points, and the Minimum Description Length (MDL) principle [11] used as a stopping criterion. It should be noted that it is possible that all candidate cut-points will be rejected and then the initial single interval defined for the entire range of values for a given attribute will remain undivided [10].

Among the group of unsupervised discretisation algorithms, the most popular are equal width binning and equal frequency binning [5], both of which also belong to top-down approaches. Equal width binning simply constructs the requested number of bins of equal width. As it completely disregards distributions of datapoints in the input space, it is widely criticised. In the investigations, two variants of equal frequency binning were used. The equal frequency algorithm divides the range of attribute values into a specific number of intervals set by the user,

ensuring that each bin contains the same number of sorted values. Equal frequency binning with weights is a variant of simple equal frequency binning. The latter is focused more on the input parameter, which is a number of bins to be constructed, while the former studies the occurrences of values closer. As a consequence, both variants can return the same numbers of intervals, but with different cut-points.

4. Conditional Processing of Stylometric Features

The research methodology illustrated in this paper, dedicated to extended transformations of continuous input features, was applied in the domain of stylometric analysis of texts. The section presents the properties and specifics of this domain with its central task of authorship attribution and its requirements, and other elements involved in the research framework.

4.1. Characteristics of Stylometric Attributes

A style is an elusive phenomenon, better recognised intuitively than defined [2]. Writing styles are characterised by the linguistic traits, habits, and preferences of the authors. Style-markers that enable the approximation of stylistic writer profiles often refer to lexical and syntactic elements of a text [6], [30]. The frequencies of occurrence are then calculated for the selected common function words and punctuation marks, which makes the descriptors real valued.

In the research, a set of 12 two-letter function words was used, as follows: as, at, by, if, in, no, of, on, or, so, to, up. Such short words are typically employed almost without conscious thought, so the patterns of their use can be treated as reliable markers of individuality, in particular when they are observed over many samples of writing.

4.2. Authorship Attribution as Classification Task

The preparation of the input corpus for analysis is one of the tricky elements, with specific demands [9]. On one hand, the more text samples, the better, as they provide insight into style variations. As rare authors write many long texts, it is a standard operating procedure to divide texts into blocks of comparable size [7]. However, as a consequence, the groups of samples are constructed, and stratification is imposed on the input space. Inside a class corresponding to one author, sub-classes can be recognised, each including samples based on some single longer text. In such conditions cross-validation cannot be used as a reliable measure of performance for a classifier, as it returns over-optimistic results [3], [25], and the use of separate test sets is safer.

On the other hand, the texts written by authors of the same gender (including cases of non-binary gender and trans-gender) show certain common characteristics [28]. This should not be ignored as it could compromise the reliability of the analysis. To ensure as unbiased construction of an authorial profile as possible, it is better to compare writers within the same gender group.

Taking these pointers into consideration led to the construction of two datasets: Wharton-Johnston dataset (in short WJ dataset) including characteristics for texts authored by Edith Wharton and Mary Johnston, and James-Hardy dataset (in short JH dataset) for comparison of Henry James and Thomas Hardy. Each dataset included a single train set and two test sets, and each set was constructed for binary classification with balanced classes. Both classes were treated as of the same importance and with the same costs in the case of misclassification. For such conditions, classification accuracy was chosen as a suitable measure of performance [26] for evaluating inducers based on the rough set processing of the data.

4.3. Information Fusion from Supervised and Unsupervised Discretisation

The main goal of top-down discretisation approaches is to save processing time: transformations stop as soon as possible, and then minimal numbers of bins are constructed for variables. In the

case of the Fayyad and Irani method [8], which was used in the research, the stopping criterion relies on the entropy calculated for the smaller number of intervals compared to the entropy after their further division. Thus, entropy serves as a measure of informative content brought by attributes into a classification task.

However, when some features are considered irrelevant based on entropy, this does not mean that they cannot be exploited. Some unsupervised discretisation procedures [10] can be used to extend transformations, to retain more information than can be offered by supervised processing, and to obtain categorical representations for all available attributes, even those found to be less important. In the research reported in this paper, a fusion of the supervised Fayyad and Irani algorithm with unsupervised equal frequency and equal frequency with weights binning was investigated, with varying the number of bins for unsupervised approaches.

In the first part of the transformations, the input datasets were discretised individually and independently on each other by the Fayyad and Irani method. In the second stage, the variables, for which only single intervals were found through supervised processing, were further transformed by the two selected unsupervised approaches with varying bin numbers, thus resulting in construction of several discrete variants of data, with particular combinations of algorithms, with all variables discretised, and all attributes with more than one bin in a discrete domain.

4.4. Induction of Decision Rules from Stylometric Data

To induce decision rules, the Rough Set Exploration System [4] requires categorical input datasets. In the research, the exhaustive algorithm for rule construction was used. It finds all minimal decision rules available. The processing starts with an analysis of decision tables in the train sets, and the algorithm infers rules that contain minimal numbers of descriptors (pairs attribute=value) in their premise parts.

Finding all rules on examples most often results in relatively high cardinalities of rule sets, which extends processing and increases storage requirements. To avoid that, it is possible to employ some approximate approaches. They include greedy algorithms, genetic algorithms, and various types of heuristics which use different criteria based on entropy, the Gini index, statistical characteristics, and many others [16]. However, such smaller rule sets that are found by heuristic algorithms may not provide satisfactory predictions when applied to test samples.

