

Article

Significance of Single-Interval Discrete Attributes: Case Study on Two-Level Discretisation

Urszula Stańczyk ^{1,*} , Beata Zielosko ²  and Grzegorz Baron ¹ 

¹ Department of Computer Graphics, Vision and Digital Systems, Silesian University of Technology, Akademicka 2A, 44-100 Gliwice, Poland; grzegorz.baron@polsl.pl

² Institute of Computer Science, University of Silesia in Katowice, Będzińska 39, 41-200 Sosnowiec, Poland; beata.zielosko@us.edu.pl

* Correspondence: urszula.stanczyk@polsl.pl

Abstract: Supervised discretisation is widely considered as far more advantageous than unsupervised transformation of attributes, because it helps to preserve the informative content of a variable, which is useful in classification. After discretisation, based on employed criteria, some attributes can be found irrelevant, and all their values can be represented in a discrete domain by a single interval. In consequence, such attributes are removed from considerations, and no knowledge is mined from them. The paper presents research focused on extended transformations of attribute values, thus combining supervised with unsupervised discretisation strategies. For all variables with single intervals returned from supervised algorithms, the ranges of values were transformed by unsupervised methods with varying numbers of bins. Resulting variants of the data were subjected to selected data mining techniques, and the performance of a group of classifiers was evaluated and compared. The experiments were performed on a stylometric task of authorship attribution.

Keywords: discretisation; pattern recognition; stylometry



Citation: Stańczyk, U.; Zielosko, B.; Baron, G. Significance of Single-Interval Discrete Attributes: Case Study on Two-Level Discretisation. *Appl. Sci.* **2024**, *14*, 4088. <https://doi.org/10.3390/app14104088>

Academic Editor: Yang Kuang

Received: 13 March 2024

Revised: 28 April 2024

Accepted: 9 May 2024

Published: 11 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Discretisation is a procedure that is typically employed as a step in initial preprocessing of data to transform continuous features into categorical ones. If it is performed, in standard approaches, all variables are treated in the same way [1]. Discretisation results in specific granulation of the input domain, as constructed intervals (also called bins) represent ranges of attributes' values. This change enables the application of such data mining methods, which can operate only on nominal features [2]. After the transformations, the data is simplified, and the descriptive properties and powers of attributes become more general in their nature. It can mean that noise or insignificant details are removed from the input domain, but it also leads to some loss of information, even if it is considered as unimportant at this stage. Subjecting discretised data to exploration implies the application of a knowledge discovery process to limited, reduced data.

For discretisation to be effective and useful, this transformation should preserve sufficient information on the values of features and their distributions. To that end, two interconnected elements play the key role: the number of bins defined and the cut points indicating borders where one interval ends and another begins. The values in an attribute domain can be analysed just by themselves in unsupervised processing or along with labels assigning them to particular classes in supervised transformations [3]. The former approach is popularly considered as too simplistic, and supervised discretisation is widely preferred. The algorithms in the latter category rely on the evaluation of the candidate cut points with some measure and choosing the best one based on some selected criteria [4].

The Fayyad and Irani [5] and Kononenko [6] methods are two well-known supervised discretisation procedures referring to the entropy and the Minimum Description Length (MDL) principle in the construction of bins [4]. Both algorithms start with assigning a

single interval to represent an entire domain for a translated variable, and they then check whether splitting this interval is beneficial. If not, then the defined bin remains the sole representative for all values. As a consequence, the variable has constant nominal value for all samples, thus making it effectively removed from any knowledge discovery process that follows discretisation.

Instead of just accepting the verdict that some attributes are judged as useless by supervised discretisation, these variables can be subjected to more complex transformations. The paper reports research works dedicated to not one- but two-level discretisation of features. In the first processing step, the features are discretised through supervised methods. In the second step, for all these attributes which are assigned single intervals, their continuous domains are transformed by unsupervised algorithms. The combination of both approaches enables obtaining meaningful discrete representations for all variables.

The experiments were executed within stylometry, the domain focused on characteristics of writing styles [7], where the task of authorship attribution is considered of paramount importance [8]. In the training phase for known authors, stylistic profiles are constructed, which are then measured against samples of writing of unknown or questioned authorship to either confirm or deny it. Profiles are obtained by referring to linguistic descriptors [9], which are most often continuous [10]. Building a profile is not trivial, as authors should be recognised regardless of a particular topic they write about. Knowledge discovery means mining stylometric features and removing any of them from considerations as a consequence of discretisation could be ill afforded. Therefore, any transformation that aims at preserving any available informative content of input features, however small or insignificant it may seem, is welcome, as it may help in recognition.

Treating authorship attribution as a supervised learning problem [11], several inducers were applied to the stylometric data to study the influence of discretisation transformations on pattern recognition. The classifiers could operate on both continuous and nominal valued features [12]. For all the classification systems, their performance was observed in the context of discretisation of the input space and the characteristics of such processing. Taking the results from supervised discretisation at face value implied that not only details of attributes' values but also entire domains can end up being disregarded. The experiments showed that it could mean ignoring some cases. In the proposed approach, where an extra step of transformation is performed, it led to discrete transformations that were advantageous to classification.

The original contributions of this paper include the following:

- The proposed framework for two-level discretisation, where all features are explored within a knowledge discovery process, thus combining supervised discretisation algorithms with unsupervised discretisation methods;
- Extensive experiments performed on datasets from the stylometry domain involving various parameters of operation: thereby involving, on one hand, inducers with different operation modes and multiple variants of data transformations on the other hand;
- Analysis of performance for selected classification systems employed to variants of input data discretised in many ways, with a specific focus on the influence of extended transformations for samples in the train and test sets.

The body of the paper is organised into seven sections. The introduction is followed by a description of discretisation approaches, the differences between them, and parameters, which is given in Section 2. The characteristics of input data space, dedicated to stylometric analysis of texts, are explained in Section 3. Section 4 contains presentation of the proposed research framework and its stages and steps. Details of the experiments performed are included in Section 5, with comments with respect to the results listed in Section 6. Conclusions and indications of possible future research directions are given in Section 7.

2. Discretisation

The knowledge discovery process plays a vital role in a wide variety of applications. One of the key steps in this process is the preparation of the data, which in turn influences the outcome of the subsequent phases of data exploration [13]. Preparation of the data includes various preprocessing methods [14]. Data cleaning covers filling missing values, smoothing noise, identifying or removing outliers, and resolving inconsistencies. Data integration and transformation involve the integration of elements that come from various sources and making them uniform in format. In addition, data transformation includes normalisation and aggregation, which are important for data mining processes. Data reduction makes it possible to obtain a reduced representation of a dataset, which is smaller in volume but produces the same analytical results. In this latter group, discretisation, attribute subset selection, and dimensionality reduction methods can be distinguished. Detailed information on discretisation approaches is presented in the following.

2.1. Translating Continuous Space into Discrete

Discretisation can be considered as a form of data reduction that is useful for the automatic generation of concept hierarchies from numerical data [12]. The main objective of a discretisation process is to reduce the number of values for a given continuous attribute by dividing the range of values into intervals [15]. Each bin is associated with a distinct discrete value. In this way, the attributes' values can be treated as nominal ones. This kind of preprocessing method is important, because there are algorithms that do not work with continuous features and require the discretised form of the attributes' values [16].

Among other advantages, the following can be distinguished: lower memory requirements for storage, increased comprehensibility of the description of the presented knowledge [17]—which is more relevant for interpretation—the possible reduction of noise present in the data, and enhanced efficiency of induction algorithms that work on a reduced domain of attribute values, which also influences time needed for operations. However, it should be noted that the discretisation process can cause some loss of information [18]; therefore, it should be used with certain caution and always adjusted to the nature of available data, its characteristics, and inherent nature [19].

2.2. Unsupervised Discretisation Transformations

In the unsupervised methods, the range of continuous values of a considered attribute is divided into subranges by the user-specified width (range of values) or frequency (number of values in each interval). In this group of algorithms, equal width binning and equal frequency binning are the most popular. Both are relatively simple, but they are sensitive to a user-provided input parameter k , which determines the number of bins to be constructed for a given attribute.

The equal width method divides the range of attribute values into k intervals of equal width. The equal frequency method divides the sorted values of a discretised attribute into k intervals so that each interval contains approximately the same number of values. The weakness of these methods is that some information can be lost after the discretisation process. It can occur in a situation of uneven distribution of the considered attribute values. If outliers of values are present in the data, they can affect the cut points between intervals of the considered attribute. In the case of the equal frequency method, it is important to keep the same values of a discretised attribute in the same interval. Consequently, it is not always possible to generate k equal frequency bins.

2.3. Entropy and MDL Principle in Supervised Discretisation Algorithms

Supervised discretisation algorithms condition the process of finding cut points and constructing intervals based on their usefulness for the distinction of concepts. Candidate cut points are firstly located, each is evaluated, and the best one (optimal) selected. In this group of methods, entropy-based measures are often employed to evaluate cut points.

The class information entropy of the candidate partitions is a measure of purity. It evaluates the amount of information needed to specify to which class an instance belongs. In the top-down approach, at the beginning, one range is considered containing all values of a discretised attribute. This interval is recursively divided into smaller bins until an optimal number of intervals is obtained or the stopping criterion is met. As a stopping criterion, the Minimum Description Length principle can be used, with roots in Bayesian inference and Kolmogorov complexity.

The MDL principle states that among the given set of hypotheses, the one with the minimum combined description lengths of both the hypothesis and the error for a given set of data is the best approximation of the mechanism behind the data and can be used to predict the future data with the best accuracy. The principle was introduced by Rissanen [20] and adopted for the induction of decision trees [21] and rules. This principle can be considered as a method of inductive inference and the basis of statistical modelling and machine learning. It can be used for model selection, prediction, and estimation problems, especially where the models under consideration are complex. The MDL principle is also used as an element of discretisation algorithms.

The Fayyad and Irani [5] and Kononenko [6] algorithms are the most popular supervised discretisation methods. The first one evaluates the midpoint between neighbouring pairs of sorted values of an attribute as a cut point and selects this candidate cut point for which the entropy is minimal. This binary discretisation is applied recursively and refers to the MDL principle as the stopping criterion. In the Kononenko method, the MDL principle is applied in the framework of selecting the most compressive attribute.

