Title: Heuristic-based feature selection for rough set approach

Author: Urszula Stańczyk, Beata Zielosko

Citation style: Stańczyk Urszula, Zielosko Beata. (2020). Heuristic-based feature selection for rough set approach. "International Journal of Approximate Reasoning" (Vol. 125 (2020), s. 187-202), DOI:10.1016/j.ijar.2020.07.005

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Heuristic-based feature selection for rough set approach

U. Stańczyk [a,*], B. Zielosko [b]

[a] Department of Graphics, Computer Vision and Digital Systems, Faculty of Automatic Control, Electronics and Computer Science, 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

ARTICLE INFO

Article history:
Received 1 July 2019
Received in revised form 13 February 2020
Accepted 13 July 2020
Available online 4 August 2020

Keywords:
Rough sets
Greedy heuristics
Decision rules
Feature selection
Discretisation
Stylometry

ABSTRACT

The paper presents the proposed research methodology, dedicated to the application of greedy heuristics as a way of gathering information about available features. Discovered knowledge, represented in the form of generated decision rules, was employed to support feature selection and reduction process for induction of decision rules with classical rough set approach. Observations were executed over input data sets discretised by several methods. Experimental results show that elimination of less relevant attributes through the proposed methodology led to inferring rule sets with reduced cardinalities, while maintaining rule quality necessary for satisfactory classification.

1. Introduction

In rough set perspective [1,2] the universe is seen as granular, with data points grouped in space into equivalence classes, imposed by indiscernibility relation. When two objects have the same values of considered features, they cannot be discerned. Detection of patterns in granules leads to inferring decision rules. Based on conditions on attributes included in the premises, the rules assign class labels to examples. A length of a rule, corresponding to the number of its conditions, is one of important indicators of rule quality [3,4]. Among other quality measures there is also used support, which gives the number of learning samples that match the rule. Length and support are frequently taken under consideration in a search for interesting rules.

Depending on the selected focus, there are many rule induction algorithms: Boolean reasoning [5,6], dynamic programming [7–9], separate-and-conquer approach [10–12], algorithms based on decision tree construction [13,14], genetic algorithms [15,16], different kinds of greedy algorithms [5,17], and various others. Each of these methods has different forms, which return rule sets with varying cardinalities and characteristics. In exhaustive search all rules on examples are inferred, minimal cover algorithm ensures only coverage of the training samples and then stops the search, greedy heuristics obtain optimal solutions in the local context. Once decision rules are inferred, the knowledge discovered in the process of their construction becomes relatively easily accessible. This knowledge, represented in the form of rules, can be used by itself to enhance understanding of patterns present in the input space, or to perform classification of unknown samples, but it can also aid in other tasks [18].

[*] Corresponding author.
E-mail address: urszula.stanczyk@polsl.pl (U. Stańczyk).

https://doi.org/10.1016/j.ijar.2020.07.005

The paper presents a research methodology that proposes to employ decision rules, induced by greedy heuristics, in construction of feature rankings [19,20], with the score function dependent on rule characteristics. The obtained rankings can be next applied in the process of feature selection and reduction.

In the research works described from the initial choice of seven algorithms, by analysis of Pareto points in the optimisation space, three heuristics were selected and four discarded. The choice was driven by observations of two rule quality indicators, namely length and support. Exploiting knowledge on attributes, stored in rules induced by the selected algorithms, all available characteristic features were ranked.

Then, the ranking was used for reduction of attributes that were considered for induction of decision rules by rough set approach. The elimination process resulted in generation of rule sets with lowered cardinalities, while keeping satisfactory level of average rule length and support, when compared to these parameters for the whole set of features. Obtained rough rule classifiers were applied in the task of authorship attribution from stylometric analysis of texts, in which authors are recognised through their writing styles [21,22].

Since stylometric characteristic features are most often continuous in nature, and both heuristics and rough set approach operate on nominal attributes, discretisation was added as a necessary step of input data preparation stage [23]. As there are various methods for transformation of real-valued into discrete variables, the experiments were extended to include observations on both supervised and unsupervised procedures.

The content of the paper is organised into six sections. Introduction is followed by Section 2 addressing all background subjects included in the described research, such as feature selection, decision rule induction with rough set approach and heuristics, stylometry as the application domain, and discretisation. The framework of executed experiments is presented in broad strokes in Section 3, with details for all constituent stages and results commented in Section 4 and Section 5. Conclusions from research works are included in Section 6.

2. Background and related works

This section presents the fields of study included in the described research and some related works dedicated to subjects of feature selection and characterisation, induction of decision rules, discretisation, and stylometry as the application domain.

2.1. Feature selection and characterisation

In recent years, a huge increase of data stored, transmitted, and processed could be observed. As a consequence, feature selection and reduction domain plays an important role in knowledge discovery and different data mining tasks [19,20], especially in areas where data sets contain a huge number of attributes, for example, sequence-pattern in bioinformatics, genes expression analysis, market basket analysis, stock trading. The main underlying objectives of variable selection are better understanding of data and improving the prediction performance of classifiers. From the point of view of knowledge induction, some attributes can be insufficient or redundant, so the problem is how to select the relevant features that allow to obtain knowledge stored in data [24].

There are different approaches and algorithms for features selection [25,26]. Usually they are divided into three categories: filter, wrapper, and embedded methods, however, algorithms can be also mixed together in different variations, leading to hybrid solutions [27].