The other processing path leads through the application of some method from the exact category of approaches for inducing decision rules. In addition to exhaustive search, in this framework Boolean reasoning and extensions of dynamic programming can be highlighted [1], [18]. After the rules are induced, they can be analysed with respect to their characteristics that betray their quality. Popular measures that allow characterising the quality of decision rules are length and support [29]. The length denotes the number of descriptors in a premise part of a rule. Support indicates the number of objects in a training set that have the same values of condition attributes and the same decision as in the rule. Experts prefer simple and short decision rules due to their ease of understanding and interpretation. Additionally, short rules have fewer storage requirements, and the time required for the classification process using such rules can be reduced. Support allows the discovery of major patterns in the data, so it is an important measure from the point of view of both knowledge representation and classification. These rule characteristics can be used in the process of rule filtering to prune rule sets [20].

5. Experimental Results

The experiments started with the preparation of the input datasets, which were then discretised by supervised, unsupervised, and combined two-level methods with varying values of the input parameters. For all discrete variants of the datasets, the classical rough set approach was used to induce decision rules by an exhaustive algorithm. Rule-based classifiers were applied to label

samples from test sets. The performance of inducers was investigated and evaluated in relation to the discretisation algorithm employed and the conflict resolution strategy applied. This section details the processing steps and the results obtained.

5.1. Supervised, Unsupervised, and Combined Discretisation of Attributes

As a consequence of top-down processing applied to attributes domains by the Fayyad and Irani algorithm, it may turn out that for some features, the single interval, assigned to represent all their values at the beginning, becomes the final categorical representation. Attributes can be seen as being characterised by discretisation, and these characteristics for the studied datasets and attributes are shown in Table 1.

Table 1. Characteristics of attributes for supervised discretisation using the Fayyad and Irani (dsF) approach of train sets.

WJ dataset		**JH dataset**	
Bins	Attributes	Bins	Attributes
1	if in no or so up	1	as of no on so to up
2	as at by	2	at if or
3	of on to	3	by in

As can be observed, for both the WJ and the JH datasets, the numbers of single-bin variables were relatively high, equal to or around 50 %. For individually and independently transformed sets, these characteristics are local and differ from set to set. This means that for some variable in one set, a single interval is used for discrete representation, while in some other set multiple bins can be found. As a consequence, discrete data models can vary so much that they have a noticeable influence on the performance of a classifier. Supporting supervised discretisation with unsupervised transformations helps to make these models closer and keep as much information in both types of processing as possible.

Unsupervised equal frequency (duf) and equal frequency with weights (dufw) binning both construct the required number of bin, but in this construction they focus on the number of instances represented by an interval for an attribute. In the experiments for both methods, the number of bins ranged from 2 to 10.

In total, for each dataset, 37 discrete variants were investigated: 1 variant from fully supervised discretisation with the Fayyad and Irani algorithm (dsF), 9 variants from fully unsupervised equal frequency binning (from duf02 to duf10), 9 variants from fully unsupervised equal frequency binning with weights (from dufw02 to duwf10), and finally 9+9=18 variants from discretisation combining supervised with the two unsupervised methods (from dsF-duf02 to dsF-duf10, and from dsF-dufw02 to dsF-dufw10).

5.2. Performance Evaluation for Rule-Based Classifiers

From all these variants of discrete data, decision rules were induced by the exhaustive algorithm. The sets of rules were used to classify samples from the corresponding variants of the test sets, employing two variants of the conflict resolution strategy in the case of several rules with conflicting decisions matching a sample [15], [22]. Simple voting denotes the case where each rule gets a single vote, regardless of the characteristics of this rule. Standard voting denotes weighted voting: each rule is given as many votes as its support.

For the case where discretisation was executed in one step, for supervised and unsupervised methods, the performance of rule-based classifiers is included in Table 2. It is worth noting that the classification accuracy for unsupported supervised discretisation was much worse than for any variant of data from unsupervised transformations. In all cases but one (duf02 for the

JH dataset), standard voting led to higher precision than simple voting. In some cases, the difference was only slight, but sometimes noticeable. These results can be treated as reference points for comparison with the extended discretisation processing.

Table 2. Performance [%] of rule-based classifiers for datasets discretised by the supervised Fayyad and Irani (dsF) approach, and unsupervised equal frequency (duf) and equal frequency with weights (dufw) binning, with varying the number of constructed bins.

	WJ dataset		**JH dataset**			**WJ dataset**		**JH dataset**	
Discret. method	Simple voting	Standard voting	Simple voting	Standard voting	Discret. method	Simple voting	Standard voting	Simple voting	Standard voting
duf02	81.53	88.82	73.06	72.50	dufw02	82.64	89.31	71.95	72.02
duf03	78.68	87.57	70.28	77.09	dufw03	79.24	87.02	70.28	76.53
duf04	85.14	89.24	75.77	79.86	dufw04	86.39	89.24	76.39	78.61
duf05	81.67	89.86	74.72	77.71	dufw05	81.11	89.79	74.72	79.59
duf06	83.82	90.91	74.03	78.61	dufw06	83.89	90.35	74.66	78.61
duf07	83.61	91.04	75.28	81.88	dufw07	82.99	91.04	75.28	81.88
duf08	87.50	90.42	76.25	80.97	dufw08	86.88	89.79	74.03	79.31
duf09	85.63	88.06	76.95	78.20	dufw09	85.07	88.06	75.28	78.20
duf10	86.39	87.02	73.75	75.56	dufw10	86.39	88.13	72.64	75.56
dsF	48.89	50.00	55.42	62.09					

For the two-step discretisation transformations, with the supervised Fayyad method combined with either unsupervised equal frequency binning (dsF-duf) or equal frequency binning with weights (dsF-dufw), the trends in performance for rule-based classifiers can be observed through plots included, respectively, in Fig. 1 and Fig. 2. In all charts, the categories for the horizontal axis reflect the numbers of bins constructed for unsupervised methods for train sets, while the series do the same, but for the test sets, and these variants were matched in the processing. The number of bins equal to 1 denotes the standard and simple one-step transformations with the supervised Fayyad method, without any additional processing. For both the JH and WJ datasets, the results were shown in the case of simple and standard voting used as conflict resolution strategies, and they were obtained as an average over test sets.