2.4. Characterisation of Attributes by Discretisation

One of the goals of subjecting a continuous-valued attribute to a discretisation process is to prepare the input data for a specific data mining algorithm and its operation mode [3]. The research carried out in this area shows that the results of discretisation may also constitute features' characteristics. It can be expressed as the number of bins assigned to the values of a given attribute by a specific discretisation algorithm [22]. On the basis of that, features can be evaluated from the point of view of the possibility of distinguishing decision classes and constructing classifiers, since the obtained bin number reflects the relationship between the attribute values and the class label. It leads to reason that higher numbers of bins indicate a higher importance of some variables. In the case of discretisation approaches based on entropy, it means that with increased numbers of intervals and lower numbers of ranges, the class entropy can be too degraded, or there may exist more complex relationships between these attributes and decision classes [23].

The attributes for which only one interval has been indicated by supervised discretisation often turn out to be the ones occupying the last places in feature rankings, i.e., their values do not affect the distinguishability of objects relative to decision classes. From a classification point of view, such attributes appear to be irrelevant. However, a deeper analysis and additional research on the "1-bin" features show that the application of the next stage of processing for such attributes causes them to no longer be considered irrelevant, and the validity of their use in the classifier construction process exists [24].

When the data are discretised, the division of samples into the training and testing parts should also be considered. When these sets are discretised independently, there is a risk that not only the cut points but the obtained numbers of bins for specific variables in the test and training sets may differ. Local variations in values and their irregularities expressed through discretisation can result in a deterioration of classification quality [25], especially in the case of attributes that by discretisation were assigned one interval. With this approach for separate sets, separate discretisation models for variables are also constructed and become a part of a learning problem, thereby adding a degree of complexity to a task.

3. Nature of Input Stylometric Space

A style is a concept that is intuitively understood and which could even be subconsciously recognised but notoriously difficult to express or define precisely. It could be characterised, but commonly rather with qualifying descriptors, or illustrated with examples but not measured. However, the application of quantitative measures to writing styles is what researchers in the field of stylometry aim to achieve [26]. The stylometric analysis of texts involves profiling authors by their linguistic habits and choices, with measurable characteristics, which can be employed as input features for data exploration methods.

3.1. Stylometric Characteristic Features

The uniqueness of writing styles is the fundamental and underlying notion of stylometry. Thoughts and ideas presented in a written form capture the author's individuality in preferences with respect to applied language features. The elements, which are observable in many text samples irrespective of the subject matter or specific genre, constitute writer prints [27]. Most often, they refer to the lexical aspects of sentence construction, and syntactic annotation also plays a role.

Unusual linguistic choices are more easily detected, so they are also more prone to falsification. Safer and more reliable descriptors of writing styles could be found among common parts of speech, in particular popular function words and regular punctuation marks. Stylometric markers typically refer to frequencies, averages, and distributions, which makes them continuous-valued. The nature of the input domain imposes some limitations on techniques and methods that can be involved in text analysis or requires discretisation to become part of the data preparation procedure [6].

3.2. Authorship Attribution as Supervised Learning Problem

Attributing authorship is considered the most important stylometric task [28]. It involves two other problems: finding authorial invariants—such sets of features that characterise studied authors; and comparing the writing styles of authors—finding similarities and differentiating the features between them. The first task would be impossible to solve without addressing the two others.

Writer profiles are constructed by the analysis of text samples of known authorship. The input data are explored to detect existing patterns, and then the discovered knowledge can be tried against characteristics calculated for works for which authorship is either questioned or unknown. The former part of processing can be seen as standard learning under supervision [29], with classes corresponding to known authors. The second part is then testing, with the aim of predicting labels for previously unknown samples based on the knowledge on styles learnt from the train sets.

3.3. Construction of Datasets

To build a reliable authorial profile, access to a sufficient number of representative samples of writing is needed [30]. These two adjectives, sufficient and representative, indicate the way of processing, but they are also kind of tricky because of their relativity. The selection of texts, their sizes, and the number of samples are crucial for the effectiveness of authorship attribution.

When there are too few samples, when samples are small, or when they are atypical, the inferred profile could be misleading. Texts can vary significantly in size [31]. The longer the text, the more time spent on writing it, and in consequence, some style variations could be observed in smaller parts or units, such as, for example, chapters in a novel. Calculating some statistics over long texts would hide these subtleties and also would result in a rather small number of available samples, as it is practically impossible to write many long works over a lifetime. To address both these issues at the same time, a different approach is widely used: long text samples are partitioned into smaller chunks of some uniform size, which also helps with the aspect of comparability of obtained statistics.

When several data samples originate from one longer text, it results in a stratified input space, with as many groupings of samples as there are base texts. Within each group, samples are more similar to each other than to those of different groups. To avoid falsely optimistic test results, for the evaluation of learnt patterns, samples should never be used based on the same texts that are used for training. Standard crossvalidation, with its foundation of random choice of samples for folds, cannot be trusted to return a reliable estimation of classification accuracy [32]. Instead, nonstandard crossvalidation, with swapping whole groups of samples (instead of individual instances) could be attempted, but it results in highly increased computational costs. A compromise could be obtained through evaluation with not one but multiple test sets and averaging results, and this kind of processing was employed in the described research.

3.4. Data Mining Techniques Applied

With the fundamental ambiguity of a concept of style and the multitude of sets of linguistic features that can be employed in tasks, stylometry is the domain that is well suited to the application of such techniques of data exploration that are effective for uncertain and incomplete data. Also, algorithms for attribute selection and reduction are welcome, as they can offer their support and help with the aspects of dimensionality [33].

The approaches typically refer to statistics or artificial intelligence [34]. In the former case, there could be constructed matrices reflecting probabilities of transitions between certain characters based on a text or texts, which means building a certain limited language model. This model is then compared with its equivalent calculated for other texts to verify how close or distant they are. On the other hand, an artificial neural network can be trained based on textual data, or decision rules could be inferred [28]. This last form of representation for discovered knowledge has the huge advantage of being easily accessible, transparent, and having intuitive interpretation.

4. Proposed Research Framework

The framework for the proposed experiments can be seen as consisting of four main parts: data preparation, data transformations, data exploration, and results analysis. This section provides the general comments to processing steps, parameters, and limitations for all stages, while the obtained results and their discussion are given in the next two sections.

4.1. Stages of the Proposed Methodology

The framework for the proposed experiments is shown in Figure 1. The four main constituent parts are marked accordingly in the given diagram. Input data preparation includes the selection of authors for attribution, their specific works, preparation of the text samples, choosing a set of features to be employed for the authorship attribution task, and obtaining their values. This stage returns the constructed input datasets with all attributes of numeric type.

After this initial part, the processing of the data follows. It encompasses the application of discretisation procedures by selected algorithms, which results in receiving several discrete variants of the input data. The fusion of information from the supervised and unsupervised discretisation approaches constitutes the central part of the research procedure, and it is highlighted in the diagram.

Data mining or exploration involves knowledge discovery with the help of the chosen inducers. In the final stage of the research, the performance of the selected classifiers is evaluated in various settings, and an analysis of the results is performed. It is focused on the observations of detected trends in performance, particularly in the context of the influence of the discretisation strategy on the quality of predictions for a classifier.

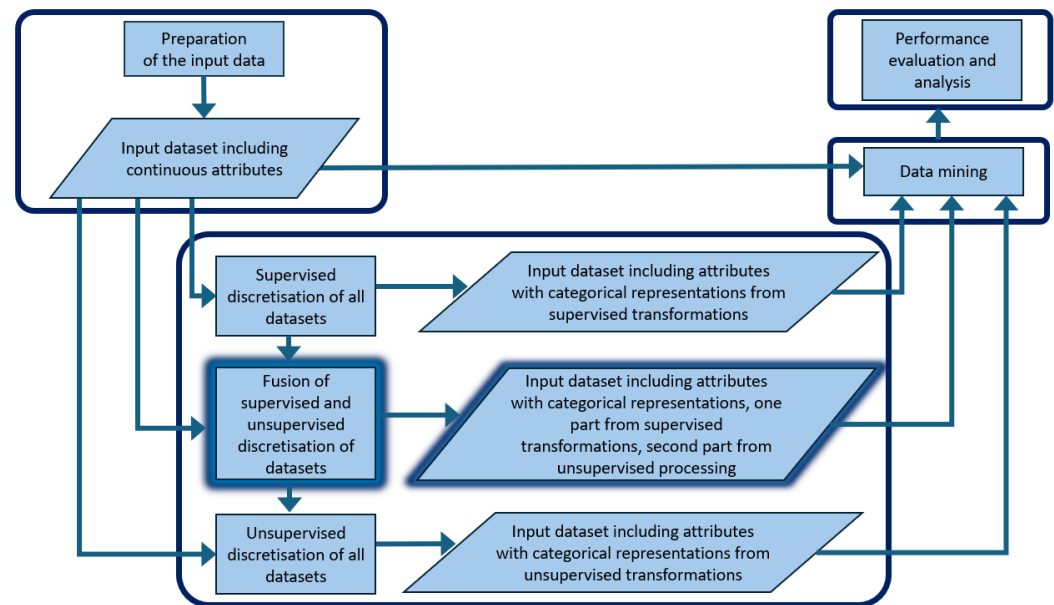


Figure 1. The procedure of the research.

4.2. Data Preparation

In the reported research, the texts to be analysed were taken from the selected literary works of four well-known writers: Edith Wharton, Mary Johnston, Henry James, and Thomas Hardy. All of these writers authored several novels that were long enough that they could provide a base for the construction of text samples. It was necessary to ensure access to sufficiently rich sources of varied information on writing styles that were visible in works with different content and topics [35], both in the training and in the testing stages. Therefore, the novels were partitioned into three separate groups: four for the learning and three and four for two test sets, respectively.

Studies show that the writing styles of male and female authors have some distinctive and common traits [36], which could obscure the observations dedicated to other aspects of the data. This is the reason the four writers were put in pairs according to their gender. This resulted in the construction of two datasets with female writers (F-writers) and male writers (M-writers). As a consequence, the classification task became binary. In each pair, both authors were considered of the same importance, with the same costs of misclassification. To further limit the number of influential factors, the data was prepared to obtain a balance of representation, in particular a balance of classes [37,38].

The novels analysed were divided into smaller parts, which were comparable in size. For the three sets included in a dataset, the following approach was adapted with respect to the number of samples and novels per author:

- Training: A total of four novels and 25 samples per novel, thus leading to 100 samples per author;
- Test 1: A total of three novels and 15 samples per novel, which returned 45 samples per writer;
- Test 2: A total of four novels and 10 samples per novel, thus totaling of 40 samples per writer.

For the selected 12 stylometric descriptors over the text chunks, the frequency of occurrence was calculated. The markers were of the lexical type, because they referred to popular two-letter function words: as, at, by, if, in, no, of, on, or, so, to, up.

4.3. Data Processing

The initial stage of experiments dedicated to data preparation returned two datasets (the female writers and male writers), with each including a single training set and two test

sets with continuous-valued features. These sets were next transformed through standard and nonstandard discretisation procedures.