Filter methods are independent from classification systems. Filters pre-process data sets without any feedback information concerning improvement or degradation of classification results. They tend to be faster, less resource demanding, and more universal than other approaches. Their main drawback is what makes them fast and easily applicable in almost all kinds of problems, i.e., neglecting the real-time influence on a classification system.

Wrapper methods can be interpreted as a system with a feedback [28]. This category of algorithms is based on the idea of examining the influence of the chosen subsets of features on the classification results. Wrapper approach typically requires large computational costs as the classification step needs to be repeated many times. On the other hand, wrappers can obtain close tailoring of feature sets to inducers, which leads to significantly enhanced predictions.

The last category of methods is known as embedded solutions [29]. Generally, they consist of mechanisms that are embedded directly into the learning algorithm and they are responsible for the feature selection process at the learning stage. An advantage of embedded methods is good performance as the solutions are dedicated to specific applications. Nevertheless, they cannot be used without knowing the learning algorithm characteristics.

Ranking is a filter mechanism that assigns each variable a certain score, basing on which features become ordered [30]. The highest ranking attributes are considered as the most important, and the lowest ranking as the least. Ranking mechanisms often refer to statistics and calculate for example entropy, mutual information, information gain. The scores depend on a definition of a ranking function, and the ranking procedure can return just the resulting order of variables, their assigned ranking positions. Such could be a case of using a wrapper as a ranker.

2.2. Rough set-based feature selection

Rough set theory (RST) was proposed by Z. Pawlak in 1982 as a way of dealing with inconsistent data [1]. The knowledge is perceived through its granular structure, i.e., some objects of the universe are indiscernible relative to a given set of attributes. Such set of all indiscernible objects forms a granule of knowledge about the universe. Granularity of knowledge causes that rough (imprecise) concepts cannot be characterised in the framework of knowledge available about their elements, and such concept is replaced by a pair of precise sets called the lower and the upper approximation of the rough concept.

Rough set methods dedicated to feature selection are mainly based on algorithms for construction of reducts, and their different modifications. From the classification point of view a reduct is a minimal subset of attributes that has the same classification power as the entire set of condition attributes. It can be also defined as a minimal set of attributes that preserves the degree of dependency of the full set of attributes. Definitions for attribute reducts can be based on different criteria [31,32], and the problem of finding various versions of reducts in data is NP-hard [17], which is the reason why heuristic approaches are often employed.

For inconsistent decision tables a unified decision table model was proposed for five representative reducts [33], along with two general heuristic algorithms for attribute reduction, based on relative discernibility measure: quick general forward and backward elimination. The efficiency of the proposed algorithms was obtained mainly by reducing the time spent on sorting.

For large-scale data sets a way for approximate reduct construction was presented [34], with the main idea based on subtables of a data set, which were considered as small granularities. Fusing together all estimated reducts on small granularities allows to obtain an approximate reduct of the original data set.

Some of feature selection methods use positive region-based dependency measure for attributes to establish how uniquely the value of an attribute determines the value of a dependent variable. The measure ranges from zero (which means no dependency of an attribute) to one (which means that an attribute fully depends on the other). However, such approach is time consuming and complex, which makes it unfeasible in case of data sets with bigger size.

An alternative to the conventional positive region-based dependency measure, called Direct Dependency Calculation, finds the number of unique and non-unique classes directly by using attribute values [35]. A unique dependency class defines a set of objects, which for the same values of condition attributes all lead to the same decision class, while a non-unique decision class is a set of objects where all lead to more than one decision class, for the same values of condition attributes.

In recent years, incremental feature selection based on rough set approach has become more and more popular. In this framework, two main tracks can be distinguished: based on discernibility matrix [36,37], and focused on entropy [38]. Fused decision tables [39] are constructed by integration of several similar decision tables, where one among them is the original, and others are added successively. These decision tables share the same attributes and have their own objects. The proposed incremental selection method is based on using quasi- and pseudo fuzzy rough approximation operators, which optimise the space constraint of storing discernibility matrix and accelerate calculations.

Still other methods are based on knowledge granulation approaches [40,41], possibly combining variation of attributes with incremental attribute reduction [42]. Roughinement operation offers a specific way of granulation for features, and then aggregation of information granules belonging to different partitions [43]. The proposed routine allows to obtain a compact representation of data, reduces the cost for evaluation of the quality of selected subset of features, while maintaining good reduction capability.

Algorithms based on neighbourhood rough sets aim to distinguish samples that belong to different decisions using neighbourhood information granules [44]. The structure of clusters can be adjusted dynamically [45] during the clustering process, even when a new set of attributes feeds the algorithm.

There are also many other heuristics based on rough set theory developed for feature selection, for example based on genetic algorithms [16], optimisation of particle swarm [46] or ant colony [47], and others [48].

Greedy approaches are frequently applied to the task of distinguishing the most important features within the entire set of available attributes. Usually such heuristics start with an empty set of features and then adopt either forward selection or backward elimination algorithm in case of full set of features. However, greedy methods do not guarantee finding an optimal or minimal feature combination, also because of the fact that many significance measures exist.

In the paper a new methodology was proposed, dedicated to employing selected greedy heuristics, typically used for optimisation of association rules, in feature selection process. Heuristics were adjusted for work with decision rules, and the step of reduct construction was omitted. Knowledge discovered based on characteristics of induced rules was used for establishing a ranking of features, and a score function was proposed.