Since purely supervised discretisation of datasets led to relatively low predictions (due to relatively high numbers of disregarded 1-bin variables and data irregularities present in independently processed sets [23]), it is not surprising that in all the studied cases, the extended transformations of train sets resulted in improved performance. The highest increase was detected when relatively few intervals were constructed for the additionally processed variables, 2 or 3, for both variants of the equal frequency binning method. For the Wharton-Johnston dataset, the improvement was observed when train as well as test sets were further transformed, while for the James-Hardy dataset it was best to process train sets as it led to more variables to induce rules from, but subject the test sets only to the first stage of discretisation, that is, just supervised transformations.

5.3. Discussion of Results

The analysis of results also included comparisons between the two variants of unsupervised discretisation methods, equal frequency binning vs. equal frequency binning with weights, and the two types of voting strategies in case of conflicts, standard weighted voting vs. simple majority voting. For each unsupervised discretisation approach, voting strategies were compared, and, on the other hand, for each conflict resolution strategy, the discretisation processes were contrasted. For the corresponding data variants, the differences in the reported classifier performance were calculated, as shown in Fig. 3, where the coloured cells indicate a positive difference.

When in this two-step processing the roles played by the two variants of unsupervised discretisation algorithms were compared (which was denoted as dsF_duf - dsF_dufw), it turned out that for equal frequency binning with weights, the performance of rule-based classifiers was

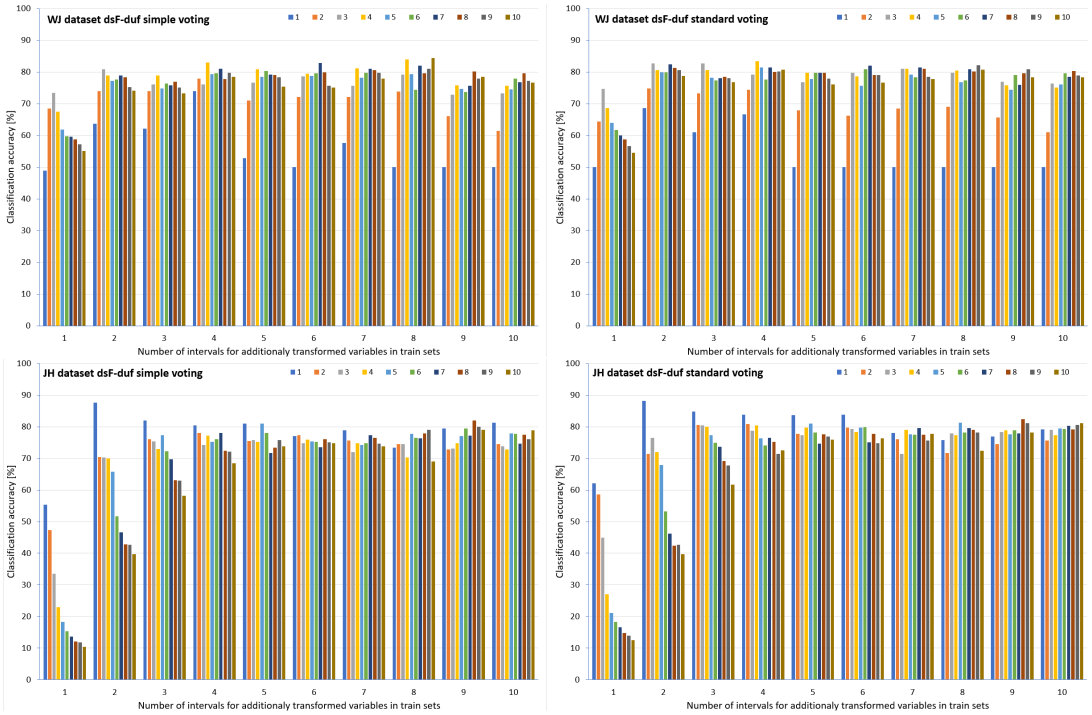


Fig. 1. Performance of rule-based classifiers for data transformed by supervised discretisation with the Fayyad and Irani algorithm combined with unsupervised equal frequency binning (dsF-duf) for the WJ dataset and the JH dataset, for simple and standard voting as a conflict resolution strategy.

higher for the matching bin numbers in train and test sets for both the WJ and JH datasets, although there were also some cases where the simple equal frequency binning reported predictions at much higher level (even 20 percentage points).

On the other hand, comparative analysis of predictions for rule-based classifiers while using two different strategies for conflict resolution (denoted as standard voting - simple voting), brought the conclusion that for both variants of the equal frequency binning algorithm, in the majority of cases standard voting worked better. In particular, for the James-Hardy dataset, it happened for almost all investigated cases, regardless of the variant of the unsupervised discretisation method used and the number of bins constructed for additionally processed attributes.