Within standard processing, the following procedures were involved:

- Unsupervised equal width (duw) binning, with the input parameter that corresponds to the number of bins to be constructed for transformed variables varying from 2 to 10, with a step of one (9 variants of the data);
- Unsupervised equal frequency (duf) binning, with the input parameter that corresponds to the number of bins to be constructed for transformed variables varying from 2 to 10, with a step of one (9 variants of the data);
- Supervised discretisation using the Fayyad and Irani algorithm (dsF), nonparametric (single variants of the data),
- Supervised discretisation using the Kononenko method (dsK), nonparametric (single variants of the data).

This processing gave a total of 20 variants for the transformed datasets.

For unsupervised approaches, the ranges of bin numbers were based on past research [25]. It was shown that further increasing the number of intervals either worsened the accuracy or resulted in negligible differences in representation of information, which led to effectively the same powers of employed inducers. Unsupervised methods consider only attribute values, thus disregarding class labels for samples, but it does not follow that there is no influence on the classification results after transformations. When knowledge is inferred from discretised data, it could be observed that an attribute can play a different role, it could be more or less important depending on the chosen discretisation strategy [39].

Supervised methods find the numbers of intervals to be constructed by themselves based on some evaluation criteria for the considered cut points. The approaches attempt to estimate the usefulness of features for classification and preserve this usefulness when the domain is translated from continuous to discrete. For the attributes considered in the research, the numbers of bins obtained from the train sets for both the female and male writer datasets are provided in Table 1.

Table 1. Characteristics of attributes for supervised discretisation of the train sets.

Dataset	Bins	Fayyad and Irani (dsF) Attributes	Kononenko (dsK) Attributes
F-writers	1	if in no or so up	if in or no so up
	2	as at by	at as by of
	3	of on to	on to
M-writers	1	as of no on so to up	as no on so to up
	2	at if or	at if of or
	3	by in	by in

The attributes that were assigned single intervals to represent the entire range of their values were the ones which were found as not contributing to recognition. Keeping them in this categorical form means excluding them from considerations in a discrete domain because they have constant values in all studied samples. It would mean a reduction of at least 50% of the features (at least 6 of the 12 attributes were assigned single bins for both the supervised discretisation methods and both datasets). This could be treated as an advantage—dimensionality reduction is most often regarded in positive terms. However, what if it were possible to gain more from these variables to be rejected? Finding the answer to this question was the motivation leading to the proposed nonstandard discretisation approaches.

In the research, a two-level discretisation procedure was proposed for data transformation. In the first step, the data were discretised by supervised algorithms, and then at the second stage, the variables, for which only single intervals were formed, were further

transformed by unsupervised methods, with varying the numbers of constructed bins. This two-level discretisation combined supervised and unsupervised approaches as follows:

- Fayyad and Irani approach combined with
 - equal width binning (denoted as dsF-duwi), with bins in the range of $i = 2, \dots, 10$;
 - equal frequency binning (denoted as dsF-dufi), with bins in the range of $i = 2, \dots, 10$;
- Kononenko algorithm combined with
 - equal width binning (denoted as dsK-duwi), with bins in the range of $i = 2, \dots, 10$;
 - equal frequency binning (denoted as dsK-dufi), with bins in the range of $i = 2, \dots, 10$.

For each combination of methods, varying a number of intervals gave nine variants of the data. The total number of such variants was equal to $2 \times (2 \times 9) = 36$ per dataset. Together with twenty variants from standard discretisation (one-level and uniform for all attributes), it gave 56 versions. Each was subjected to data exploration using the selected methods (presented in the next subsection), and the influence of discretisation algorithms on classifier performance was studied.

4.4. Inducers Employed

To provide a wider range for observations in the research, several classifiers were used with different modes of operation. All inducers were capable of operating on both continuous and discrete attributes. They included Naive Bayes (NB), Bayes Network (BNet), J48, k-Nearest Neighbours (kNN), and Random Forest (RndF), which were all implemented in the popular WEKA environment and used with default settings [40].

Naive Bayes and the Bayes Network are known as probabilistic classification models, because they use the Bayes' probability theory [12] to present the relationships between attributes and class labels. NB classifiers are based on the assumption of the conditional independence of variables. In the case of BNet, this assumption is considered in a more flexible way, so it makes it possible to capture more general representations of conditional independence among attributes by using simple schematic, graphical representations.

J48 is a decision tree classifier. It is an implementation of the C4.5 algorithm [41] in the WEKA workbench. A decision tree learning algorithm approximates a target concept using a tree representation, where each internal node corresponds to an attribute, and each terminal node (called a leaf) corresponds to a class label. The root node is at the top, and the leaves are at the bottom of a tree. Reading a tree from root to leaves, a decision is proposed for a considered object. In this way, it is possible to see the reasons for reaching a given decision. Decision trees are considered not only good classifiers but also as a popular form of knowledge representation. For the Random Forest classifiers, each constituent decision tree is induced from a bootstrap sample of the training data [42].

The k-Nearest Neighbour inducer belongs to instance-based learning approaches [43], which are based on the calculation of similarity (distance) between instances. A parameter k refers to the number of nearest neighbours to be included in the voting process during classification, so each test object has assigned a decision based on the voting of k -nearest neighbours of the considered object. This type of classifier is categorised as a lazy learning algorithm, where the learning process is postponed until the classification point. In the case of decision trees, the learning process is connected with model creation.

4.5. Evaluation of Performance

To evaluate the performance of an inducer, it is necessary to assume some measure. In the problems studied, where classification was binary, classes and data balanced, and both recognised classes of the same importance, with the same misclassification costs, classification accuracy was the suitable measure [44,45]. It was given as the percentage of

correctly classified samples from the test sets, regardless of class (the number of correctly labelled examples divided by the total number of samples expressed as a percentage).

The strategy employed in the selection of samples for evaluation is yet another important issue. Popularly used variants of crossvalidation randomly choose some samples for training and for testing, and such selection is repeated a number of times. The average of the partial results then leads to the final outcome. In the stylometric domain, this approach proves problematic due to the existing stratification of the input space [32]. Data points are grouped by the original long works they are based on. Consequently, random selection leads to many cases where more similar samples are used in both stages, training as well as testing, thus returning over-optimistic results. To address this issue, nonstandard cross-validation can be employed. Then, specific groups of samples are swapped between the training and test sets to make it similar to regular folds. However, such processing involves very high additional costs. Instead, as a compromise, a different attitude can be attempted, with a single training and multiple test sets, from which the average is calculated. This approach was used in the described research works.

5. Performed Experiments

Once the datasets were prepared, experiments involving inducers could be started, and they were divided into three groups. The first batch concerned the application of the classifiers operating in the continuous domain. The second group of tests was focused on the discrete data, but only with the implementation of standard discretisation (one-level and uniform for all attributes) approaches. In the third phase, all classification systems were employed to work on the data discretised by the combined supervised–unsupervised two-level transformations.

5.1. Operation in Continuous Domain

Not all classification systems can handle continuous values of features. The discretisation of data makes it possible to use such systems. However, even for learners working sufficiently well in both domains, discretisation can bring enhanced predictions. Therefore, the performance in the continuous domain should be treated as one of the reference points for comparisons. The results for the selected inducers are given in Table 2. They were calculated as the classification accuracy averaged over the test sets.

Table 2. Average performance [%] of the selected inducers working on the datasets in the continuous input domain.

Dataset	Inducer				
	NB	BNet	J48	kNN	RndF
F-writers	93.34	91.60	89.79	85.56	91.16
M-writers	84.03	77.43	75.63	77.29	75.77

Overall, it could be stated that the male writer dataset proved to be a more difficult task, as for all inducers, the performance was noticeably worse than for the female writer dataset. For both the female and male writers, the highest level of correct predictions was detected for the Naive Bayes classifier. On the other hand, for the female writers, kNN was the worst at attributing samples, while for the male writers, it happened for J48 (but Random Forest was close).

5.2. Standard Discretisation Approaches

As stated before, discretisation involves a transformation of features that can result in an increase in performance, but also a decrease is possible; everything depends on the data and the selected discretisation algorithm. In standard approaches to the task, the same method is applied to all available features, and the popular opinion is that supervised procedures lead to better results than unsupervised ones. However, past studies have showed that, depending on the data, this bad reputation of the latter is not always deserved [25].

Furthermore, in the case when several separate sets with the same features (such as training and test sets) are needed, the methodology of their transformations can play a significant role. Separate sets can be discretised entirely independently, and then intervals and cut points are constructed based only on the local context of each set. The other path of transformations leads through learning discrete data models from the training data, and then the formed intervals are enforced on the test data. In the research described, an independent transformation of all sets was used.

For discretisation using supervised MDL methods, the performance of the classification systems considered is shown in Table 3. In the majority of cases, discretisation resulted in worsened accuracy—sometimes rather slight—but for some inducers, degradation was severe, which happened in consequence of data irregularities existing in the sets. Naive Bayes and Bayesian Net shared some similarities. Therefore, they reported close results, and they suffered the most. Some cases of improvement could be observed for the J48 and kNN classifiers, in particular for the male writer dataset. When the results from the Fayyad and Irani method were compared against those received when the Kononenko method was employed, the latter led to an overall higher accuracy.

Table 3. Performance [%] of inducers for the datasets discretised by the standard supervised Fayyad and Irani (dsF) and Kononenko (dsK) algorithms.

Dataset	Inducer				
	NB	BNet	J48	kNN	RndF
Fayyad and Irani (dsF) method					
F-writers	50.00	50.00	87.78	62.22	62.22
M-writers	69.44	67.22	80.76	83.54	77.22
Kononenko (dsK) method					
F-writers	62.22	62.22	93.40	62.22	74.24
M-writers	68.89	67.22	80.76	82.43	76.11

For the two unsupervised methods examined, the input parameter of a required number of intervals to be constructed was always studied in the range from 2 to 10, and nine variants of the data were explored. For equal width binning, the performance of all inducers was included in Table 4.

NB classifier again suffered for both the female and male writer datasets, regardless of the number of bins considered for equal width binning, yet it was not so bad as in the case of supervised discretisation. For BNet, the results were varied depending on a number of intervals constructed and a dataset. For M-writers, mostly some improved performance was observed, while such a situation for F-writers happened only once when seven bins were formed. For J48, M-writers fared better than F-writers through discretisation. For kNN, the results showed some cases of improvement for the female writers, but for the male writer dataset, mostly worse recognition was noted. Random Forest worked surprisingly well (in comparison to other learners) when higher numbers of intervals were constructed.