In the research works reported, the ranking constructed through greedy heuristics was employed as a filter mechanism. For sets of decision rules, found by exhaustive algorithm implemented in Rough Sets Exploration System (RSES) [49,50], classification results as well as the number and characteristics of inferred rules were examined.

Usefulness of the presented methodology was shown for data sets and tasks from stylometry domain. Experimental results indicate that elimination of less relevant attributes through the proposed framework led to inferring rule sets with reduced cardinalities, while maintaining rule quality necessary for satisfactory classification.

2.3. Induction of decision rules with greedy heuristics

In search for decision rules there are considered various quality measures, probably the most popular of which are length and support, analysed also in the research described in the paper. Unfortunately, the problems of minimisation of length and maximisation of support of decision rules are NP-hard [51,52]. With the exception of brute-force, Boolean reasoning, and extensions of dynamic programming, for the most part the approaches cannot guarantee the construction of optimal rules (i.e., rules with minimum length or maximum support).

Based on the results of U. Feige [53], a greedy algorithm was shown to be close to the best polynomial approximate algorithms for minimisation of decision rule length, under reasonable assumptions on the class NP [54]. Greedy heuristics for construction of association rules were also studied and compared from the point of view of length and support of obtained rules [55]. It was reported how on average the output of each greedy algorithm is close to optimal rules obtained by extensions of dynamic programming approach. The experimental results showed that the average relative difference between length of rules constructed by the best heuristic and minimum length of rules is at most 4%. Interestingly, the same situation was obtained for support.

In this work, a methodology of research and an application of seven greedy heuristics in feature selection domain was proposed.

2.3.1. Main notions

In rough set theory, the main structure for data representation is an *information system*, and a special case of information system—a *decision table* [2].

Information system is a pair of the form $S = (U, A)$, where U is a nonempty finite set of objects, and $A = \{f_1, \ldots, f_{n+1}\}$ is a nonempty finite set of attributes, i.e., $f : U \to V_f$, where V_f is the set of values of attribute f, called the domain of f. Decision table is a pair of the form $S = (U, A \bigcup \{d\})$, with a decision attribute $d \notin A$, and a is a value of the decision attribute (also called a decision), $a \in V_d$, where V_d is the domain of d. In the case of a decision table the attributes belonging to $A = \{f_1, \ldots, f_n\}$ are called *condition attributes*.

The expression

$$(f_{i_1} = a_1) \wedge \ldots \wedge (f_{i_m} = a_m) \to d = a \tag{1}$$

is called a *decision rule over* T if $f_{i_1}, \ldots, f_{i_m} \in \{f_1, \ldots, f_n\}$, $a_1, \ldots, a_m$ are values of corresponding attributes, and a is a decision.

Let $T = (U, A \cup \{d\})$ be a decision table. $N(T)$ denotes the number of rows in the table T. $N(T, a)$ gives the number of rows r from T with a value of a decision attribute equal a, with $M(T, a) = N(T) - N(T, a)$. A decision a, such that $N(T, a)$ has maximum value and a has minimum index, is called the *most common decision for* T, and denoted by $mcd(T)$. Not constant condition attributes in T form the set denoted as $E(T)$.

A *subtable* of a decision table T is obtained by removal of some rows from T. A subtable of T that consists of rows, which at the intersection with columns $f_{i_1}, \ldots, f_{i_m}$ have values $a_1, \ldots, a_m$, is denoted as $T(f_{i_1}, a_1), \ldots, (f_{i_m}, a_m)$. A decision rule over T (1) corresponds to the subtable $T' = T(f_{i_1}, a_1), \ldots, (f_{i_m}, a_m)$ of T.

If a row r belongs to T', then the rule (1) is called *realisable for a row* r. When each row of T', for which the rule (1) is realisable, has the decision a attached to it, then the rule is called *true* for T. If the considered rule is true for T and realisable for r, then it is a *rule for* T *and* r.

The length of the rule (1) is defined by the number of descriptors from the left hand-side of the rule, and is denoted as m. The number of rows in T', which are labelled with the decision a gives the support of the rule (1). If a rule is true for T, then its support equals $N(T')$.

2.3.2. Description of heuristics

Taking into account the way in which decision rules are constructed, the seven heuristics presented below can be divided into two groups:

- heuristics with fixed decision: M, RM, $Poly$, Log, $MaxSupp$,
- heuristics with the most common decision: Me, Mep.

In the case of fixed decision, each heuristic H constructs a decision rule for the table T and a given row r with assigned decision a. Greedy heuristic H starts with a decision rule in which the left hand-side is empty, $\to d = a$.

In the case of the most common decision, each heuristic H constructs a decision rule for the table T and a given row r. It starts with an empty decision rule, $\to$, and at the end of the work the right hand-side of a decision rule is denoted by $d = a$, where a is the most common decision for corresponding T'.

For both types of heuristics, in each iteration such attribute $f_i \in \{f_1, \ldots, f_n\}$ fulfilling heuristic H is selected, which has the minimum index. Each heuristic is applied sequentially, for each row r of T, so at the end of the work, the number of induced rules equals $|U|$. Algorithm 1 lists a pseudo-code for the greedy heuristic H with fixed decision, for the procedure of constructing a decision rule for a row r from T.