Furthermore, for the entire range of possible combinations of train and test sets investigated, without and with extended processing dedicated to discretisation, selected statistics were calculated, as shown in Table 3. For the Wharton-Johnston dataset, for both combinations of dsF-duf and dsF-dufw, the minima were the same for both voting types, while other characteristics showed differences. On the other hand, for the James-Hardy dataset, the maxima were the same, and for other elements different values were obtained. For a dataset and an overall characteristic chosen, the differences (if existing) were relatively small in scale between the two variants of the duf method. Comparison of voting strategies led to the observation that for the JH dataset almost all values were higher, with the exception of lower standard deviation. For the WJ dataset, rather surprisingly, the maxima were higher for simple voting, but the averages, minima, and unfortunately also the standard deviation were larger for standard voting.

The results from the experiments, which illustrate the proposed methodology for conditional discretisation, combining supervised and unsupervised methods, show many cases of improved predictions compared to the reference points of standard supervised or unsupervised transformations of the input data. These observations led to the conclusion that supervised discretisation procedures should not always be considered superior. The extensive tests performed validate

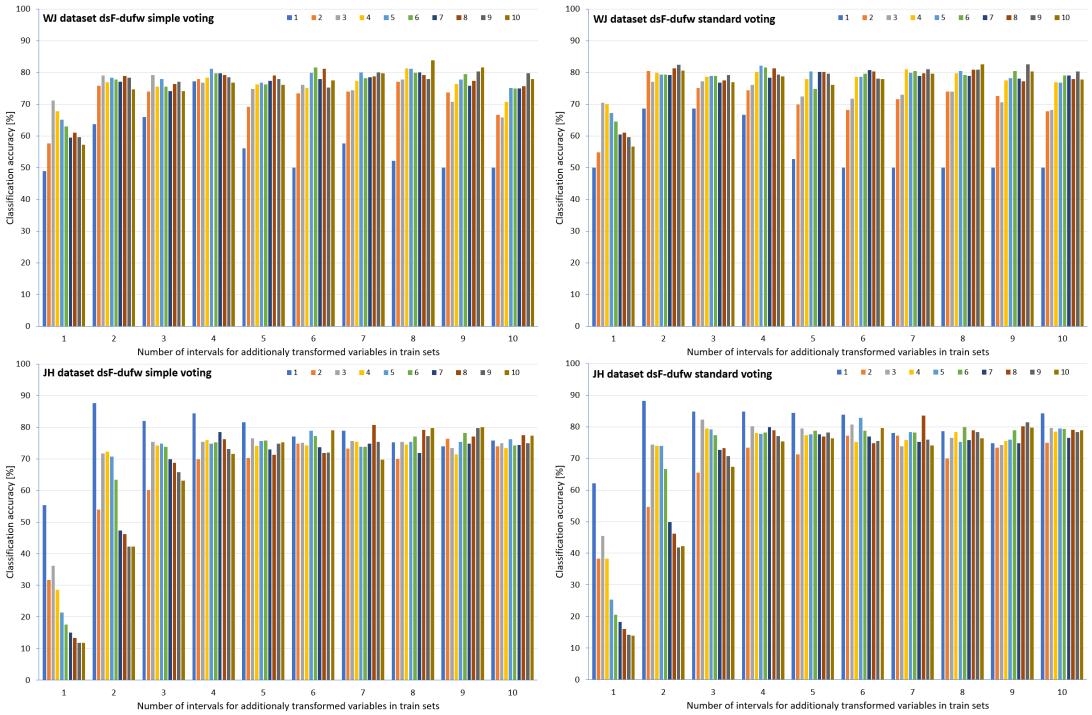


Fig. 2. Performance of rule-based classifiers for data transformed by supervised discretisation with the Fayyad and Irani algorithm combined with unsupervised equal frequency with weights binning (dsF-dufw) for the WJ dataset and the JH dataset, for simple and standard voting as a conflict resolution strategy.