Table 5 lists the performance of the inducers for equal frequency binning. Both unsupervised discretisation methods can be compared against each other and measured against supervised approaches, but firstly a look back to the continuous domain was needed.

For the male writers, NB always obtained worse results, but for the female writers in roughly half of the variants, some improvement was noted. BNet showed enhanced results only once for F-writers, yet for M-writers, it showed enhanced results in six out of nine cases. J48 recorded mostly enhanced results for the male writers, but for the female writers, it happened rarely. The performance of kNN was more often worse than better, and the opposite was true for the RndF classifier. Again, the improvement here was more noticeable than for other inducers.

When the unsupervised methods were compared against each other, it turned out that the resulting performance was often better for equal frequency binning. It confirmed some

initial expectations and popular opinions. For most of the inducers, some numbers of bins constructed led to higher accuracy.

For the female and male writer datasets, the discretisation method and, when applicable (as for unsupervised algorithms), the value of the input parameter, which led to the best performance for each studied classifier, were analysed. They are presented in the section dedicated to the analysis of the obtained research results.

Table 4. Performance [%] of inducers for the datasets discretised by the standard unsupervised equal width binning.

Nr of Bins	Dataset	Inducer				
		NB	BNet	J48	kNN	RndF
02	F-writers	87.02	87.50	85.35	88.06	86.32
	M-writers	73.82	73.96	72.71	70.14	73.68
03	F-writers	89.38	91.60	83.68	85.84	87.92
	M-writers	78.68	81.60	77.29	72.36	76.32
04	F-writers	91.67	89.24	86.95	82.78	90.97
	M-writers	80.63	82.92	79.24	71.32	84.52
05	F-writers	90.56	91.11	80.42	84.10	89.86
	M-writers	79.38	79.59	75.77	69.52	76.32
06	F-writers	92.22	88.06	85.70	85.21	89.24
	M-writers	82.22	79.79	75.21	77.22	82.16
07	F-writers	91.11	93.41	89.66	86.46	92.85
	M-writers	81.04	77.57	71.95	72.29	81.74
08	F-writers	91.67	88.27	90.97	85.77	92.78
	M-writers	77.64	81.39	75.42	73.61	82.22
09	F-writers	90.56	91.74	90.35	88.68	94.59
	M-writers	81.53	76.74	76.46	74.59	85.14
10	F-writers	91.67	86.95	87.02	86.32	93.96
	M-writers	80.00	80.77	76.25	73.61	81.04

Table 5. Performance [%] of inducers for the datasets discretised by the standard unsupervised equal frequency binning.

Nr of Bins	Dataset	Inducer				
		NB	BNet	J48	kNN	RndF
02	F-writers	91.18	90.63	91.18	83.89	87.64
	M-writers	75.35	76.32	73.82	68.47	75.97
03	F-writers	92.85	88.68	85.63	86.95	90.35
	M-writers	75.70	81.11	79.72	71.25	79.31
04	F-writers	93.41	92.85	86.46	88.61	91.67
	M-writers	81.60	81.74	69.31	75.21	84.38
05	F-writers	92.85	85.63	86.18	85.07	92.92
	M-writers	81.11	75.77	78.96	73.54	82.71
06	F-writers	94.10	89.86	88.20	89.17	94.03
	M-writers	80.35	81.11	74.79	73.47	80.35
07	F-writers	95.77	89.31	81.74	85.63	95.84
	M-writers	80.42	81.67	78.75	77.85	82.78
08	F-writers	92.29	86.25	88.20	88.47	92.78
	M-writers	80.49	78.20	77.92	75.28	81.60
09	F-writers	94.59	89.93	88.27	86.81	93.96
	M-writers	80.42	80.49	79.79	77.57	83.96
10	F-writers	93.47	91.60	87.16	86.88	92.85
	M-writers	79.86	77.02	80.91	75.35	83.41

5.3. Two-Level Discretisation

As could be observed above in the reported research results, supervised discretisation can return such variants of the data where noticeably many input features are treated as useless in a discrete domain. Also, unsupervised algorithms can lead to more advantageous performance of classifiers working on the discretised datasets. These observations provided motivation for a more detailed study which combined both approaches.

In the proposed research framework, the fundamental notion was to get more out of variables than supervised discretisation could offer. Therefore, for all features, for which through supervised processing only single intervals were found to represent their values, an additional transformation step was executed this time employing unsupervised methods. Consequently, in a discrete domain, all variables had at least two bins assigned, and all could be mined in data exploration processes.

Unsupervised discretisation algorithms were used with varying the input parameter that defined the number of bins requested, and they were paired with both the Fayyad and Irani and Kononenko discretisation methods, which resulted in multiple versions of the datasets. All these variants were next mined with the group of selected classification systems, and their performance was observed. Some inducers showed similar trends, while others varied greatly. In the presented results, the bin number equal one denotes the situation where a dataset was transformed only by supervised methods; thus, some variables were assigned single intervals. When the number given is greater than one, it means additional transformations by some unsupervised algorithm, and then the number corresponds to the input parameter. The combination of the Fayyad and Irani method with equal width binning was denoted as dsF-duw and with equal frequency binning as dsF-duf. In the same convention, dsK-duw and dsK-duf stand for the combination of the Kononenko algorithm with equal width binning and equal frequency binning, respectively.

BNet stood out from the other learners. Its performance was included in Table 6. For the female and male writer dataset, for all combinations of supervised and unsupervised methods, the change in performance was observed only with varying the bin number in the test sets. Changing the number of intervals in the training sets with the constant number of bins in the test sets brought the same accuracy. For each pairing, it can be noted that additional transformations by unsupervised discretisation algorithms caused increased performance. For F-writers, the trends were monotonic; for M-writers, they were close to monotonic.

Table 6. Performance [%] of BNet inducer for the datasets transformed through supervised discretisation by the Fayyad and Irani (dsF) and Kononenko (dsK) algorithms combined with unsupervised equal frequency (duf) or equal width (duw) binning.

ds-du	Number of Bins in Test Sets									
	1	2	3	4	5	6	7	8	9	10
F-writers dataset										
dsF-duf	50.00	61.39	74.31	81.18	83.47	86.94	89.38	89.38	90.49	91.04
dsF-duw	50.00	57.29	68.40	75.63	82.85	86.39	89.93	92.22	92.22	92.22
dsK-duf	62.22	79.17	85.76	88.19	90.42	91.60	92.22	92.22	92.22	92.22
dsK-duw	62.22	71.25	85.21	90.42	91.67	92.22	92.78	93.40	93.40	93.40
M-writers dataset										
dsF-duf	67.22	71.67	74.65	77.22	77.78	77.78	77.85	77.29	76.67	77.85
dsF-duw	67.22	73.61	77.85	79.72	79.72	77.29	78.47	79.10	79.10	78.47
dsK-duf	67.22	71.67	74.65	77.22	77.78	77.78	77.85	77.29	76.67	77.85
dsK-duw	67.22	73.61	77.85	79.72	79.72	77.29	78.47	79.10	79.10	78.47

For the F-writers, the Kononenko method used as a base always led to better results than the Fayyad and Irani algorithm: dsK-duf and dsK-duw resulted in better performance of the BNet than that obtained for dsF-duf and dsF-duw, regardless of the number of bins constructed in

additional second-level transformations. However, for the M-writers, the visible trend depended on the unsupervised method—for equal width binning combined with either of the supervised discretisation processes, the accuracy was higher than for equal frequency binning.

For the remaining classifiers, the performance was much more varied, as shown in the plots in Figures 2–5. In each chart, the categories describing the horizontal axis correspond to the number of bins for the additionally transformed variables in the training sets, while the series was defined by the numbers of bins in the test sets. The top four plots are for F-writers, and the bottom ones for the male writer datasets. Combinations of discretisation methods are commented on in the chart titles.

For the NB classifier (Figure 2), the trends indicate that for the female writers, some similarities could be observed with respect to the performance of BNet. It is visible that additional transformations of variables in the test sets brought more noticeable results than when the training sets were subjected to the second level of discretisation. Such a statement for the M-writers would not be true; here, differences were visible for both directions.

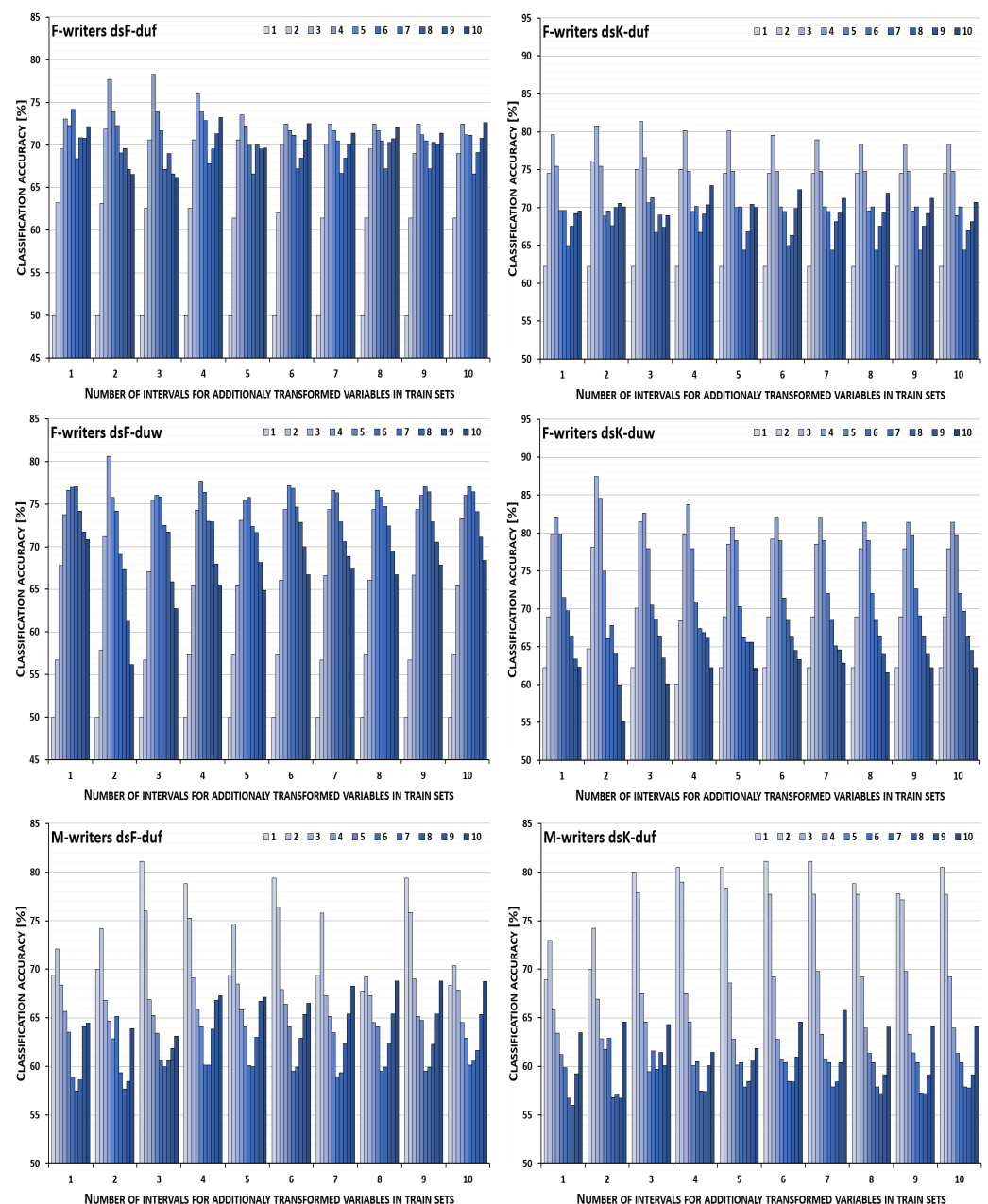


Figure 2. Cont.