To describe the work of heuristics, the following notation is introduced: $T^{(j+1)} = T^{(j)}(f_i, b_i)$, with j giving an index of the subsequently obtained subtable in the execution of heuristic H.

For heuristics with fixed decision,
$M(f_i, r, a) = M(T^{(j+1)}, a) = N(T^{(j+1)}) - N(T^{(j+1)}, a)$,

Algorithm 1 Greedy heuristic H with fixed decision for construction of a decision rule for T and r.

```
Require: Decision table $T$ with condition attributes $f_1,\dots,f_n$, row $r=(b_1,\dots,b_n)$
Ensure: Decision rule for $T$ and $r$
  begin
  $Q \leftarrow \emptyset$;
  $j \leftarrow 0$;
  $T^{(j)} \leftarrow T$;
  while all rows in $T^{(j)}$ are not assigned the same decision $a$ do
      select $f_i \in \{f_1,\dots,f_n\}$ with the minimum index fulfilling the heuristic $H$;
      $T^{(j+1)} \leftarrow T^{(j)}(f_i, b_i)$;
      $Q \leftarrow Q \cup \{f_i\}$;
      $j = j+1$;
  end while
  $\bigwedge_{f_i \in Q}(f_i = b_i) \rightarrow d = a$, where $a$ is a decision value.
  end
```

$RM(f_i, r, a) = \big(N(T^{(j+1)}) - N(T^{(j+1)}, a)\big)/N(T^{(j+1)})$,
$\alpha(f_i, r, a) = N(T^{(j)}, a) - N(T^{(j+1)}, a)$ and
$\beta(f_i, r, a) = M(T^{(j)}, a) - M(T^{(j+1)}, a)$,
the attribute $f_i \in E(T^{(j)})$ is selected by each heuristic H through:

- minimisation of
  - for M—the value of $M(f_i, r, a)$,
  - for RM—the value of $RM(f_i, r, a)$,
  - for $MaxSupp$—the value of $\alpha(f_i, r, a)$ given that $\beta(f_i, r, a) > 0$,
- maximisation of
  - for $Poly$—the value of $\frac{\beta(f_i,r,a)}{\alpha(f_i,r,a)+1}$,
  - for Log—the value of $\frac{\beta(f_i,r,a)}{\log_2(\alpha(f_i,r,a)+2)}$.

For heuristics with the most common decision, heuristic H selects the attribute $f_i \in E(T^{(j)})$ by minimisation of:

- for Me—the value of
  $Me(f_i, r_i) = N(T^{(j+1)}) - N\left((T^{(j+1)}), mcd(T^{(j+1)})\right)$,
- for Mep—the value of
  $Mep(f_i, r_i) = \left(N(T^{(j+1)}) - N\left((T^{(j+1)}), mcd(T^{(j+1)})\right)\right)/N(T^{(j+1)})$.

Example 1. *This example presents how heuristic H constructs a decision rule for the decision table T_0 and row r_1.*

$T_0 =$

	f_1	f_2	f_3	d
r_1	1	1	0	Yes
r_2	2	0	1	No
r_3	2	0	0	No
r_4	2	1	0	Yes

The decision table T_0 has three condition attributes, which leads to considerations for three subtables:

$T_1^{(1)} = T_0^{(0)}(f_1, 1) =$

	f_1	f_2	f_3	d
r_1	1	1	0	Yes

$T_2^{(1)} = T_0^{(0)}(f_2, 1) =$

	f_1	f_2	f_3	d
r_1	1	1	0	Yes
r_4	2	1	0	Yes

$T_3^{(1)} = T_0^{(0)}(f_3, 0) =$

	f_1	f_2	f_3	d
r_1	1	1	0	Yes
r_3	2	0	0	No
r_4	2	1	0	Yes

- *Heuristics with fixed decision*
  At the beginning the decision rule for r_1 has the form: $\rightarrow d = Yes$.
  - *Heuristic M:*
    $M(f_1, r_1, Yes) = 0$, $M(f_2, r_1, Yes) = 0$, $M(f_3, r_1, Yes) = 1$,
    so the rule $f_1 = 1 \rightarrow d = Yes$ is obtained.
  - *Heuristic RM:*

$RM(f_1, r_1, Yes) = 0$, $RM(f_2, r_1, Yes) = 0$, $RM(f_3, r_1, Yes) = \frac{1}{3}$,
so the rule $f_1 = 1 \rightarrow d = Yes$ is obtained.

- *Heuristic $MaxSupp$:*
  $\alpha(f_1, r_1, Yes) = 1$, $\beta(f_1, r_1, Yes) = 2$,
  $\alpha(f_2, r_1, Yes) = 0$, $\beta(f_2, r_1, Yes) = 2$,
  $\alpha(f_3, r_1, Yes) = 0$, $\beta(f_3, r_1, Yes) = 1$,
  so the rule $f_2 = 1 \rightarrow d = Yes$ is obtained.
- *Heuristic $Poly$:*
  $\frac{\beta(f_1,r_1,Yes)}{\alpha(f_1,r_1,Yes)+1} = \frac{2}{2}$, $\frac{\beta(f_2,r_1,Yes)}{\alpha(f_2,r_1,Yes)+1} = \frac{2}{1}$, $\frac{\beta(f_3,r_1,Yes)}{\alpha(f_3,r_1,Yes)+1} = \frac{1}{1}$,
  so the rule $f_2 = 1 \rightarrow d = Yes$ is obtained.
- *Heuristic Log:*
  $\frac{\beta(f_1,r_1,Yes)}{\log_2(\alpha(f_1,r_1,Yes)+2)} = \frac{2}{\log_2 3}$, $\frac{\beta(f_2,r_1,Yes)}{\log_2(\alpha(f_2,r_1,Yes)+2)} = \frac{2}{\log_2 2}$,
  $\frac{\beta(f_3,r_1,Yes)}{\log_2(\alpha(f_3,r_1,Yes)+2)} = \frac{1}{\log_2 2}$,
  so the rule $f_2 = 1 \rightarrow d = Yes$ is obtained.

