

**Title:** On Approaches to Discretisation of Stylometric Data and Conflict Resolution in Decision Making

**Author:** Urszula Stańczyk, Beata Zielosko

**Citation style:** Stańczyk Urszula, Zielosko Beata. (2019). On approaches to discretisation of stylometric data and conflict resolution in decision making. "Procedia Computer Science" (2019, Vol. 159C, s. 1811-1820), doi 10.1016/j.procs.2019.09.353

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2019) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

# On Approaches to Discretisation of Stylometric Data and Conflict Resolution in Decision Making

Urszula Stańczyk[a], Beata Zielosko[b,*]

[a]*Institute of Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland*
[b]*Institute of Computer Science, University of Silesia in Katowice, Będzińska 39, 41-200 Sosnowiec, Poland*

**Abstract**

The paper presents research on unsupervised and supervised discretisation of input data used in execution of stylometric tasks of authorship attribution. Basing on numeric characterisation of writing styles, recognition of authorship is performed by decision rules, as their transparent structure enhances understanding of discovered knowledge. The performance of rule classifiers, constructed in rough set approach, is studied in the context of a strategy employed for resolving conflicts. It is also contrasted with that of other selected inducers.



*Keywords:* discretisation; rough sets; decision rule; conflict; classification; stylometry

## 1. Introduction

Decision rules are often preferred forms of knowledge representation due to their transparent structure that enhances understanding of described patterns by providing explicit premises leading to decisions [2, 17]. A set of rules, obtained through some induction process from the input data [4], can be used to classify new objects as long as some strategy for resolving possible conflicts is adopted.

A conflict occurs when several rules match one example and they do not agree upon the decision—the same premises lead to more than one decision. In such case the final verdict needs to be found by auxiliary procedures [14, 13]. Rejecting all ambiguous decisions is one way of dealing with conflicts. However, the cost of this approach can be prohibitively high, in particular when processing rule sets with high cardinalities. It is possible that for all new samples only ambiguous decisions are made and then their rejection means that no sample is classified.

* Corresponding author.
*E-mail address:* beata.zielosko@us.edu.pl

Another strategy of solving conflicts is to allow for some kind of voting among matching rules, either considering all of them as of the same merit and disregarding their properties, or with taking into account their quality [18]. The former approach is a simple voting, where the majority makes the decision, while the latter involves weighting votes. In the research presented in this paper the effects of these two voting strategies were compared, employed in classification with decision rules inferred in rough set approach.

Rough set data mining is well suited to cases of incomplete and uncertain data [16]. It enables to perceive the universe in granular manner, by grouping objects into equivalence classes constructed through indiscernibility relation. If two objects are characterised by the same values of considered attributes they are indiscernible and belong to the same class. Approximations of classes allow to induce decision rules, which leads to representation of knowledge learned from examples, and classification of previously unseen objects.

Classical rough set approach (CRSA) allows only for nominal classification as it works on discrete value sets. In cases of input data sets with real values, their transformation from continuous into discrete domain becomes the necessary step, with required choice of some discretisation method [8].

Discretisation algorithms are put into several categories depending on the selected focus, and one of possible distinctions relies on the information on recognised classes. When this information has no bearing on the process of building the discretisation model, then it is called *unsupervised*. When class information has some influence on construction of intervals, discretisation is termed *supervised*. In the research described one example from each category of discretisation algorithms was used, namely equal width binning and Kononenko's [11].

The input data sets contained samples providing quantitative characteristics of writing styles in tasks of binary authorship attribution from the stylometry as the application domain [1]. Numerical nature of data leads either to techniques of data mining that are capable of dealing with continuous values [12], or to discretisation to be implemented as a part of initial pre-processing. In the latter approach the performance of obtained rough rule classifiers was confronted with a selected set of other inducers, operating on the same discretised data sets.

The performed experiments show that in most cases weighted voting, used as the strategy for solving conflicts, resulted in increased classification accuracy, as compared with simple voting. For unsupervised discretisation, for relatively small numbers of intervals defined, the power of rule classifiers was comparable to other type of inducers, while for higher numbers of bins they were significantly outperformed. On the other hand, for supervised Kononenko discretisation, which resulted in relatively small numbers of defined intervals, for one of input data sets CRSA classifiers gave recognition at the level of the best, while for the other as the worst, if still acceptable.

The structure of the paper is organised as follows. Section 2 provides theoretical background on rough sets and decision rules. Section 3 describes selected discretisation approaches. Section 4 explains experimental setup, while Section 5 presents results from the experiments. Section 6 contains conclusions and comments on feature research.

## 2. Rough sets

Rough set theory was proposed by Z. Pawlak in 1982 as a way of dealing with inconsistency and incompleteness in data [15]. One of the main notions of this theory is indiscernibility relation defined relative to a given set of attributes. Objects characterised by the same values of attributes are indiscernible (similar) from the point of view of the available knowledge about them. A set of all indiscernible objects is called an elementary set and forms a granule (atom) of knowledge about the universe.