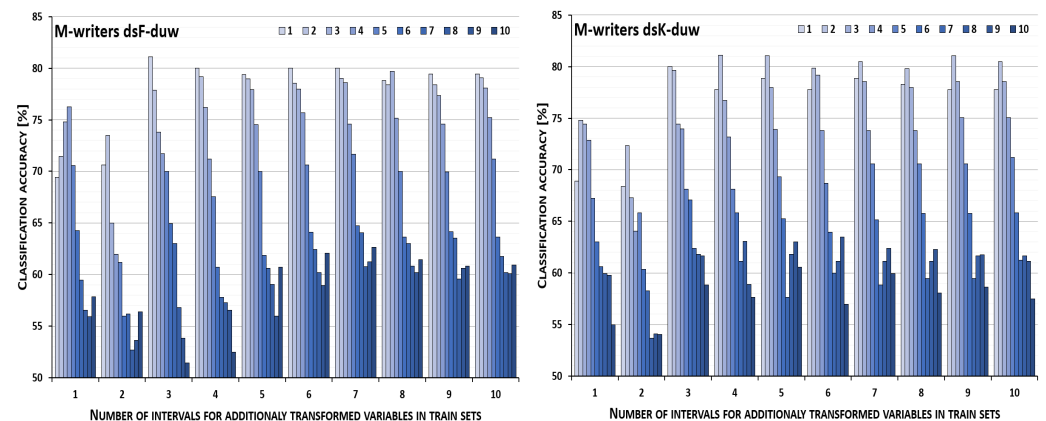


Figure 2. Performance [%] of the Naive Bayes classifier for the data transformed through supervised discretisation by the Fayyad and Irani (dsF) and Kononenko (dsK) algorithms combined with unsupervised equal frequency (duf) or equal width (duw) binning.

For both the female and male writer datasets, additional processing of variables caused enhanced performance in many cases but rarely for the higher ranges of bin numbers. In particular, for M-writers, fewer bins led to enhanced predictions, while more intervals brought worsened accuracy. For F-writers, the trend lines were much clearer and showed that for the Fayyad-based data variants, increasing the number of bins in the test sets above one caused enhanced results. For the Kononenko-based versions, the opposite situation was observed. For M-writers, the changes occurred in both directions.

An analysis of the plots for J48 (Figure 3) resulted in the conclusion that for the female writers the trends visible for categories were very similar, with hardly any differences. It indicated a higher degree of dependence on transformations of the test sets than of the training sets. For the male writers, much more variation was noticed, with yet again many cases of improved predictions. As previously for NB, for F-writers, closer similarities were observed between the plots based on the same supervised discretisation method, while for M-writers, the closeness depended on the unsupervised method employed.

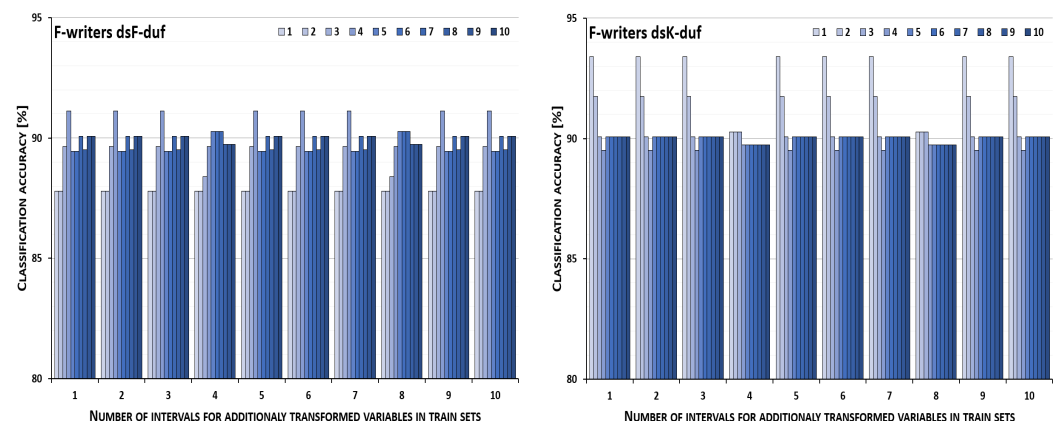


Figure 3. Cont.

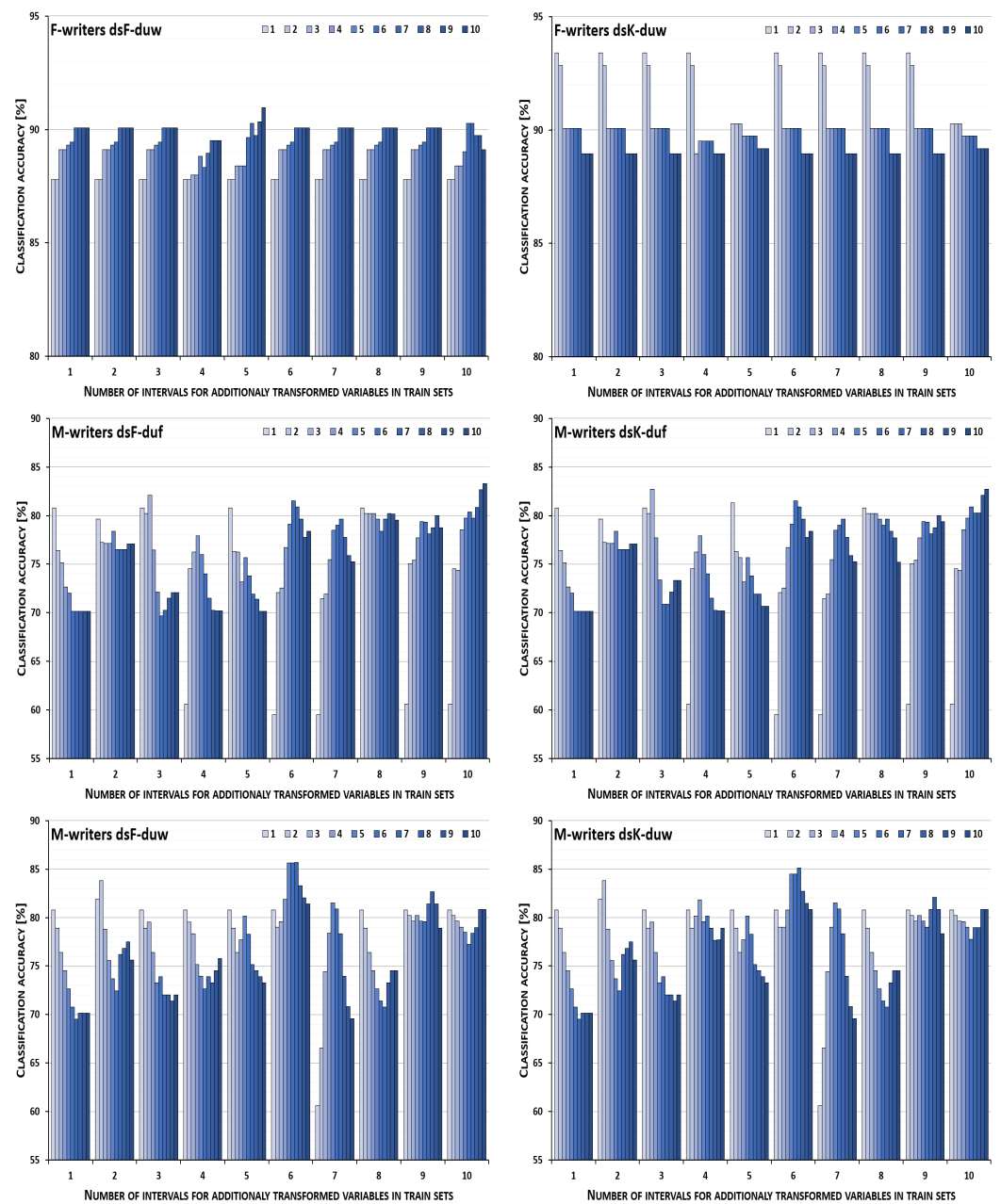


Figure 3. Performance [%] of the J48 classifier for the data transformed through supervised discretisation by the Fayyad and Irani (dsF) and Kononenko (dsK) algorithms combined with unsupervised equal frequency (duf) or equal width (duw) binning.

The kNN classifier (Figure 4) behaved in a rather distinctive manner. For the female writer dataset, the second-level discretisation of the training sets rarely brought better results, but additional transformations of the test sets almost always increased the accuracy by a noticeable degree. For the male writer dataset, in both groups of processing, the predictive power was mostly degraded, and hardly any improvement could be found among all tested variants of the data.