- *Heuristics with the most common decision*
  At the beginning the decision rule for r_1 has the form: $\rightarrow$.
  - *Heuristic Me:*
    $Me(f_1, r_1) = 0$, $Me(f_2, r_1) = 0$, $Me(f_3, r_1) = 1$,
    so the rule $f_1 = 1 \rightarrow d = Yes$ is obtained.
  - *Heuristic Mep:*
    $Mep(f_1, r_1) = 0$, $Mep(f_2, r_1) = 0$, $Mep(f_3, r_1) = \frac{1}{3}$,
    so the rule $f_1 = 1 \rightarrow d = Yes$ is obtained.

In the research works described in this paper, the presented seven heuristics were used to discover knowledge on characteristic features employed in the task of authorship attribution, with stylometry as the application domain.

2.4. Stylometric authorship attribution

In stylometric analysis authorship attribution is a task of paramount importance [21]. It involves obtaining such definition of a writing style that can be used in pattern recognition by modern data mining approaches, machine learning algorithms, statistic calculations [56,57]. For this purpose quantitative style descriptors are required and they often refer to individual linguistic habits and preferences of writers, displayed in employed words (called lexical markers) and patterns of sentence formulation indicated with punctuation marks (syntactic descriptors). As a consequence, stylometric characteristic features frequently are of continuous type.

Calculation of selected markers is typically executed over sets including many text samples of comparable size [58], as with the size not only characteristics can vary, but language elements differ too. In the popularly employed practice longer works are divided into smaller chunks of text, but then in the knowledge discovery process it is important not to use samples originating in the same whole both for training and testing, as it artificially increases recognition [59], falsifying results. Hence the need for completely separate test sets for evaluation of performance when authorship attribution is treated as a classification task.

2.5. Discretisation

Discretisation transforms the input continuous space into discrete by the process of controlled loss of information [60]. One of the popular discretisation approaches are supervised and unsupervised algorithms. In case of supervised methods, information about class labels is taken into account while searching for intervals among ranges of attribute values. Some heuristic measures, e.g. entropy, can be used to determine the best cut-points. In case of unsupervised methods information about class labels is omitted during discretisation process.

In the experiments reported, Fayyad and Irani [61], and Kononenko [62] algorithms were used as representatives of supervised discretisation algorithms (denoted as DsF and DsK). In both the process of finding cut-points starts from one interval containing all values of each discretised attribute, and then partitioning is repeated recursively, until a stopping criterion is met. The methods are based on class entropy of considered intervals for evaluating cut-points and Minimum Description Length (MDL) [63] principle as a stopping criterion.

As representatives of unsupervised discretisation algorithms equal width binning (denoted as Duw), and equal frequency binning (denoted as Duf) [23] were used. In both methods the input parameter k, defined by a user, determines the number of bins and each bin is associated with a distinct discrete value. The disadvantage of these methods is that in cases where the values of continuous attribute are not distributed evenly, even some relevant information can be lost after the discretisation process.

3. Steps of proposed methodology

The framework of executed experiments included:

1. Preparation of input data sets for the task of authorship attribution, which consisted of
   - construction of text samples for analysis,
   - selection of stylometric features and obtaining their values for all samples,
   - discretisation of all sets by chosen methods;
2. Induction of decision rules with
   - exhaustive algorithm in classical rough set approach,
   - seven greedy heuristics described in the paper;
3. Analysis of induced rule sets which led to
   - limiting considerations to heuristics with most promising parameters by observing Pareto points in the optimisation space,
   - constructing feature rankings based on characteristics of rule sets returned by the three chosen heuristics,
   - selecting for further processing these variants of discrete data sets, for which RST approach found rules with the best characteristics;
4. Induction of decision rules with rough set approach for reduced subsets of features;
5. Evaluation of performance for rule classifiers with test sets.

The details for all constituent stages and comments to obtained experimental results are given in the next two sections of the paper.

4. Experiments leading to attribute characterisation

The section presents specifics of executed experiments from data preparation step to obtaining rankings for features, while evaluation of performance for rule classifiers is discussed in the next section.

4.1. Input data sets and features

Two data sets were prepared for the research, each enabling comparison and recognition of two authors, a pair of two female and a pair of two male writers, namely Edith Wharton and Mary Johnston, and Jack London and James Oliver Curwood. Their works were divided into three groups, corresponding to one training and two test sets. All novels were partitioned into smaller text samples, over which frequencies of occurrence were calculated for a hundred of stylometric descriptors, selected function words and punctuation marks.

Next, to both training sets (respectively for female and male authors) several feature rankings were applied (implemented in WEKA [64] environment) and their results compared. The attributes which were at least once considered as irrelevant (were assigned a rank of 0) were rejected. The variables left were found as relevant by all ranking mechanisms. This initial processing led to the set of 24 characteristic features (22 lexical and 2 syntactic markers), which were used in all following stages of the executed experiments. The values of condition attributes, reflecting frequencies of occurrence of linguistic elements, ranged between 0 and 1, while class labels corresponded to recognised authors.