Perception of knowledge through its granular structure causes that any imprecise (rough) concept is replaced by a pair of precise concepts called the lower and the upper approximation of this concept. Imprecision of a concept is expressed by employing a boundary region which is a difference between the upper and the lower approximation of the concept. If the boundary region of a set is non-empty, it means that our knowledge about the set is insufficient to define the set precisely.

### *2.1. Basic notions*

In the rough set theory, the main structure for data representation is an *information system*, and a special case of the information system—*decision table* [16].

An information system is a pair of the form $S = (U, A)$ where $U$ is a nonempty, finite set of objects and $A = \{a_1, \ldots, a_m\}$ is a nonempty, finite set of attributes, i.e., $a_i : U \to V_a$, where $V_a$ is the set of values of attribute $a_i$ called the domain of $a_i$.

A decision table is a pair of the form $S = (U, A \bigcup \{d\})$ with the distinguished attribute $d \notin A$. In case of decision table the attributes belonging to $A$ are called *condition attributes*, while $d$ is called a *decision*. We assume that the set of decision values is finite $V_d = \{d_1, \ldots, d_{|V_d|}\}$. Is it possible to interpret the decision attribute as a classifier on the universe of objects given by an expert. The decision $d$ determines a partition $\{Class_1, \ldots, Class_{|v_d|}\}$ of the universe $U$, where $Class_i = \{x \in U : d(x) = d_i\}$ is called the $i$-th decision class of $S$, for $1 \leq i \leq |V_d|$.

Decision rules are known and popular form of knowledge representation. Their significant advantages are simplicity and ease in being understood and interpreted by humans, which is why decision rules are used in many areas connected with data mining and knowledge discovery. In the paper, decision rules are formulas presented in the form: $(a_{i_1} = v_1) \wedge \ldots \wedge (a_{i_k} = v_k) \to d = v_d$, where $1 \leq i_1 < \ldots < i_k \leq m, v_i \in V_{a_i}$, and $1 \leq v_d \leq |V_d|$.

With the rule some numerical characteristics can be connected [20]. Length of the rule is the number of descriptors (pairs $attribute = value$) in the premise part of rule. Support is the number of training objects that their attribute values satisfy the premise and have the same decision as the one attached to the rule.

There are many decision rule construction methods based on rough set theory [4, 6, 7, 15, 20, 22]. In the research there was used exhaustive algorithm, implemented in Rough Sets Exploration System (RSES) [5]. It constructs all minimal decision rules, i.e., rules with minimal number of descriptors in a premise part, which can be induced from a training decision table.

### *2.2. Decision algorithms and conflicts*

Sets of induced decision rules can be used as decision algorithms and allow to classify new objects. For any given new object $x$ the algorithm attempts to create a decision for this object using only values of condition attributes on $x$. When no rule matches the object, it is not covered, there are no premises to lead to the decision and thus the object cannot be classified.

If only one rule matches the object $x$, the decision is straightforward. If more than one decision rule covers $x$, the case becomes more complex. When all matching rules point to one and only one decision class, then the decision made by the rule classifier is unambiguous. Otherwise, there is a conflict. Only together with some chosen strategy employed for resolving conflicts the set of inferred rules can be treated as a decision algorithm [14, 13].

It is possible to reject all conflicting verdicts and treat samples with such decisions as incorrectly recognised. Yet such attitude, especially in case of high cardinalities of rule sets, can cause that for all examples decisions are ambiguous, and the decision is made for none of new samples.

Conflicts can be resolved by voting. In the RSES system [5], used in the described research, there are two types of voting available: simple and standard. Simple voting means that the decision is chosen by counting votes casted in favor of a certain decision class and each matching rule has one and only one vote. With this approach all rules are treated as if they were of exactly the same quality [18], because all their properties are disregarded at the voting stage.

Standard voting means that not only the number of rules is taken into account, but also the properties of rules are considered in the aspect of their supports. Each rule has as many votes as supporting objects, so in other words the votes of rules are weighted by their supports. Such attitude enables acknowledgment of higher importance of rules supported by more objects.

## 3. Unsupervised vs. supervised discretisation

Discretisation can be considered as a process of simplification of data. Instead of dealing with all subtleties, noting infinite details, the continuous input space is transformed into granular. Each granule corresponds to some category, a specific range partitioned from the continuous domain [8].

Discretisation approaches are grouped into various types, depending on their focus on some data properties. The construction of discretisation model for data, or definitions for recognised categories, can be executed while disregarding information on class recognition in unsupervised approach, or with taking this information into account in supervised procedures [9].