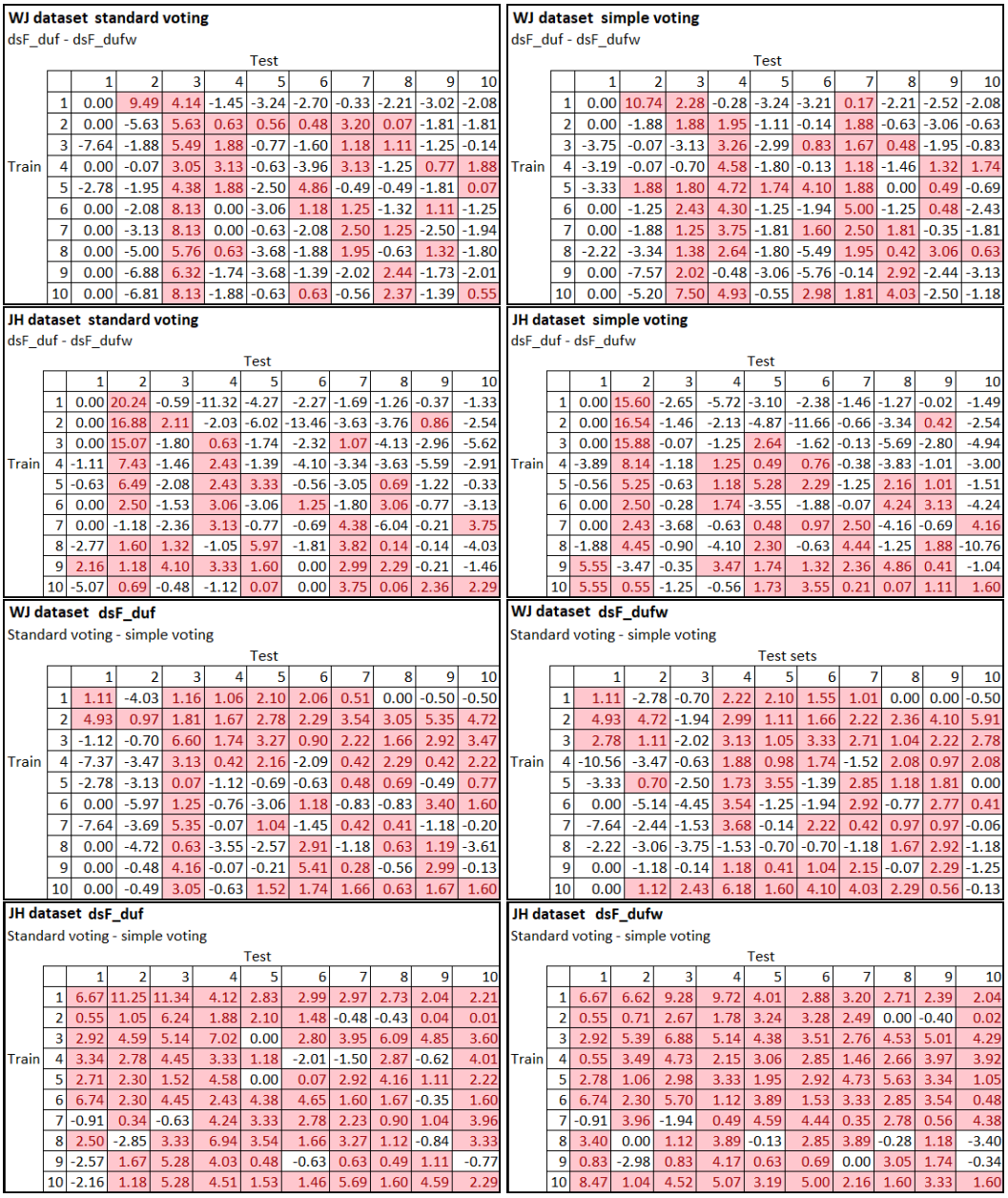
Table 3. Statistics of performance [%] of rule-based classifiers for datasets discretised by the supervised Fayyad and Irani approach combined with unsupervised equal frequency binning (dsF-duf) or equal frequency binning with weights (dsF-dufw), with averaging over the varying number of constructed bins.

	Simple voting			**Standard voting**		
	Avg. $\pm$ St.dev.	Min	Max	Avg.$\pm$St.dev.	Min	Max
	dsF-duf					
WJ dataset	73.83 $\pm$ 08.59	48.89	84.38	74.30 $\pm$ 09.18	50.00	83.34
JH dataset	68.48 $\pm$ 17.48	10.38	87.64	70.95 $\pm$ 17.19	12.58	88.20
	dsF-dufw					
WJ dataset	73.77 $\pm$ 08.12	48.89	83.75	74.35 $\pm$ 08.75	50.00	82.57
JH dataset	68.24 $\pm$ 16.97	11.87	87.64	70.98 $\pm$ 16.64	13.91	88.20

the notion that the extended processing of some variables helps to preserve higher informative content than straightforward one-step transformations, and enables to combine advantages of both transformation paths.

6. Conclusions

The paper presents an illustration for the proposed extension of transformations of the input data in the cases when the change from continuous to discrete domain is required to enable or simplify data mining. Some application domains bring their characteristics and limitations into the knowledge discovery phase. In the stylometric analysis of texts that was studied, style-markers, which are used to define authorial profiles, often are continuous valued, as they reflect lexical and syntactic patterns observed in text samples. The input space is stratified as a consequence



WJ dataset standard voting
dsF_duf - dsF_dufw

Train \ Test	1	2	3	4	5	6	7	8	9	10
1	0.00	9.49	4.14	-1.45	-3.24	-2.70	-0.33	-2.21	-3.02	-2.08
2	0.00	-5.63	5.63	0.63	0.56	0.48	3.20	0.07	-1.81	-1.81
3	-7.64	-1.88	5.49	1.88	-0.77	-1.60	1.18	1.11	-1.25	-0.14
4	0.00	-0.07	3.05	3.13	-0.63	-3.96	3.13	-1.25	0.77	1.88
5	-2.78	-1.95	4.38	1.88	-2.50	4.86	-0.49	-0.49	-1.81	0.07
6	0.00	-2.08	8.13	0.00	-3.06	1.18	1.25	-1.32	1.11	-1.25
7	0.00	-3.13	8.13	0.00	-0.63	-2.08	2.50	1.25	-2.50	-1.94
8	0.00	-5.00	5.76	0.63	-3.68	-1.88	1.95	-0.63	1.32	-1.80
9	0.00	-6.88	6.32	-1.74	-3.68	-1.39	-2.02	2.44	-1.73	-2.01
10	0.00	-6.81	8.13	-1.88	-0.63	0.63	-0.56	2.37	-1.39	0.55