Thus constructed input data sets provided examples of a binary classification task, with balanced classes (100 samples per author in a training set, and 45 per author in both test sets) and continuous features. As for the purposes of intended processing discrete attributes were needed, the next step of experiments was dedicated to discretisation.

4.2. Discretisation approaches employed

With the aim at extended observations, in the research four discretisation methods were used: two supervised and two unsupervised. As representatives of supervised category, the algorithms by Fayyad and Irani (DsF), and Kononenko (DsK) were chosen. These discretisation approaches do not require any input from a user and return single versions for each discretised set, with possibly varying numbers of intervals assigned for different variables.

Equal width binning (Duw) and equal frequency binning (Duf) were chosen as representatives of unsupervised discretisation category. In both cases the input parameter defines the number of constructed bins, the same for all features in a set. For both methods this number was varied from 2 to 10.

The step devoted to discretisation resulted in obtaining 20 variants for each of input sets (9 per each unsupervised method plus 1 per each supervised). All sets were discretised independently on others.

4.3. Induction of decision rules by RST

All versions of learning sets for both data sets (male and female writers) were next subject to rough set processing. The decision rules were induced by exhaustive algorithm implemented in RSES system [49]. It constructs all minimal decision

Table 1
Characteristics of rule sets induced in rough set approach for both data sets. Columns list: (a) for unsupervised discretisation the number of bins, (b) number of rules, (c) average support of rules, (d) average rule length.

	Female writer data set			Male writer data set		
(a)	(b)	(c)	(d)	(b)	(c)	(d)
			Unsupervised equal width binning (Duw)			
2	2094	**6.5**	5.5	1509	**7.5**	4.6
3	26025	3.6	5.6	32447	3.3	5.7
4	46480	2.6	5.0	47574	2.9	5.0
5	67054	2.1	4.5	79561	2.3	4.6
6	75888	2.0	4.1	77033	2.0	4.2
7	70152	1.8	3.9	75733	1.9	4.0
8	60422	1.8	3.7	72675	1.8	3.7
9	59332	1.7	3.5	68722	1.7	3.5
10	54187	1.6	**3.4**	61920	1.6	**3.4**
			Unsupervised equal frequency binning (Duf)			
2	103645	**3.7**	5.7	138910	**3.6**	5.6
3	122527	2.1	4.5	135696	2.0	4.4
4	81723	1.8	3.8	96327	1.8	3.7
5	68994	1.7	3.5	76240	1.7	3.4
6	58327	1.6	3.2	65184	1.5	3.2
7	49026	1.5	3.1	55184	1.5	3.1
8	42490	1.5	3.0	47511	1.5	3.0
9	37750	1.5	2.9	42278	1.5	2.9
10	34155	1.4	**2.8**	38670	1.4	**2.8**
DsF	4121	**6.6**	**4.8**	15283	**5.9**	**5.1**
DsK	10190	**5.4**	**5.3**	20815	**5.5**	**5.1**

rules, i.e., rules with minimal number of descriptors (pairs attribute=value) in their premise parts. The characteristics of rule sets found are listed in Table 1. For unsupervised discretisation the results are given for all versions of training sets, corresponding to the numbers of bins defined for all variables. For supervised discretisation there were single results, listed in the two bottom rows of the table, and denoted as DsF for Fayyad algorithm, and DsK for Kononenko.

It can be observed that for both data sets for equal width binning with increased numbers of bins the number of induced rules firstly steeply rises, then slowly decreases. For equal frequency binning these numbers of rules are at the maximum for small bin numbers and then become smaller. What is of the highest interest in this table, are listed values of average support and length of rules. As higher support offers higher probability of matching to unknown samples and good predictions, we would like this parameter to be as high as can be obtained (the maximum is shown in bold). On the other hand, shorter rules are more general and they have a better chance at causing a hit in tests, thus the length is preferred as low as possible (the minimum is shown in bold). Here the lowest average lengths exist only for cases with low support, and the best ratio between these two elements is for unsupervised equal width binning with just two bins defined for attributes. Only for these versions of training sets the inferred rule sets showed comparable characteristics to those obtained for data sets discretised in supervised Fayyad and Kononenko approaches.

These considerations led to selection of the three variants of training sets for both data sets: supervised discretisation based on Fayyad method, supervised discretisation with Kononenko algorithm, and unsupervised equal width binning with two bins (denoted as Duw02). These versions of data sets were used for further research dedicated to feature selection and reduction.

4.4. Characteristics of rule sets induced through heuristics

The previously described seven heuristics were employed for the task of decision rule induction for all 20 versions of the two training sets. Heuristics found a single rule for each row of each decision table, however, some of these rules were duplicated, thus the numbers of unique rules were much smaller (in particular when considered in comparison with cardinalities of rule sets found by rough set approach in exhaustive algorithm). The most interesting characteristics, that is averaged support and length, are given for all heuristics in Table 2 for female writer data set and in Table 3 for male writers.