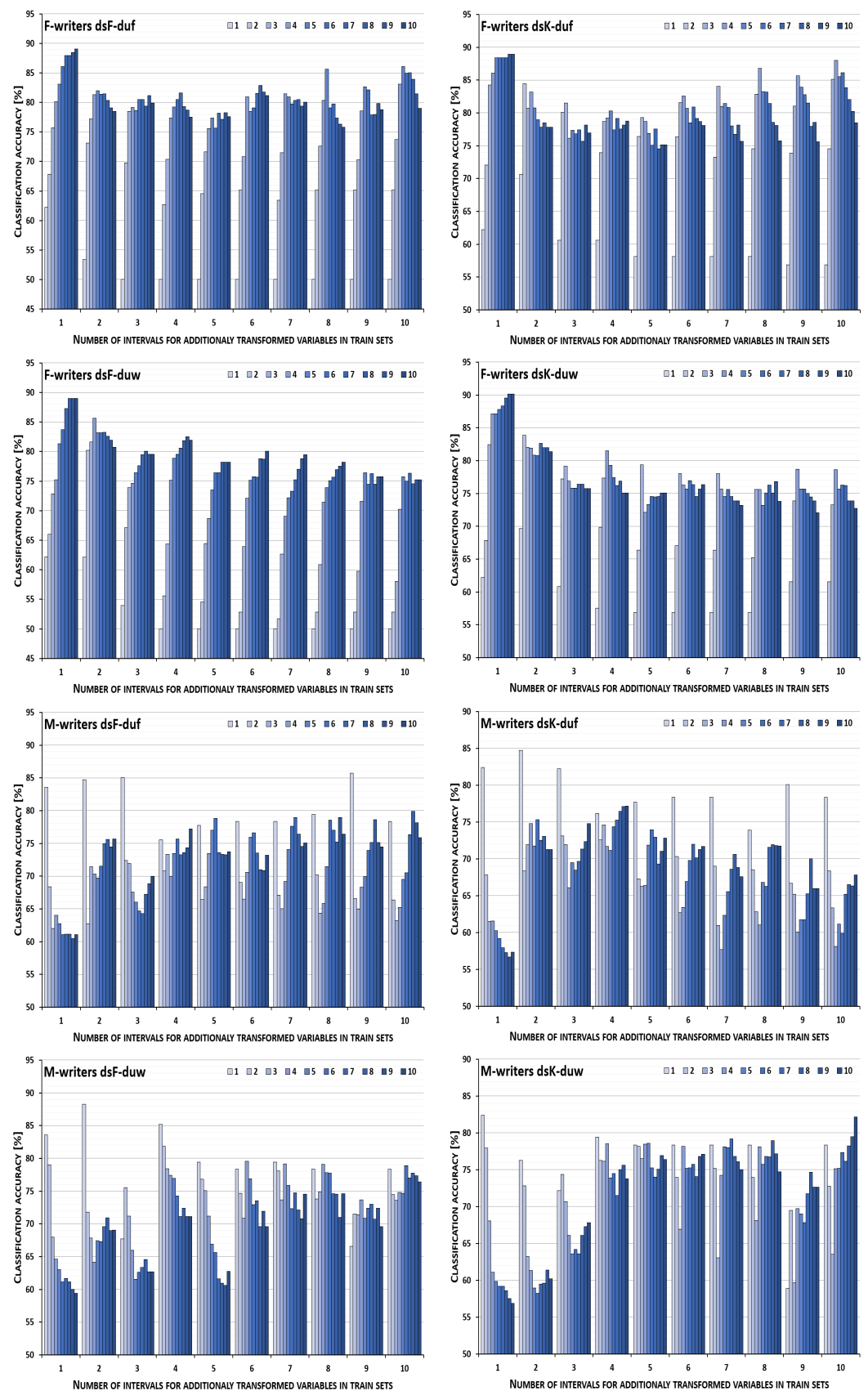


Figure 4. Performance [%] of the kNN classifier for the data transformed through supervised discretisation by the Fayyad and Irani (dsF) and Kononenko (dsK) algorithms combined with unsupervised equal frequency (duf) or equal width (duw) binning.

The performance of Random Forest is shown in Figure 5. For the female writer dataset, high dependence on the number of bins in the test sets was visible for all combinations of discretisation methods. The opposite statement was true for M-writers, where transformations of the test sets caused rather worsening results.

The presented research results showed that two-level (instead of standard one-level) discretisation of attributes in the training and test sets, by combining supervised with unsupervised transformations, can be expected to have some influence on the performance of the classifiers working on discrete versions of the data. However, the charts presented differences between trends that were noticeable for classifiers, for datasets, and for combinations of methods. Therefore, the proposed methodology proved its merit, yet the answer to such a question as which method or combination of methods for each inducer is most advantageous, or which classifier works best, requires more analysis and is addressed in the next section of the paper.

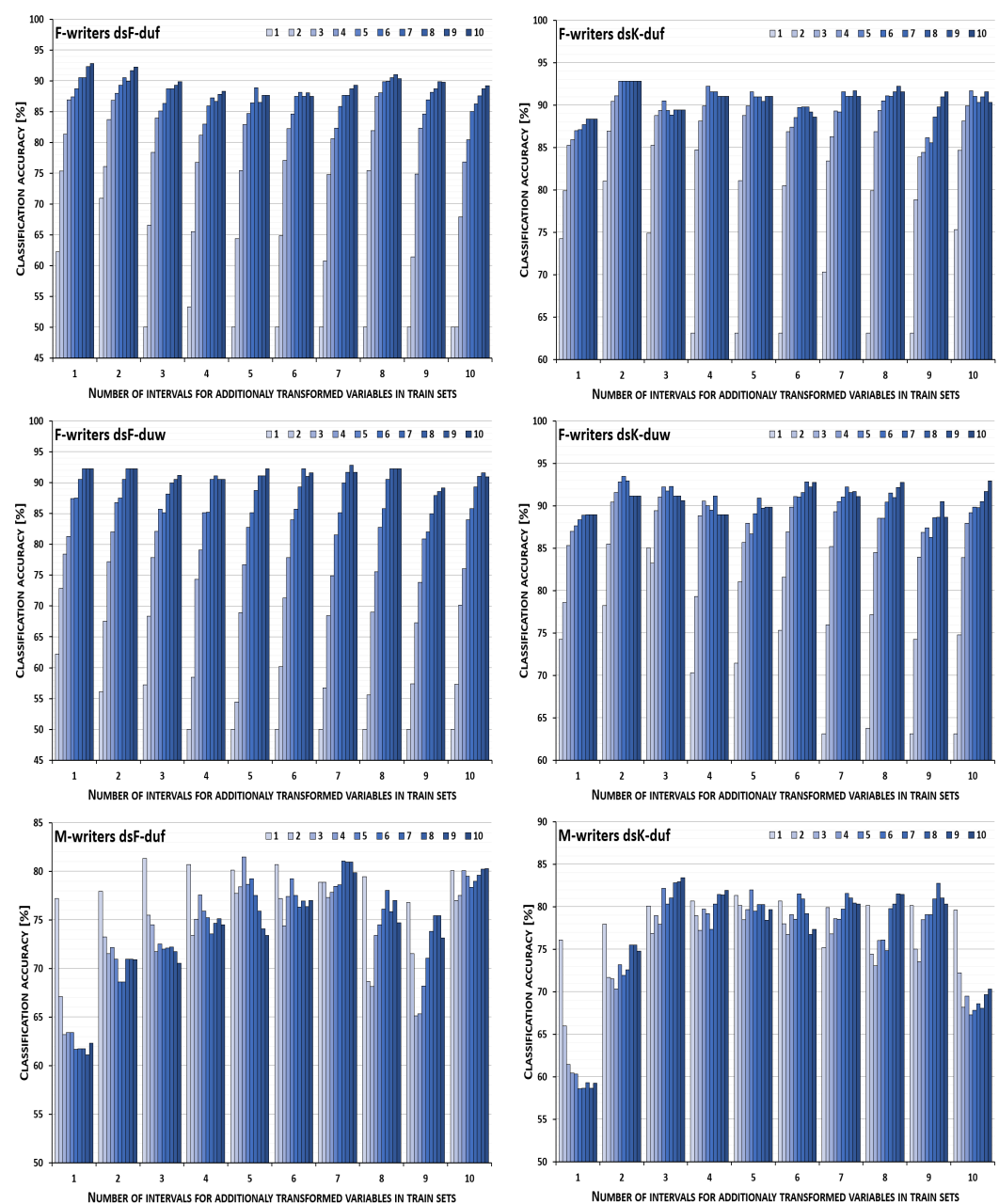


Figure 5. Cont.

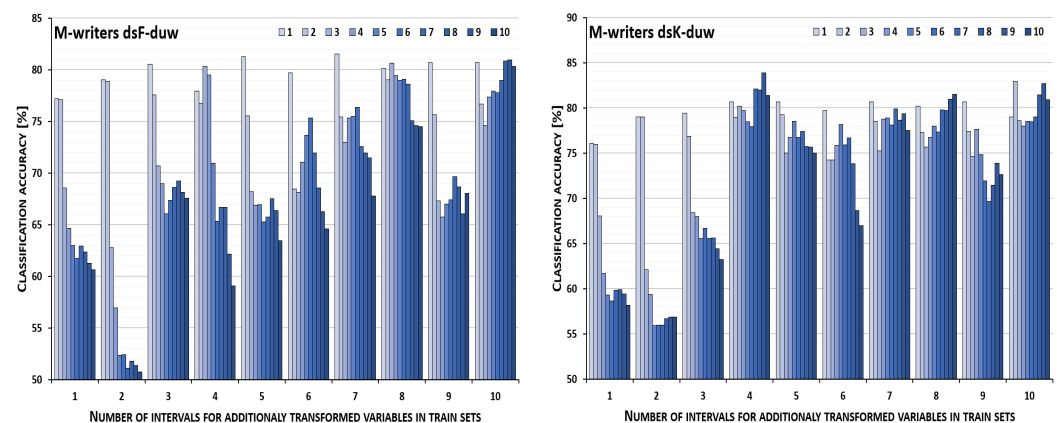


Figure 5. Performance [%] of the Random Forest (RndF) classifier for the data transformed through supervised discretisation by the Fayyad and Irani (dsF) and Kononenko (dsK) algorithms combined with unsupervised equal frequency (duf) or equal width (duw) binning.

6. Discussion of Obtained Results

The choice of a particular discretisation method for transformation of the input data, which would be most advantageous for the task and machine learning methods employed to solve it, is not trivial. Popular beliefs in the quality of approaches should not be accepted blindly. Firstly, the obtained results should be considered in terms of some characteristics and statistics, and only with such support an informed decision can be made.

In the research, there were calculated such elements as the minimum, maximum, average, and standard deviation. For any detected maximum, conditions leading to it were also considered, which were understood as the number of bins in either the training or test (or both) sets. These statistics were found for the standard unsupervised discretisation algorithms applied to the data, and also for all two-level discretisation approaches.

For standard unsupervised discretisation with both methods, the characteristics are gathered in Table 7. The rows $i-T_{Max}$ list the number of bins constructed in both the training and the test sets that corresponded to the data variant for which the performance was maximal. For all classifiers, the calculated standard deviation was relatively low. For Naive Bayes and Random Forest for F-writers, a better predictive power was found after discretisation with equal frequency binning, and for M-writers, it happened for the equal width binning method. For BNet for both the female and male datasets, the results were generally better when equal width discretisation was executed. For J48 and kNN classifiers, equal frequency binning was more advantageous. For F-writers, in noticeably more cases, higher bin numbers were associated with maximal performance, while for M-writers, the difference was not so distinctive.