WJ dataset simple voting
dsF_duf - dsF_dufw

Train \ Test	1	2	3	4	5	6	7	8	9	10
1	0.00	10.74	2.28	-0.28	-3.24	-3.21	0.17	-2.21	-2.52	-2.08
2	0.00	-1.88	1.88	1.95	-1.11	-0.14	1.88	-0.63	-3.06	-0.63
3	-3.75	-0.07	-3.13	3.26	-2.99	0.83	1.67	0.48	-1.95	-0.83
4	-3.19	-0.07	-0.70	4.58	-1.80	-0.13	1.18	-1.46	1.32	1.74
5	-3.33	1.88	1.80	4.72	1.74	4.10	1.88	0.00	0.49	-0.69
6	0.00	-1.25	2.43	4.30	-1.25	-1.94	5.00	-1.25	0.48	-2.43
7	0.00	-1.88	1.25	3.75	-1.81	1.60	2.50	1.81	-0.35	-1.81
8	-2.22	-3.34	1.38	2.64	-1.80	-5.49	1.95	0.42	3.06	0.63
9	0.00	-7.57	2.02	-0.48	-3.06	-5.76	-0.14	2.92	-2.44	-3.13
10	0.00	-5.20	7.50	4.93	-0.55	2.98	1.81	4.03	-2.50	-1.18

JH dataset standard voting
dsF_duf - dsF_dufw

Train \ Test	1	2	3	4	5	6	7	8	9	10
1	0.00	20.24	-0.59	-11.32	-4.27	-2.27	-1.69	-1.26	-0.37	-1.33
2	0.00	16.88	2.11	-2.03	-6.02	-13.46	-3.63	-3.76	0.86	-2.54
3	0.00	15.07	-1.80	0.63	-1.74	-2.32	1.07	-4.13	-2.96	-5.62
4	-1.11	7.43	-1.46	2.43	-1.39	-4.10	-3.34	-3.63	-5.59	-2.91
5	-0.63	6.49	-2.08	2.43	3.33	-0.56	-3.05	0.69	-1.22	-0.33
6	0.00	2.50	-1.53	3.06	-3.06	1.25	-1.80	3.06	-0.77	-3.13
7	0.00	-1.18	-2.36	3.13	-0.77	-0.69	4.38	-6.04	-0.21	3.75
8	-2.77	1.60	1.32	-1.05	5.97	-1.81	3.82	0.14	-0.14	-4.03
9	2.16	1.18	4.10	3.33	1.60	0.00	2.99	2.29	-0.21	-1.46
10	-5.07	0.69	-0.48	-1.12	0.07	0.00	3.75	0.06	2.36	2.29

JH dataset simple voting
dsF_duf - dsF_dufw

Train \ Test	1	2	3	4	5	6	7	8	9	10
1	0.00	15.60	-2.65	-5.72	-3.10	-2.38	-1.46	-1.27	-0.02	-1.49
2	0.00	16.54	-1.46	-2.13	-4.87	-11.66	-0.66	-3.34	0.42	-2.54
3	0.00	15.88	-0.07	-1.25	2.64	-1.62	-0.13	-5.69	-2.80	-4.94
4	-3.89	8.14	-1.18	1.25	0.49	0.76	-0.38	-3.83	-1.01	-3.00
5	-0.56	5.25	-0.63	1.18	5.28	2.29	-1.25	2.16	1.01	-1.51
6	0.00	2.50	-0.28	1.74	-3.55	-1.88	-0.07	4.24	3.13	-4.24
7	0.00	2.43	-3.68	-0.63	0.48	0.97	2.50	-4.16	-0.69	4.16
8	-1.88	4.45	-0.90	-4.10	2.30	-0.63	4.44	-1.25	1.88	-10.76
9	5.55	-3.47	-0.35	3.47	1.74	1.32	2.36	4.86	0.41	-1.04
10	5.55	0.55	-1.25	-0.56	1.73	3.55	0.21	0.07	1.11	1.60

WJ dataset dsF_duf
Standard voting - simple voting

Train \ Test	1	2	3	4	5	6	7	8	9	10
1	1.11	-4.03	1.16	1.06	2.10	2.06	0.51	0.00	-0.50	-0.50
2	4.93	0.97	1.81	1.67	2.78	2.29	3.54	3.05	5.35	4.72
3	-1.12	-0.70	6.60	1.74	3.27	0.90	2.22	1.66	2.92	3.47
4	-7.37	-3.47	3.13	0.42	2.16	-2.09	0.42	2.29	0.42	2.22
5	-2.78	-3.13	0.07	-1.12	-0.69	-0.63	0.48	0.69	-0.49	0.77
6	0.00	-5.97	1.25	-0.76	-3.06	1.18	-0.83	-0.83	3.40	1.60
7	-7.64	-3.69	5.35	-0.07	1.04	-1.45	0.42	0.41	-1.18	-0.20
8	0.00	-4.72	0.63	-3.55	-2.57	2.91	-1.18	0.63	1.19	-3.61
9	0.00	-0.48	4.16	-0.07	-0.21	5.41	0.28	-0.56	2.99	-0.13
10	0.00	-0.49	3.05	-0.63	1.52	1.74	1.66	0.63	1.67	1.60