When these two characteristics were compared among all versions and all heuristics it became clear that some of heuristics were much more promising than others, which is best visible in terms of Pareto points in optimisation space with two dimensions, one for support and the other for length. Firstly, for all rule sets returned by heuristics averages were calculated, given in the bottom rows of Tables 2 and 3 respectively for each data set. Then the corresponding data points in the optimisation space were observed, as shown in Fig. 1.

Analysis of points led to the immediate conclusion that heuristic MaxSupp was the worst of all, generating long rules with low supports. On the other hand, Log offered the highest support thus was the obvious choice as one Pareto point, but it was not the only point as this heuristic did not give the shortest rules. With respect to lowest averaged length two

Table 2
Characteristics of rule sets generated by heuristics for female writer data set. Columns list: (a) for unsupervised discretisation the number of bins, (b) average support of rules, (c) average rule length.

	Heuristic													
	LOG		M		MaxSupp		ME		MEP		POLY		RM	
(a)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)
Unsupervised equal width binning (Duw)														
2	11.1	4.1	3.5	2.2	4.8	11.2	2.9	2.4	6.9	5.8	9.1	6.3	6.9	2.3
3	7.4	3.5	3.7	1.9	2.6	9.6	3.7	2.0	4.5	2.2	4.5	7.5	4.5	1.9
4	8.3	3.3	2.7	1.7	1.5	9.0	2.7	1.7	3.2	1.7	3.6	6.2	3.4	1.7
5	6.9	2.8	3.3	1.6	1.7	6.7	3.2	1.7	4.1	1.6	4.4	4.6	4.3	1.6
6	5.7	2.7	3.2	1.5	1.6	6.1	3.2	1.5	3.6	1.5	2.6	4.5	3.6	1.5
7	6.9	2.5	2.9	1.5	1.9	5.7	2.8	1.5	3.0	1.5	3.1	4.3	3.1	1.5
8	6.5	2.3	2.9	1.4	1.7	5.8	2.9	1.4	3.3	1.4	2.6	4.2	3.3	1.4
9	5.9	2.3	2.8	1.4	1.8	5.5	2.8	1.4	3.0	1.3	2.7	4.2	3.0	1.3
10	5.6	2.5	2.7	1.4	1.8	5.6	2.7	1.4	3.0	1.4	2.4	4.3	3.1	1.4
Unsupervised equal frequency binning (Duf)														
2	18.2	2.5	10.0	2.1	9.5	4.5	7.1	3.4	10.5	5.3	14.3	3.1	12.5	2.1
3	8.0	2.8	5.0	2.1	3.8	4.8	4.9	2.2	5.4	2.7	5.9	3.6	5.6	2.2
4	9.1	2.3	3.7	1.9	3.0	4.7	3.6	2.0	4.0	2.1	5.1	3.5	4.1	1.9
5	6.5	2.4	4.4	1.8	1.9	5.1	4.3	1.8	4.4	1.9	3.2	3.9	4.7	1.8
6	7.1	2.3	4.3	1.8	1.9	5.1	4.1	1.8	4.3	1.8	3.2	4.0	4.7	1.7
7	7.1	2.2	3.6	1.8	1.7	5.2	3.3	1.8	3.9	1.7	2.6	4.0	4.1	1.7
8	7.7	2.1	3.7	1.7	1.7	5.2	3.7	1.7	4.2	1.6	2.2	4.1	4.2	1.6
9	6.7	2.3	4.1	1.5	1.7	5.2	4.1	1.5	4.3	1.5	2.3	4.1	4.4	1.5
10	4.9	2.2	3.8	1.4	1.6	5.3	3.8	1.4	3.8	1.4	2.2	4.3	3.9	1.4
DsF	14.3	3.0	7.4	1.9	6.9	7.1	6.7	2.1	7.7	2.8	11.8	5.4	7.7	1.8
DsK	14.3	3.0	7.4	1.9	6.1	8.3	6.7	2.1	7.7	2.8	11.8	5.5	7.7	1.8
Avg	11.0	2.8	5.7	1.8	4.5	6.9	5.2	2.0	6.1	2.5	8.0	5.0	6.2	1.7

Table 3
Characteristics of rule sets generated by heuristics for male writer data set. Columns list: (a) for unsupervised discretisation the number of bins, (b) average support of rules, (c) average rule length.