One of unsupervised discretisation algorithms is equal width binning. For each feature its values are analysed and the minimum and maximum found. The resulting range of values is divided into sub-ranges of equal width, with their number provided as an input parameter. With this algorithm the frequencies or distributions of input real values are disregarded and as a result there can be defined such categories that have no representatives in the original data sets, which is often considered a drawback of this method. On the other hand this approach results in construction of data model that most closely resembles the original space, it is only sufficiently simplified by change of recognised scale, uniform reduction of accuracy of description provided for objects. In case of discretisation of several separate data sets, as minimum and maximum values can and most likely do vary, there can be some differences in obtained definitions of categories, but their number will remain as required by the same input parameters.

Kononenko method belongs with supervised discretisation [11]. The process of constructing definitions of categories starts with assigning a single interval for the whole range of encountered values. Then there are considered possible candidates for cut-points to be used in splitting the range into smaller sub-ranges. The algorithm is executed recursively until the stopping criterion, based on *Minimum Description Length* (MDL) principle, is met. It is possible that all candidate cut-points are rejected and then the initial single interval defined for the whole range of values of some variable will remain undivided [19].

Since such procedures rely on numbers of samples, numbers of instances for recognised classes, and values of attributes in the context of decisions, calculations have highly local context. As a consequence, the same process applied to separate data sets will most likely result in obtaining different data models—different definitions of ranges, even different numbers of these ranges.

## 4. Experimental setup

Experiments performed in the research presented in this paper consisted of stages, as described below:

i) construction of input data sets with continuous valued features,
ii) discretisation of all input data sets through selected methods,
iii) induction of decision rules for all versions of training decision tables,
iv) classification of test samples by rule classifiers with simple and standard voting strategies in case of encountered conflicts,
v) training the set of selected inducers and testing them,
vi) comparison and analysis of obtained test results.

### *4.1. Analysis of texts with respect to style*

A style is an elusive phenomenon, difficult to be expressed by precise definitions, yet detectable as long as there are sufficiently high numbers of examples of undisputed authorship by the same creator [1]. In textual analysis it means that access to several texts is required, as many as possible. Then longer works are divided into smaller parts of comparable size to construct representative text samples, over which stylometric descriptors are to be calculated.

The writing style should be recognised regardless of the subject content of some text, thus instead of looking for keywords or phrases, rather some lexical and syntactic properties are often analysed, such as frequencies of usage of function words and punctuation marks [12]. Discriminative properties of such characteristic features are sufficient for recognition of authorship executed with the help of some modern machine learning approaches [10].

Two data sets were employed in the research, one for recognition between two male writers, and the second for a pair of female writers. This grouping of authors of the same sex into one data set was dictated by the fact that their writing styles share some of linguistic characteristics, which would falsify the recognition in case of authors of opposite gender, as then attribution is simpler.

In the preliminary preprocessing step, frequencies of usage of a hundred function words were calculated, the elements selected from the list of most often used words in English language. Next, for the obtained sets there were applied several ranking mechanisms, implemented in WEKA workbench [21]. These ranking scores were then used to select a subset of such features that were never considered as irrelevant (never received the score of zero).

The described processing led to obtaining sets with 24 stylometric features, comprised of 2 syntactic, and 22 lexical markers. The syntactic descriptors reflected frequencies of usage for semi-colon and comma, lexical for the following words: after, almost, any, around, before, but, by, during, how, never, on, same, such, that, then, there, though, until, what, whether, who, within. Due to their character the range of values for all considered attributes was $< 0, 1)$.

### *4.2. Discretisation of input data sets*

Selected discretisation algorithms were employed independently to all constructed data sets with continuous features, unsupervised equal width binning with varying the input parameter of the numbers of required intervals, and supervised Kononenko.

For equal width binning 36 variants of each set were obtained: with a step of 1 from 2 to 10, with a step of 10 from 10 to 100, with a step of 100 from 100 to 1000, and with a step of 1000 from 1000 to 10000. The supervised Kononenko method is non-parameteric, thus resulted in the single version for each discretised data set.

The discretised data can take the form of representation of values as ranges to which the values belong to, but such representation is highly inconvenient for further processing. Instead all constructed intervals were simply enumerated and this integer numbers taken as nominal attribute values defining categories.

Independent processing of sets in case of unsupervised discretisation can cause differences in definitions of cut-points. In case of supervised discretisation, models of data obtained for various sets can differ not only in borders of intervals, but their established numbers.

### *4.3. Decision algorithms and performance evaluation*

The sets of rules used in the research were induced with the help of Rough Set Exploration System (RSES) for each variant of the discretised input training sets. In the initial steps the rules were inferred with Lem2 algorithm, but the resulting rule sets were rather small, provided very low coverage, and thus unacceptably low performance. Therefore for the main experiments exhaustive algorithms were generated.