WJ dataset dsF_dufw
Standard voting - simple voting

Train \ Test sets	1	2	3	4	5	6	7	8	9	10
1	1.11	-2.78	-0.70	2.22	2.10	1.55	1.01	0.00	0.00	-0.50
2	4.93	4.72	-1.94	2.99	1.11	1.66	2.22	2.36	4.10	5.91
3	2.78	1.11	-2.02	3.13	1.05	3.33	2.71	1.04	2.22	2.78
4	-10.56	-3.47	-0.63	1.88	0.98	1.74	-1.52	2.08	0.97	2.08
5	-3.33	0.70	-2.50	1.73	3.55	-1.39	2.85	1.18	1.81	0.00
6	0.00	-5.14	-4.45	3.54	-1.25	-1.94	2.92	-0.77	2.77	0.41
7	-7.64	-2.44	-1.53	3.68	-0.14	2.22	0.42	0.97	0.97	-0.06
8	-2.22	-3.06	-3.75	-1.53	-0.70	-0.70	-1.18	1.67	2.92	-1.18
9	0.00	-1.18	-0.14	1.18	0.41	1.04	2.15	-0.07	2.29	-1.25
10	0.00	1.12	2.43	6.18	1.60	4.10	4.03	2.29	0.56	-0.13

JH dataset dsF_duf
Standard voting - simple voting

Train \ Test	1	2	3	4	5	6	7	8	9	10
1	6.67	11.25	11.34	4.12	2.83	2.99	2.97	2.73	2.04	2.21
2	0.55	1.05	6.24	1.88	2.10	1.48	-0.48	-0.43	0.04	0.01
3	2.92	4.59	5.14	7.02	0.00	2.80	3.95	6.09	4.85	3.60
4	3.34	2.78	4.45	3.33	1.18	-2.01	-1.50	2.87	-0.62	4.01
5	2.71	2.30	1.52	4.58	0.00	0.07	2.92	4.16	1.11	2.22
6	6.74	2.30	4.45	2.43	4.38	4.65	1.60	1.67	-0.35	1.60
7	-0.91	0.34	-0.63	4.24	3.33	2.78	2.23	0.90	1.04	3.96
8	2.50	-2.85	3.33	6.94	3.54	1.66	3.27	1.12	-0.84	3.33
9	-2.57	1.67	5.28	4.03	0.48	-0.63	0.63	0.49	1.11	-0.77
10	-2.16	1.18	5.28	4.51	1.53	1.46	5.69	1.60	4.59	2.29

JH dataset dsF_dufw
Standard voting - simple voting

Train \ Test	1	2	3	4	5	6	7	8	9	10
1	6.67	6.62	9.28	9.72	4.01	2.88	3.20	2.71	2.39	2.04
2	0.55	0.71	2.67	1.78	3.24	3.28	2.49	0.00	-0.40	0.02
3	2.92	5.39	6.88	5.14	4.38	3.51	2.76	4.53	5.01	4.29
4	0.55	3.49	4.73	2.15	3.06	2.85	1.46	2.66	3.97	3.92
5	2.78	1.06	2.98	3.33	1.95	2.92	4.73	5.63	3.34	1.05
6	6.74	2.30	5.70	1.12	3.89	1.53	3.33	2.85	3.54	0.48
7	-0.91	3.96	-1.94	0.49	4.59	4.44	0.35	2.78	0.56	4.38
8	3.40	0.00	1.12	3.89	-0.13	2.85	3.89	-0.28	1.18	-3.40
9	0.83	-2.98	0.83	4.17	0.63	0.69	0.00	3.05	1.74	-0.34
10	8.47	1.04	4.52	5.07	3.19	5.00	2.16	1.60	3.33	1.60

Fig. 3. Differences in performance of rule-based classifiers for data transformed by supervised discretisation with the Fayyad and Irani algorithm combined with unsupervised equal frequency binning (dsF-duf), and equal frequency binning with weights (dsF-dufw) for the WJ dataset and the JH dataset, for simple and standard voting as a conflict resolution strategy.

of the sample construction process, and the existence of sub-classes causes the need for separate test sets to be employed for reliable performance estimation of a classifier exploring train sets.

To infer meaningful and interpretable descriptions of writing styles, such as can be given by decision rules induced in the rough set approach, a translation into categorical representation is needed first. However, simple processing by supervised and unsupervised approaches is not necessarily satisfactory. Supervised discretisation can lead to discarding too much information and problems in transformations for multiple independent sets. Unsupervised methods disregard the information on class labels for samples, and thus cannot be considered as supporting predictions. Combining supervised with unsupervised approaches brings advantages of both processing paths into equation, and in the extensive investigations reported, such additional

transformations resulted in many cases of noticeably improved performance, which confirmed the merits of the proposed approach.

The results obtained were based on specific input data characteristics, selected discretisation methods, and chosen classifiers. In the future works, wider studies of two-step conditional discretisation will be conducted using other types of classifiers, in particular with different mathematical backgrounds, other discretisation methods, and more varied application domains.

Acknowledgements

The research works presented in the paper were carried out within the statutory project of the Department of Graphics, Computer Vision and Digital Systems (RAU-6, 2023), at the Silesian University of Technology, Gliwice, Poland, and at the Institute of Computer Science, University of Silesia in Katowice, Sosnowiec, Poland.