Achieving the highest predictive power is a typical goal of any data exploration process, so for all standard (one-level and uniform for all variables) discretisation approaches, the maximal classification accuracy was found. It is displayed in Table 8, where it can be observed that for both the female and male writer dataset for all classifiers, some supervised discretisation algorithm led to the best performance only once. For J48 and F-writers, it was the Kononenko algorithm, and for kNN and M-writers, it was the Fayyad and Irani method. In all remaining cases, unsupervised discretisation was more advantageous; in particular, equal frequency binning quite often resulted in the highest predictions.

Similar statistics were calculated for all four main discretisation procedures for two-level transformations, that is, for the combination of either the Fayyad and Irani or the Kononenko algorithms with either equal width or equal frequency binning. They are summarised in Tables 9 and 10. The rows specifying $i-Train_{Max}$ and $i-Test_{Max}$ provide information about the numbers of bins in the training and test sets correspondingly, for which the maximal predictive power was detected. $i = 1$ denotes a situation when a set was subjected only to the first level of transformations, that is, supervised discretisation.

When maxima occurred for several variants, the numbers of bins are either separated by a comma or given as a range.

Table 7. Statistics of performance [%] for the inducers working on the datasets discretised by unsupervised algorithms (duf and duw).

Statistics	Inducer				
	NB	BNet	J48	kNN	RndF
F-writers, equal frequency binning (duf)					
Avg	93.39	89.42	87.00	86.83	92.45
St.dev.	01.34	02.33	02.56	01.75	02.37
Min	91.18	85.63	81.74	83.89	87.02
Max	95.77	92.85	91.18	89.17	95.84
<i>i-T_{Max}</i>	07	04	02	06	07
F-writers, equal width binning (duw)					
Avg	90.65	89.76	86.68	85.91	90.94
St.dev.	01.60	02.26	03.39	01.81	02.83
Min	87.02	86.95	80.42	82.78	86.32
Max	92.22	93.41	90.97	88.68	94.59
<i>i-T_{Max}</i>	06	07	08	09	09
M-writers, equal frequency binning (duf)					
Avg	79.48	79.27	77.11	74.22	81.61
St.dev.	02.30	02.43	03.75	02.97	02.69
Min	75.35	75.77	69.31	68.47	75.97
Max	81.60	81.74	80.91	77.85	84.38
<i>i-T_{Max}</i>	04	04	10	07	04
M-writers, equal width binning (duw)					
Avg	79.44	79.37	75.59	72.74	80.35
St.dev.	02.54	02.81	02.21	02.36	03.98
Min	73.82	73.96	71.95	69.52	73.68
Max	82.22	82.92	79.24	77.22	85.14
<i>i-T_{Max}</i>	06	04	04	06	09

Table 8. The best performance [%] and the corresponding discretisation method for the standard approaches applied to the data.

Dataset	Inducer				
	NB	BNet	J48	kNN	RndF
F-writers	duf07	duw07	dsK	duf06	duf07
	95.77	93.41	93.40	89.17	95.84
M-writers	duw06	duw04	duf10	dsF	duw09
	82.22	82.92	80.91	83.54	85.14

The characteristics calculated for the dsF-duf and dsF-duw procedures (Table 9) show that only for the J48 learner did the standard deviation remain at a relatively low level. For the other inducers, greater differences among test results caused increased values for this statistic. For both the unsupervised discretisation methods employed on the second level of transformations, the minima were the same for the female writers for all inducers, but for the male writer dataset, differences could be noted. The maxima were close for both equal frequency and equal width algorithms, for both F-writers and M-writers and all learners. The provided bin numbers that were found accompanying the maxima indicate that, in the majority of cases, second level discretisation was advantageous. For some of the inducers, that was true for either the training or test sets, but for others, that was true also for both types of sets.

Table 10 includes the summarising characteristics for processing involving the Kononenko algorithm combined with unsupervised methods. Overall, the standard deviation for this group of discretisation approaches was rather lower than for the Fayyad and Irani methods used for the first level of discretisation. Yet still, J48 stood out from the other classifiers as the one with the lowest values of this characteristics.

Table 9. Statistics of performance [%] for the inducers working on the datasets discretised by two-level transformations combining the supervised Fayyad and Irani (dsF) method with unsupervised algorithms.

	Inducer									
	NB	BNet	J48	kNN	RndF	NB	BNet	J48	kNN	RndF
Statistics	F-writers, equal frequency binning (dsF-duf)					F-writers, equal width binning (dsF-duw)				
Avg	67.75	79.76	89.47	75.54	80.83	68.15	78.72	89.18	72.27	79.64
St.dev.	06.80	13.22	00.97	09.71	12.06	08.51	14.66	00.89	10.52	13.37
Min	50.00	50.00	87.78	50.00	50.00	50.00	50.00	87.78	50.00	50.00
Max	78.26	91.04	91.11	89.03	92.85	80.63	92.22	90.97	89.03	92.85
<i>i-Train</i> _{Max}	03	01-10	01-03 05-07 09-10	01	01	02	01-10	05	01	07
<i>i-Test</i> _{Max}	04	10	04	10	10	04	08-10	10	08-10	09
Statistics	M-writers, equal frequency binning (dsF-duf)					M-writers, equal width binning (dsF-duw)				
Avg	65.69	75.60	75.49	72.03	74.44	67.23	77.06	76.89	71.86	70.55
St.dev.	05.22	03.36	05.04	05.72	05.15	08.56	03.69	04.38	06.09	07.80
Min	57.50	67.22	59.51	60.49	61.11	51.46	67.22	60.63	59.38	50.76
Max	81.11	77.85	83.26	85.69	81.46	81.11	79.72	85.69	88.26	81.53
<i>i-Train</i> _{Max}	03	01-10	10	09	05	03	01-10	06	02	07
<i>i-Test</i> _{Max}	01	07, 10	10	01	04	01	04-05	07	01	01

Table 10. Statistics of performance [%] for the inducers working on the datasets discretised by two-level transformations combining the supervised Kononenko (dsK) method with unsupervised approaches.

	Inducer									
	NB	BNet	J48	kNN	RndF	NB	BNet	J48	kNN	RndF
Statistics	F-writers, equal frequency binning (dsK-duf)					F-writers, equal width binning (dsK-duw)				
Avg	70.47	86.62	90.39	78.01	87.00	70.35	86.60	90.19	74.60	86.82
St.dev.	04.83	09.05	01.03	07.22	06.95	07.34	10.40	01.37	07.77	06.99
Min	62.22	62.22	89.51	56.88	63.13	55.07	62.22	88.96	50.00	63.13
Max	81.32	92.22	93.40	88.96	92.85	87.43	93.40	93.40	90.14	93.47
<i>i-Train</i> _{Max}	03	01-10	01-03 05-07 09-10	01	01	02	01-10	01-04	01	02
<i>i-Test</i> _{Max}	03	07-10	01	09-10	05-10	03	08-10	01	09-10	06
Statistics	M-writers, equal frequency binning (dsK-duf)					M-writers, equal width binning (dsK-duw)				
Avg	64.76	75.60	76.51	69.03	76.09	67.78	77.06	77.20	71.80	73.77
St.dev.	07.23	03.36	04.95	06.03	06.30	08.03	03.69	04.27	06.88	07.78
Min	56.04	67.22	59.51	56.67	58.61	53.68	67.22	60.63	56.88	55.90
Max	81.11	77.85	82.71	84.72	83.40	81.11	79.22	85.14	82.43	83.89
<i>i-Train</i> _{Max}	06-07	01-10	03, 10	02	03	04	01-10	06	01	04
<i>i-Test</i> _{Max}	01	07, 10	03, 10	01	10	02	04-05	07	01	09

When the Kononenko-based two-level discretisation was compared with the Fayyad and Irani employed on the first level, it was observed that for most classifiers, the former was slightly better than the latter, in particular for the female writer dataset. The values of calculated averages, minima, and maxima were typically higher, while the standard deviation was smaller. In addition, for these variants of the data, in most cases additional transformations of 1-bin variables in sets resulted in enhanced predictions.

All the detailed results from the substantial experiments performed, given in the previous section in the tables and charts and supported with summaries and characteristics included in this section, illustrate the advantages of employing the two-level approach to

discretisation and combining supervised and unsupervised methods. When the characteristics of the input domain, as well as specific modes of operation of inducers, are taken into account, enhanced predictions can follow. Such increased efficiency in recognition is typically treated as the most important factor in machine learning tasks.

The presented research methodology can be applied in any domain characterised by the continuous input space, where, after a supervised discretisation process, such attributes exist that are assigned a single categorical representation. The main goal of the proposed two-step procedure is to prevent the loss of information that occurs in such conditions. The algorithm causes increased complexity in the data preparation stage, because discretisation is closer adapted to the data. Instead of relatively straightforward transformations, one-step and uniform for all features, discretisation is performed in two subsequent stages. In the first step, some standard supervised algorithm is applied to all attributes. Then, instead of treating single-bin attributes as irrelevant, for the second step, they are retrieved from the set of available variables and subjected to unsupervised discretisation. If there are no such attributes, or when their subset is relatively small (compared to the entire set of available features), the extended processing may bring no visible gain. However, through the fusion of supervised and unsupervised discretisation, the procedure limits the information loss and makes it possible to preserve more informative content of the attributes than any standard one-level method can offer. As a consequence, meaningful categorical representations are obtained for all features. When transforming procedures are adapted closer to a particular dataset in this way, a classifier working on it has access to more information and can learn more, which can result in a noticeable improvement in performance. This potential gain should be measured against the costs of the extended processing in the form of additional transformations of attribute domains.

7. Conclusions

Discretisation is often considered to be a part of initial preprocessing of the input data, thereby leading to data cleaning and possibly some reduction in dimensionality. In standard approaches, supervised methods are widely preferred, and the same type of transformation is applied to all available variables.

This paper presents research in which the methodology for data processing was proposed, thereby employing a combination of supervised with unsupervised discretisation algorithms. The former find intervals to represent attribute values by evaluating with some measure the possible placement of borders between these intervals. In a top-down approach, it means starting with assigning a single interval to represent the entire attribute domain and then looking to some stopping criterion. When this criterion is satisfied, the processing stops, and, as a consequence, some variables can remain represented in a discrete domain by these single bins. It makes them irrelevant to classification, and the information they provide in the continuous domain becomes inaccessible or lost in translation. To avoid that, in the proposed methodology, such special 1-bin variables were subjected to additional transformations with unsupervised methods. The extensive experiments, executed in the stylometric domain on the binary authorship attribution problems, showed the merits of such two-level discretisation, as the deeper processing of some variables led to many cases of improved performance of the group of considered classifiers.