As training sets were constructed from groups of samples originating from the same source texts, some objects showed higher similarity than others. As a consequence, using popular cross-validation for evaluation of classifiers performance would result in over-optimistic results [3]. Instead this performance was evaluated with test sets, which were build from samples based on separate texts than those used in the training step.

Using separate test sets caused the necessity of their discretisation, which was also performed independently on learning sets. In this way the pre-processing of input data sets was much simplified, but at the cost of making the recognition more complex and harder, as the original input space was then perceived through two different discretisation models, the one obtained from training, and the second calculated over test data.

Decision algorithms tested were obtained while inducing all rules on examples, which meant that the cardinalities were relatively high. In such situations it is possible to filter out some of the decision rules in search of such subset that includes only rules of the highest quality. Filtering can be driven directly by some rule parameters or other defined measures [19, 18]. Yet selection of rules requires additional processing time and studies on imposed constraints. The results presented in the paper were obtained for the entire sets of decision rules.

### *4.4. Classification systems used for comparison*

The performance of rule classifiers was contrasted with the set of four other inducers, often used in comparisons, all available in WEKA environment [21]. The group included Naive Bayes (denoted as Bayes), k-Nearest Neighbour (kNN), Radial Basis Function network (RBF), and PART.

Naive Bayes is a statistical classifier, quite powerful if relatively simple. It relies on the rule of conditional entropy, for calculation of which independence of considered attributes is assumed. kNN is an instance-based learner, with the number of considered neighbours treated as an input parameter. The decision about the class is based on the majority of decisions among studied neighbours. RBF is a type of an artificial neural network in which radial basis function is used as an activation function for neurons. PART constitutes a variant of C4.5 decision tree learner.

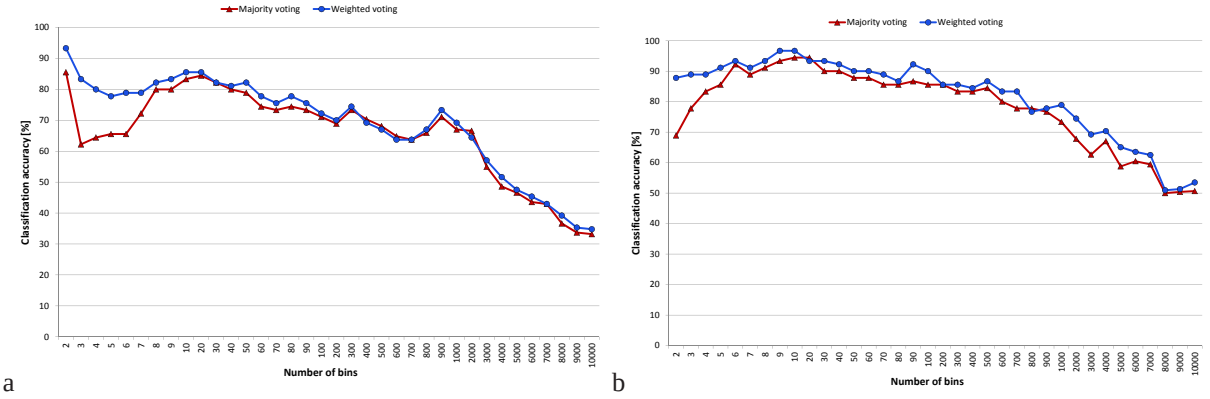
All these inducers were employed while using only their default parameters, without any fine tuning, which could lead to enhanced recognition. In case of neural network the training was executed multiple times and the results presented in the next section correspond to the calculated averaged performance.

## 5. Results from experiments

Experiments performed in the described research were divided into two parts. In the first part there were compared results of two voting strategies applied for solving conflicts occurring in classification by rule sets. The decision algorithms were induced from the learning sets and evaluated with test sets, the pairs obtained from unsupervised and supervised discretisation. In the second group of tests the other four inducers were trained and then tested while working on the same discretised data sets that were previously used for rule classifiers. The details for both processes are commented below.

### *5.1. Simple majority vs. weighted voting*

For both data sets used in research there were obtained 36 variants from unsupervised discretisation with equal width binning, and one from Kononenko's supervised discretisation approach. For all these learning sets the inferred sets of rules together with two voting strategies were next employed for classification of new samples from the corresponding test sets. The classification accuracies obtained in the whole range of numbers of intervals are shown in Fig. 1a) for male writer data set, and in Fig. 1b) for female writers.