References

1. Amin, T., Chikalov, I., Moshkov, M., Zielosko, B.: Dynamic programming approach for partial decision rule optimization. Fundam. Informaticae **119**(3-4), 233–248 (2012)
2. Argamon, S., Burns, K., Dubnov, S. (eds.): The structure of style: Algorithmic approaches to understanding manner and meaning. Springer, Berlin (2010)
3. Baron, G., Stańczyk, U.: Standard vs. non-standard cross-validation: evaluation of performance in a space with structured distribution of datapoints. In: Wątróbski, J., Salabun, W., Toro, C., Zanni-Merk, C., Howlett, R.J., Jain, L.C. (eds.) Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES-2021, Szczecin, Poland, 8-10 September 2021, Procedia Computer Science, vol. 192, pp. 1245–1254. Elsevier (2021)
4. Bazan, J., Szczuka, M.: The rough set exploration system. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets III, Lecture Notes in Computer Science, vol. 3400, pp. 37–56. Springer, Berlin, Heidelberg (2005)
5. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and unsupervised discretization of continuous features. In: Machine Learning Proceedings 1995: Proceedings of the 12th International Conference on Machine Learning. pp. 194–202. Elsevier (1995)
6. Eder, M.: Style-markers in authorship attribution a cross-language study of the authorial fingerprint. Studies in Polish Linguistics **6**(1), 99–114 (2011)
7. Eder, M.: Mind your corpus: systematic errors in authorship attribution. Literary and Linguistic Computing **28**, 603–614 (12 2013)
8. Fayyad, U., Irani, K.: Multi-interval discretization of continuous valued attributes for classification learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence. vol. 2, pp. 1022–1027. Morgan Kaufmann Publishers (1993)
9. Franzini, G., Kestemont, M., Rotari, G., Jander, M., Ochab, J., Franzini, E., Byszuk, J., Rybicki, J.: Attributing authorship in the noisy digitized correspondence of Jacob and Wilhelm Grimm. Frontiers in Digital Humanities **5**, 4 (2018)
10. Garcia, S., Luengo, J., Saez, J., Lopez, V., Herrera, F.: A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering **25**(4), 734–750 (2013)
11. Grünwald, P.D.: The Minimum Description Length Principle (Adaptive Computation and Machine Learning). The MIT Press (2007)
12. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2011)
13. Huan, L., Farhad, H., Lim, T., Manoranjan, D.: Discretization: An enabling technique. Data Mining and Knowledge Discovery **6**(4), 393–423 (2002)

14. Jockers, M., Witten, D.: A comparative study of machine learning methods for authorship attribution. Literary and Linguistic Computing **25**(2), 215–223 (2010)
15. Lindgren, T.: Methods for rule conflict resolution. In: Boulicaut, J., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) Machine Learning: ECML 2004, Lecture Notes in Computer Sicence, vol. 3201, pp. 262–273. Springer, Berlin Heidelberg (2004)
16. Liu, H., Cocea, M.: Induction of classification rules by gini-index based rule generation. Information Sciences **436-437**, 227–246 (2018)
17. Moshkov, M.J., Piliszczuk, M., Zielosko, B.: On construction of partial reducts and irreducible partial decision rules. Fundamenta Informaticae **75**(1-4), 357–374 (2007)
18. Pawlak, Z.: Rough sets and intelligent data analysis. Information Sciences **147**, 1–12 (2002)
19. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences **177**(1), 3–27 (2007)
20. Sikora, M., Matyszok, P., Wróbel, L.: SCARI: separate and conquer algorithm for action rules and recommendations induction. Inf. Sci. **607**, 849–868 (2022)
21. Stańczyk, U.: On unsupervised and supervised discretisation in mining stylometric features. In: Gruca, A., Czachórski, T., Deorowicz, S., Har ężlak, K., Piotrowska, A. (eds.) Man-Machine Interactions 6. ICMMI 2019, Advances in Intelligent Systems and Computing, vol. 1061, pp. 156–166. Springer, Cham (2020)
22. Stańczyk, U., Zielosko, B.: On approaches to discretisation of stylometric data and conflict resolution in decision making. In: Rudas, I.J., Csirik, J., Toro, C., Botzheim, J., Howlett, R.J., Jain, L.C. (eds.) Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES-2019, Budapest, Hungary, 4-6 September 2019, Procedia Computer Science, vol. 159, pp. 1811–1820. Elsevier (2019)
23. Stańczyk, U., Zielosko, B.: Data irregularities in discretisation of test sets used for evaluation of classification systems: A case study on authorship attribution. Bulletin of the Polish Academy of Sciences: Technical Sciences **69**(4), 1–12 (2021)
24. Stańczyk, U., Zielosko, B., Baron, G.: Discretisation of conditions in decision rules induced for continuous data. PLoS ONE **15**(4), 1–33 (2020)
25. Stąpor, K.: Evaluation of classifiers: current methods and future research directions. In: Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS). ACSIS, vol. 13, pp. 37–40 (2017)
26. Stąpor, K., Ksieniewicz, P., García, S., Woźniak, M.: How to design the fair experimental classifier evaluation. Applied Soft Computing **104**, 107219 (2021)
27. Stepaniuk, J., Skowron, A.: Three-way approximation of decision granules based on the rough set approach. International Journal of Approximate Reasoning **155**, 1–16 (2023)
28. Weidman, S.G., O'Sullivan, J.: The limits of distinctive words: Re-evaluating literature's gender marker debate. Digital Scholarship in the Humanities **33**, 374–390 (2018)
29. Wróbel, L., Sikora, M., Michalak, M.: Rule quality measures settings in classification, regression and survival rule induction — an empirical approach. Fundamenta Informaticae **149**, 419–449 (2016)
30. Wu, H., Zhang, Z., Wu, Q.: Exploring syntactic and semantic features for authorship attribution. Applied Soft Computing **111**, 107815 (2021)
31. Zielosko, B., Żabiński, K.: Selected approaches for decision rules construction-comparative study. In: Wątróbski, J., Salabun, W., Toro, C., Zanni-Merk, C., Howlett, R.J., Jain, L.C. (eds.) Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES-2021, Szczecin, Poland, 8-10 September 2021, Procedia Computer Science, vol. 192, pp. 3667–3676. Elsevier (2021)