The future research will be dedicated to application of the methodology to other application domains to evaluate the influence of data characteristics on the whole procedure. Other unsupervised discretisation algorithms will also be tested, for example, equal frequency binning with weights. Another path to tread is to establish some pointers for selection of a specific combination of discretisation methods that would be expected to offer improved performance.

Author Contributions: Conceptualization, U.S.; methodology, U.S., B.Z., and G.B.; software, G.B.; validation, U.S., B.Z., and G.B.; formal analysis, U.S. and B.Z.; data curation, U.S. and G.B.; writing—original draft preparation, U.S. and B.Z.; writing—review and editing, U.S. and B.Z.; visualization, U.S. and B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Texts exploited in the experiments are available for online reading and download thanks to Project Gutenberg (www.gutenberg.org, accessed on 8 May 2024). The prepared datasets are available upon request.

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

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dash, R.; Paramguru, R.L.; Dash, R. Comparative analysis of supervised and unsupervised discretization techniques. *Int. J. Adv. Sci. Technol.* **2011**, *2*, 29–37.
2. Anh, C.T.; Kwon, Y.K. Mutual Information Based on Multiple Level Discretization Network Inference from Time Series Gene Expression Profiles. *Appl. Sci.* **2023**, *13*, 11902. [\[CrossRef\]](#)
3. Toulabinejad, E.; Mirsafaei, M.; Basiri, A. Supervised discretization of continuous-valued attributes for classification using RACER algorithm. *Expert Syst. Appl.* **2024**, *244*, 121203. [\[CrossRef\]](#)
4. Grzymala-Busse, J.W. Discretization Based on Entropy and Multiple Scanning. *Entropy* **2013**, *15*, 1486–1502. [\[CrossRef\]](#)
5. 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, Chambéry, France, 28 August–3 September 1993; Morgan Kaufmann Publishers: San Francisco, CA, USA, 1993; Volume 2, pp. 1022–1027.
6. Kononenko, I.; Kukar, M. Data Preprocessing. In *Machine Learning and Data Mining*; Kononenko, I., Kukar, M., Eds.; Woodhead Publishing: Sawston, Verenigd Koninkrijk, 2007; Chapter 7, pp. 181–211.
7. Argamon, S.; Burns, K.; Dubnov, S. (Eds.) *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning*; Springer: Berlin, Germany, 2010.
8. 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. *Front. Digit. Humanit.* **2018**, *5*, 4. [\[CrossRef\]](#)
9. Eder, M.; Rybicki, J. Do birds of a feather really flock together, or how to choose training samples for authorship attribution. *Lit. Linguist. Comput.* **2013**, *28*, 229–236. [\[CrossRef\]](#)
10. Kalaivani, K.; Kuppuswami, S. Exploring the use of syntactic dependency features for document-level sentiment classification. *Bull. Pol. Acad. Sci. Tech. Sci.* **2019**, *67*, 339–347. [\[CrossRef\]](#)
11. Koppel, M.; Schler, J.; Argamon, S. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 9–26. [\[CrossRef\]](#)
12. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*; Morgan Kaufmann: San Francisco, CA, USA, 2011.
13. Sharma, P.; Neeli, S. A systematic review of discretisation methods for time-delay systems. *J. Control. Decis.* **2023**, 1–16.
14. Danyal, M.M.; Khan, S.S.; Khan, M.; Ullah, S.; Ghaffar, M.B.; Khan, W. Sentiment analysis of movie reviews based on NB approaches using TF-IDF and count vectorizer. *Soc. Netw. Anal. Min.* **2024**, *14*, 87. [\[CrossRef\]](#)
15. Cios, K.J.; Pedrycz, W.; Świniarski, R.W.; Kurgan, L. *Data Mining. A Knowledge Discovery Approach*; Springer: New York, NY, USA, 2007.
16. Peker, N.; Kubat, C. Application of Chi-square discretization algorithms to ensemble classification methods. *Expert Syst. Appl.* **2021**, *185*, 115540. [\[CrossRef\]](#)
17. Dhont, M.; Tsiorkova, E.; Boeva, V. Advanced Discretisation and Visualisation Methods for Performance Profiling of Wind Turbines. *Energies* **2021**, *14*, 6216. [\[CrossRef\]](#)
18. Kliegr, T.; Izquierdo, E. QCBA: improving rule classifiers learned from quantitative data by recovering information lost by discretisation. *Appl. Intell.* **2023**, *53*, 20797–20827. [\[CrossRef\]](#)
19. Saeed, N.; Manguri, A.; Szczepanski, M.; Jankowski, R. Non-Linear Analysis of Structures Utilizing Load-Discretization of Stiffness Matrix Method with Coordinate Update. *Appl. Sci.* **2022**, *12*, 2394. [\[CrossRef\]](#)
20. Rissanen, J. Modeling by shortest data description. *Automatica* **1978**, *14*, 465–471. [\[CrossRef\]](#)
21. Ross Quinlan, J.; Rivest, R.L. Inferring decision trees using the minimum description length principle. *Inf. Comput.* **1989**, *80*, 227–248. [\[CrossRef\]](#)

22. Stańczyk, U.; Zielosko, B. On Combining Discretisation Parameters and Attribute Ranking for Selection of Decision Rules. *Lect. Notes Comput. Sci.* **2017**, *10313*, 329–349.
23. de Sá, C.R.; Soares, C.; Knobbe, A. Entropy-based discretization methods for ranking data. *Inf. Sci.* **2016**, *329*, 921–936. [\[CrossRef\]](#)
24. Stańczyk, U. Evaluating Importance for Numbers of Bins in Discretised Learning and Test Sets. In *Intelligent Decision Technologies 2017, Proceedings of the 9th KES International Conference on Intelligent Decision Technologies (KES-IDT 2017)—Part II*, Vilamoura, Portugal, 21–23 June 2017; Czarnowski, I., Howlett, J.R., Jain, C.L., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 72, pp. 159–169.
25. Stańczyk, U.; Zielosko, B. Data irregularities in discretisation of test sets used for evaluation of classification systems: A case study on authorship attribution. *Bull. Pol. Acad. Sci. Tech. Sci.* **2021**, *69*, 1–12. [\[CrossRef\]](#)
26. Rybicki, J.; Eder, M.; Hoover, D. Computational stylistics and text analysis. In *Doing Digital Humanities: Practice, Training, Research*, 1st ed.; Crompton, C., Lane, R., Siemens, R., Eds.; Routledge: London, UK, 2016; pp. 123–144.
27. Eder, M.; Górski, R.L. Stylistic Fingerprints, POS-tags, and Inflected Languages: A Case Study in Polish. *J. Quant. Linguist.* **2022**, *30*, 86–103. [\[CrossRef\]](#)
28. Misini, A.; Kadriu, A.; Canhasi, E. A Survey on Authorship Analysis Tasks and Techniques. *Seeu Rev.* **2022**, *17*, 153–167. [\[CrossRef\]](#)
29. Stamatatos, E. A Survey of Modern Authorship Attribution Methods. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 538–556. [\[CrossRef\]](#)
30. Eder, M. Does size matter? Authorship attribution, small samples, big problem. *Digit. Scholarsh. Humanit.* **2015**, *30*, 167–182. [\[CrossRef\]](#)
31. Škorić, M.; Stanković, R.; Ikonić Nešić, M.; Byszuk, J.; Eder, M. Parallel Stylometric Document Embeddings with Deep Learning Based Language Models in Literary Authorship Attribution. *Mathematics* **2022**, *10*, 838. [\[CrossRef\]](#)
32. Baron, G.; Stańczyk, U. Standard vs. non-standard cross-validation: evaluation of performance in a space with structured distribution of datapoints. *Procedia Comput. Sci.* **2021**, *192*, 1245–1254.
33. Zielosko, B.; Piliszczuk, M. Greedy Algorithm for Attribute Reduction. *Fundam. Informaticae* **2008**, *85*, 549–561.
34. He, X.; Lashkari, A.H.; Vombatkere, N.; Sharma, D.P. Authorship Attribution Methods, Challenges, and Future Research Directions: A Comprehensive Survey. *Information* **2024**, *15*, 131. [\[CrossRef\]](#)
35. Sbalchiero, S.; Eder, M. Topic modeling, long texts and the best number of topics. Some Problems and solutions. *Qual. Quant.* **2020**, *54*, 1095–1108. [\[CrossRef\]](#)
36. Weidman, S.G.; O’Sullivan, J. The limits of distinctive words: Re-evaluating literature’s gender marker debate. *Digit. Scholarsh. Humanit.* **2018**, *33*, 374–390. [\[CrossRef\]](#)
37. Lai, J.; Yang, X.; Luo, W.; Zhou, L.; Li, L.; Wang, Y.; Shi, X. RumorLLM: A Rumor Large Language Model-Based Fake-News-Detection Data-Augmentation Approach. *Appl. Sci.* **2024**, *14*, 3532. [\[CrossRef\]](#)
38. Fernández, A.; García, S.; Galar, M.; Prati, R.C.; Krawczyk, B.; Herrera, F. Data Level Preprocessing Methods. In *Learning from Imbalanced Data Sets*; Springer International Publishing: Cham, Switzerland, 2018; pp. 79–121.
39. Zielosko, B.; Stańczyk, U.; Jabłoński, K. Filtering Decision Rules Driven by Sequential Forward and Backward Selection of Attributes: An Illustrative Example in Stylometric Domain. *Ann. Comput. Sci. Inf. Syst.* **2023**, *35*, 833–842.
40. Witten, I.; Frank, E.; Hall, M. *Data Mining. Practical Machine Learning Tools and Techniques*, 3rd ed.; Morgan Kaufmann: San Francisco, CA, USA, 2011.
41. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1993.
42. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
43. Lv, Y.; Zhang, B.; Yue, X.; Dencœux, T.; Yue, S. Selecting reliable instances based on evidence theory for transfer learning. *Expert Syst. Appl.* **2024**, *250*, 123739. [\[CrossRef\]](#)
44. Stańpor, K. Evaluation of classifiers: current methods and future research directions. *ACSIS* **2017**, *13*, 37–40.
45. Stańpor, K.; Ksieniewicz, P.; García, S.; Woźniak, M. How to design the fair experimental classifier evaluation. *Appl. Soft Comput.* **2021**, *104*, 107219. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.