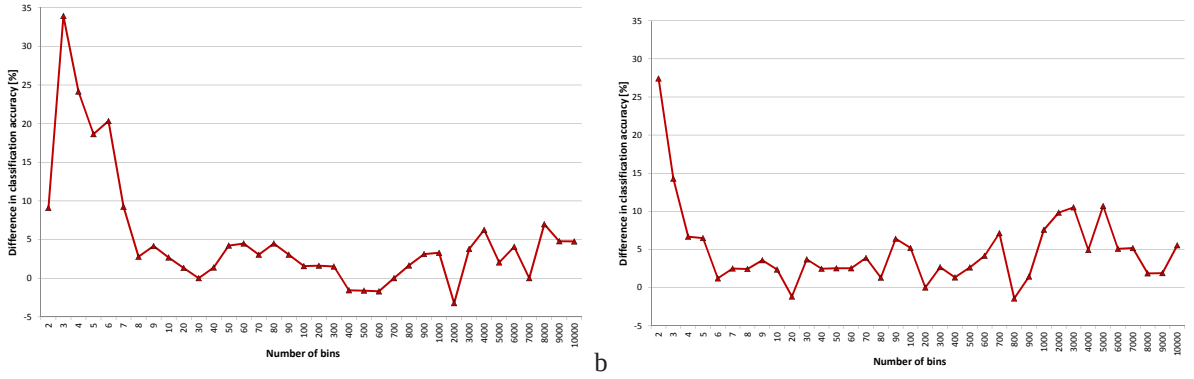

a b

Figure 1. Performance of rule classifiers induced for input data sets obtained from unsupervised discretisation with equal width binning for: (a) male writers, (b) female writers.

Overall the charts allow to conclude that male writer data set provides a more difficult case for classification, as the detected level of correct predictions was lower than for the female writers. Furthermore, for both data sets the performance for the most part decreases with the increase of the numbers of constructed bins. The exception to this observation lies in the beginning parts of charts, for relatively few intervals considered, in particular for female writer data set. The maximum performance was achieved for male writers for just two bins where the recognition was 93.33%, and for female for either 9 or 10 with 96.67%.

Both plots included in Fig. 1 clearly indicate that application of weighted voting, that is assigning to each matching rule as many votes as it had supporting instances in the learning data, as opposed to the single votes assigned to all rules, regardless of their parameters, causes outperforming of the latter strategy in most cases. The differences in obtained results can be observed in Fig. 2.

The plots show the difference calculated by subtracting the classification accuracy for simple majority voting from the results obtained for weighted voting, and divided by the former. The result is given then as percentage. In the beginning the difference is highly in favour of weighted voting, then with the increasing numbers of bins it decreases, in few cases reaching zero, or even below zero, which then denotes cases where majority voting caused outperforming of weighted voting.




a b

Figure 2. The percentage difference observed in performance for the two voting strategies employed for rule classifiers induced from data sets obtained by unsupervised discretisation with equal width binning for: (a) male writers, (b) female writers.

These results to some extent can be explained by characteristics of the inferred rule sets, shown in Table 1. With the increasing numbers of bins the cardinalities of the rule sets increase steeply to reach some maximum and then slowly decrease. The minimal support of rules was always equal 1, which is why it is omitted. On the other hand, the average support values decrease from the initial maximum till an almost constant level close to 1 is reached. These two elements indicate that for just a few intervals defined there is a significant portion of rules with high support values and naturally their votes have high influence over the voted decision which accounts for the significant difference for the two voting strategies. When the average support is lower, few rules with higher support can be outvoted by many more rules with lower supports and the difference in the voting strategies gets smaller and smaller, close to zero.

Kononenko's supervised discretisation, as non-paramateric, returned a single variant for each input data set, in which all attributes had a relatively few intervals defined, from the minimum of 1 to the maximum of 3 for male writers, and from 1 to 4 for female writers. Rule sets inferred from these training sets together with two voting strategies implemented, resulted in the performance shown in Table 2. For female writer data set the classification accuracy for weighted voting was equal to the highest level previously detected for unsupervised discretisation, while for male writers it was lower than the previous maximum.

Also in this case weighted voting gave better correct prediction results than simple voting. When statistics of rules sets are consulted, provided in Table 3, the previously made observations are confirmed—the differences between two conflict resolving strategies are noticeable as the average supports indicate the presence of significantly large groups of decision rules with higher support values.

When rule sets are generated through exhaustive algorithms, typically rules supported by single instances in the training decision table constitute a large percentage. If such rules with low supports were filtered out by imposing hard constraints on the minimal support required of rules to be included in the set taken for classification of new samples, it would greatly influence the possible difference between two discussed voting methods.

### *5.2. Comparison with other inducers*

For the same variants of discretised input sets there were next employed other selected inducers, often used in comparisons. The results of classification for unsupervised discretisation are shown in Fig. 3, and for supervised Kononenko's approach in Table 4. In few cases the correct recognition was slightly better then the previously found maxima for rule classifiers, however, these inducers were more constant in their power considered in the context of the number of bins defined for attributes. The recognition was not decreased with the increase of the number of intervals.

For equal width binning the maximum classification accuracy of 92.22% for Bayes was detected for 2 bins for male writer data set, and 97,78% for either 30, 60, or 2000 bins for female writers. For kNN the maximum was of 94.44% for male writers for either 5, 8, or 90 intervals, and 97.78% for female with 8, or 9 bins. The best results of 91.17% for RBF were obtained for 100, 3000, or 10000 bins for male writer data set, and 96.17% for female for 10 bins. And finally for PART the maximum was 92.22% for 30, 60, 200, 300, 500 or 600 bins in male writer data sets, and 96.67%

Table 1. Parameters of rule sets obtained for training data discretised by unsupervised approach

| | Data set | | | | | |
|---|---|---|---|---|---|---|
| | Male writers | | | Female writers | | |
| Number of bins | Number of rules | Maximal support | Average support | Number of rules | Maximal support | Average support |
| 2 | 1509 | 78 | 7.5 | 2094 | 86 | 6.5 |
| 3 | 32447 | 66 | 3.3 | 26025 | 75 | 3.6 |
| 4 | 47574 | 49 | 2.9 | 46480 | 58 | 2.6 |
| 5 | 79561 | 51 | 2.3 | 67054 | 54 | 2.1 |
| 6 | 77033 | 51 | 2.0 | 75888 | 44 | 2.0 |
| 7 | 75733 | 43 | 1.9 | 70152 | 42 | 1.8 |
| 8 | 72675 | 33 | 1.8 | 60422 | 39 | 1.8 |
| 9 | 68722 | 31 | 1.7 | 59332 | 33 | 1.7 |
| 10 | 61920 | 30 | 1.6 | 54187 | 32 | 1.6 |
| 20 | 46394 | 22 | 1.5 | 40232 | 32 | 1.5 |
| 30 | 41721 | 21 | 1.5 | 35837 | 32 | 1.5 |
| 40 | 39756 | 21 | 1.5 | 34304 | 32 | 1.4 |
| 50 | 38129 | 21 | 1.5 | 34825 | 32 | 1.4 |
| 60 | 36075 | 21 | 1.5 | 32786 | 32 | 1.4 |
| 70 | 34878 | 21 | 1.5 | 32474 | 32 | 1.4 |
| 80 | 34023 | 21 | 1.5 | 31145 | 32 | 1.4 |
| 90 | 33957 | 21 | 1.5 | 29734 | 32 | 1.4 |
| 100 | 32435 | 21 | 1.5 | 30944 | 32 | 1.4 |
| 200 | 26467 | 21 | 1.5 | 25227 | 32 | 1.4 |
| 300 | 22780 | 21 | 1.4 | 21938 | 32 | 1.4 |
| 400 | 19953 | 21 | 1.4 | 19518 | 32 | 1.4 |
| 500 | 17782 | 21 | 1.4 | 18138 | 32 | 1.3 |
| 600 | 17078 | 21 | 1.4 | 16700 | 32 | 1.3 |
| 700 | 15460 | 21 | 1.3 | 15626 | 32 | 1.3 |
| 800 | 14856 | 21 | 1.3 | 14746 | 32 | 1.3 |
| 900 | 13966 | 21 | 1.3 | 14300 | 32 | 1.3 |
| 1000 | 12956 | 21 | 1.3 | 13482 | 32 | 1.3 |
| 2000 | 9777 | 21 | 1.3 | 10118 | 32 | 1.3 |
| 3000 | 8308 | 21 | 1.2 | 8284 | 32 | 1.3 |
| 4000 | 7439 | 21 | 1.2 | 7551 | 32 | 1.2 |
| 5000 | 7302 | 21 | 1.2 | 7236 | 32 | 1.2 |
| 6000 | 6966 | 21 | 1.2 | 6949 | 32 | 1.2 |
| 7000 | 6941 | 21 | 1.2 | 6873 | 32 | 1.2 |
| 8000 | 6826 | 21 | 1.2 | 6808 | 32 | 1.2 |
| 9000 | 6833 | 21 | 1.2 | 6736 | 32 | 1.2 |
| 10000 | 6776 | 21 | 1.2 | 6719 | 32 | 1.2 |

Table 2. Performance of rule classifiers for input data sets obtained from supervised discretisation [%]

| | Voting strategy | | |
|---|---|---|---|
| Data set | Majority | Weighted | Percentage difference |
| Male writers | 72.22 | 86.67 | 20.01 |
| Female writers | 91.11 | 96.67 | 6.10 |

Table 3. Parameters of rule sets obtained for training data discretised by supervised approach

| Data set | Number of rules | Maximal support | Average support |
|---|---|---|---|
| Male writers | 20815 | 75 | 5.5 |
| Female writers | 10190 | 88 | 5.4 |

for 20 intervals in female writer data set. Comparison of these results brought the conclusion that for male writers kNN provided the best classification accuracy, and for female writers it was either Bayes or once again kNN.

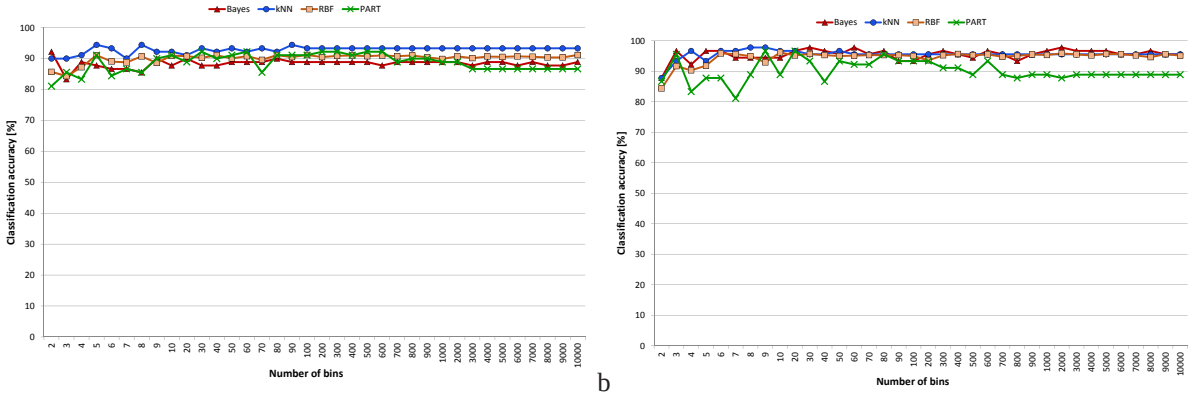



Figure 3. Performance of selected classifiers for input data obtained from unsupervised discretisation with equal width binning for: (a) male writers, (b) female writers.

For supervised discretisation Bayes classifier was the undisputed champion for both data sets tested, with RBF closely behind for male writer data set (however, it was the worst for female writers), and kNN slightly degraded for female writers. These good predictive properties of Bayes classifier can be explained to some extent by similarity of the elements studied by Kononenko's supervised discretisation approach to the Bayes rule of conditional entropy.

Table 4. Performance of selected classifiers for input data sets obtained from supervised discretisation [%]

| | Classifier | | | |
|---|---|---|---|---|
| Data set | Bayes | kNN | RBF | PART |
| Male writers | 96.67 | 90.00 | 95.67 | 88.89 |
| Female writers | 96.67 | 94.44 | 76.78 | 91.11 |

However, none of the studied inducers outperformed rule classifiers in a significant degree. The knowledge mined with their help is hidden in the internal structures and is not so easily accessible as in induced decision rules. As performance is at the comparable level and available rule sets also enhance understanding of detected patterns, it is understandable that they can be preferred over other learners. On the other hand, the ways of optimisation for rule classifiers could be studied, leading to rejection of weaker rules of lower quality, as it would possibly improve not only the structure by dimensionality reduction but also recognition.

## 6. Concluding remarks

The paper presents research on performance of rule classifiers induced from data discretised by unsupervised and supervised approaches. The power of inducers was studied in the context of two strategies applied to solving encountered conflicts, when matching rules point to different decisions. In simple voting each rule has a single vote, while in weighted voting each rule has as many votes as many instances support the rule.

The decision rules were induced by exhaustive algorithm in classical rough set processing, which perceives the universe through granules of objects that cannot be discerned, basing on values of the considered attributes. Rough set theory is often applied to cases of incomplete and uncertain knowledge, and provides mechanisms for data mining, allowing for approximations of recognised concepts.

The power of rule classifiers was contrasted with the set of chosen other learners, often used in comparisons. None of them showed noticeably better predictive accuracies, while they lacked the transparency of representation of knowledge discovered in the learning process of decision rules.

All inducers were employed in the problem of authorship attribution that belongs with stylometry as the application domain. Stylometric characteristic features provide descriptions of writing styles by numeric attributes, enabling to treat the recognition of authorship as a classification task.

In the future research other discretisation approaches to data will be investigated, along with other ways of obtaining discretised test sets, different from independent processing that was used in the tests shown in this paper.

## Acknowledgments

The research described was performed within the project BK/RAu2/2019 at the Silesian University of Technology, Gliwice, and at the University of Silesia in Katowice, Sosnowiec, within the project – Methods of artificial intelligence in information systems. For data processing RSES system was used (logic.mimuw.edu.pl), and WEKA workbench [21]. Texts exploited in experiments are available thanks to Project Gutenberg (www.gutenberg.org).

## References

[1] Argamon, S., Burns, K., Dubnov, S. (Eds.), 2010. The structure of style: Algorithmic approaches to understanding manner and meaning. Springer, Berlin.

[2] Azad, M., Zielosko, B., Moshkov, M., Chikalov, I., 2013. Decision rules, trees and tests for tables with many-valued decisions-comparative study, in: Watada, J., Jain, L.C., Howlett, R.J., Mukai, N., Asakura, K. (Eds.), Proceedings of the 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, KES 2013. Elsevier. volume 22 of *Procedia Computer Science*, pp. 87–94.

[3] Baron, G., 2016. Comparison of cross-validation and test sets approaches to evaluation of classifiers in authorship attribution domain, in: Czachórski, T., Gelenbe, E., Grochla, K., Lent, R. (Eds.), Proceedings of the 31st International Symposium on Computer and Information Sciences. Springer, Cracow. volume 659 of *Communications in Computer and Information Science*, pp. 81–89.

[4] Bazan, J., Nguyen, H., Nguyen, S., Synak, P., Wróblewski, J., 2000. Rough set algorithms in classification problem, in: Polkowski, L., Tsumoto, S., Lin, T. (Eds.), Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems. Physica-Verlag HD, Heidelberg, pp. 49–88.

[5] Bazan, J., Szczuka, M., 2005. The rough set exploration system, in: Peters, J.F., Skowron, A. (Eds.), Transactions on Rough Sets III. Springer, Berlin, Heidelberg. volume 3400 of *Lecture Notes in Computer Science*, pp. 37–56.

[6] Błaszczyński, J., Słowiński, R., Szeląg, M., 2011. Sequential covering rule induction algorithm for variable consistency rough set approaches. Inf. Sci. 181, 987–1002.

[7] Chikalov, I., Lozin, V., Lozina, I., Moshkov, M., Nguyen, H., Skowron, A., Zielosko, B., 2013. Three Approaches to Data Analysis —- Test Theory, Rough Sets and Logical Analysis of Data. volume 41 of *Intelligent Systems Reference Library*. Springer, Berlin, Heidelberg.

[8] Garcia, S., Luengo, J., Saez, J., Lopez, V., Herrera, F., 2013. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering 25, 734–750.

[9] Huan, L., Farhad, H., Lim, T., Manoranjan, D., 2002. Discretization: An enabling technique. Data Mining and Knowledge Discovery 6, 393–423.

[10] Jockers, M., Witten, D., 2010. A comparative study of machine learning methods for authorship attribution. Literary and Linguistic Computing 25, 215–223.

[11] Kononenko, I., 1995. On biases in estimating multi-valued attributes, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence IJCAI'95, Morgan Kaufmann Publishers Inc.. pp. 1034–1040.

[12] Koppel, M., Schler, J., Argamon, S., 2009. Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology 60, 9–26.

[13] Lindgren, T., 2004. Methods for rule conflict resolution, in: Boulicaut, J., Esposito, F., Giannotti, F., Pedreschi, D. (Eds.), Machine Learning: ECML 2004. Springer, Berlin Heidelberg. volume 3201 of *Lecture Notes in Computer Sicence*, pp. 262–273.

[14] Lindgren, T., Boström, H., 2004. Resolving rule conflicts with double induction. Intelligent Data Analysis 8, 457–468.

[15] Pawlak, Z., Skowron, A., 2007a. Rough sets and boolean reasoning. Information Sciences 177, 41–73.

[16] Pawlak, Z., Skowron, A., 2007b. Rudiments of rough sets. Information Sciences 177, 3–27.

[17] Sikora, M., Wróbel, L., Gudyś, A., 2019. Guider: A guided separate-and-conquer rule learning in classification, regression, and survival settings. Knowledge-Based Systems 173, 1 – 14.

[18] Stańczyk, U., 2011. Reduct-based analysis of decision algorithms: Application in computational stylistics, in: Grana Romay, M., Corchado, E., Garcia-Sebastian, M. (Eds.), Hybrid Artificial Intelligence Systems. Part 1. Springer, Berlin. volume 6679 of *LNCS (LNAI)*, pp. 295–302.

[19] Stańczyk, U., Zielosko, B., 2017. On combining discretisation parameters and attribute ranking for selection of decision rules, in: Polkowski, L., Yao, Y., Artiemjew, P., Ciucci, D., Liu, D., Ślęzak, D., Zielosko, B. (Eds.), Proceedings of the International Joint Conference on Rough Sets, IJCRS 2017. Part I, Springer, Olsztyn, Poland. pp. 329–349.

[20] Stańczyk, U., Zielosko, B., Żabiński, K., 2018. Application of greedy heuristics for feature characterisation and selection: A case study in stylometric domain, in: Nguyen, H., Ha, Q., Li, T., Przybyla-Kasperek, M. (Eds.), Proceedings of the International Joint Conference on Rough Sets, IJCRS 2018, Springer, Quy Nhon, Vietnam. pp. 350–362.

[21] Witten, I., Frank, E., Hall, M., 2011. Data Mining. Practical Machine Learning Tools and Techniques. 3rd ed., Morgan Kaufmann.

[22] Zielosko, B., 2016. Application of dynamic programming approach to optimization of association rules relative to coverage and length. Fundamenta Informaticae 148, 87–105.