	Heuristic													
	LOG		M		MaxSupp		ME		MEP		POLY		RM	
(a)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)	(b)	(c)
Unsupervised equal width binning (Duw)														
2	6.5	3.4	2.7	2.4	4.3	11.5	2.3	2.6	3.6	6.2	6.5	7.0	4.2	2.4
3	7.4	3.1	3.8	1.9	1.7	10.8	3.5	2.0	3.6	2.7	4.3	6.0	4.1	1.9
4	5.6	2.7	2.3	1.9	1.5	9.5	2.2	1.9	2.6	2.2	3.6	5.3	2.7	1.8
5	5.3	2.3	2.9	1.7	1.7	7.1	2.6	1.7	3.1	1.9	3.0	5.0	3.2	1.6
6	5.0	2.5	3.0	1.5	1.7	6.5	3.0	1.5	3.3	1.5	2.7	4.5	3.3	1.5
7	4.5	2.0	2.6	1.6	1.7	6.4	2.6	1.6	2.9	1.6	2.4	4.4	2.9	1.5
8	4.5	1.8	2.5	1.5	1.8	5.6	2.5	1.5	3.1	1.5	2.2	4.2	3.1	1.4
9	4.8	1.9	2.6	1.4	1.7	5.2	2.6	1.5	2.7	1.5	2.2	3.7	2.7	1.4
10	4.9	1.7	2.6	1.4	1.6	5.1	1.6	1.4	2.7	1.4	2.2	3.8	2.7	1.3
Unsupervised equal frequency binning (Duf)														
2	5.4	3.1	4.8	3.0	3.1	5.1	3.8	4.2	3.8	5.7	5.9	3.5	4.5	3.0
3	6.7	2.5	5.1	2.3	3.4	4.1	3.4	4.1	5.3	3.3	5.4	3.5	5.9	2.3
4	5.3	2.1	3.5	2.0	2.0	4.6	3.1	2.1	3.6	2.4	3.2	3.6	4.3	2.0
5	5.3	2.1	3.4	2.0	1.8	4.7	3.2	2.0	3.8	2.1	2.6	3.7	4.0	2.0
6	5.3	1.9	3.7	1.9	1.7	4.5	3.6	1.9	4.2	1.9	2.3	3.7	4.2	1.9
7	5.6	1.9	3.4	1.9	1.7	4.6	3.3	1.9	3.7	1.9	2.2	3.6	3.8	1.9
8	4.5	1.9	3.3	1.9	1.7	4.8	3.1	1.9	3.6	1.9	2.2	3.7	3.7	1.9
9	4.7	1.8	3.6	1.8	1.6	4.8	3.4	1.8	3.6	1.8	2.2	3.7	3.7	1.8
10	4.3	1.8	3.6	1.7	1.5	4.8	3.5	1.7	3.7	1.7	1.9	3.7	3.7	1.7
DsF	12.5	3.5	7.1	2.3	3.7	7.0	6.5	2.6	7.1	3.6	7.1	4.6	8.0	2.3
DsK	13.3	3.1	6.9	2.2	3.4	7.0	6.1	2.5	6.2	3.8	8.3	4.5	7.4	2.3
Avg	5.3	2	3.6	2.0	1.9	4.6	3.4	2.0	3.9	2.2	2.8	3.7	4.3	2.0

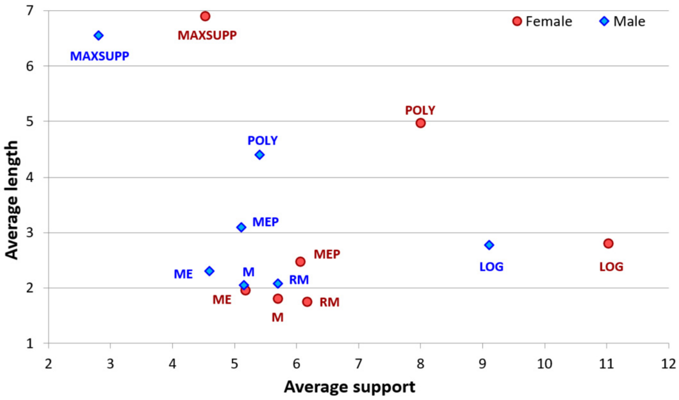

Fig. 1. Points in the optimisation space for rule sets generated by all heuristics for both data sets.

Table 4
Rankings of features obtained through rule sets induced by heuristics.

Ranking position	F-AllH	M-AllH	Ranking position	F-AllH	M-AllH
	Attribute	Attribute		Attribute	Attribute
1	atr23	atr0	13	atr5	atr10
2	atr1	atr23	14	atr7	atr22
3	atr17	atr3	15	atr19	atr19
4	atr0	atr1	16	atr11	atr9
5	atr2	atr2	17	atr4	atr13
6	atr13	atr16	18	atr12	atr12
7	atr22	atr17	19	atr14	atr11
8	atr20	atr6	20	atr16	atr14
9	atr3	atr18	21	atr21	atr5
10	atr10	atr21	22	atr9	atr20
11	atr8	atr7	23	atr18	atr4
12	atr6	atr8	24	atr15	atr15

heuristics gave very close results, RM and M, which led to including them in further considerations as well, while discarding the four that were so much weaker. These conclusions confirmed the previously reported findings with respect to optimality of greedy heuristics [55], especially for M and RM heuristics as presented as the best from the point of view of minimisation of rule lengths.

The rule sets induced through the selected three heuristics were next analysed with respect to included features.

4.5. Characteristics of attributes by rule sets constructed through heuristics

The rule sets inferred by all heuristics varied in numbers of unique rules and rule characteristics, such as support and length. They also varied in use of attributes included as conditions in their premises. In fact, some of attributes were employed much more often than others, which reflected the degree of importance of features as perceived by algorithms mining knowledge from them. When discovered knowledge is represented in the form of obtained rules it can be exploited as a source of information on particular attributes and their relative relevance. This line of reasoning led to construction of rankings for all considered features. The proposed scoring function was defined as follows.

Firstly, for all 20 variants of train sets, for the rule sets generated by the three chosen heuristics, there were calculated relative frequencies of occurrence of attributes in rules, giving the ratio of conditions including a feature to the total number of conditions in all rules in the set. Secondly, for unsupervised discretisation for each of the two methods, the overall averages were calculated. It led to obtaining one averaged characteristic for each of four discretisation methods used in research, for all features and heuristics. Thirdly, averages over all four discretisation methods were calculated, returning scores from each heuristic for all attributes. And lastly, the final overall score function was calculated as the average from the three heuristics. These resulting rankings, based on characteristics of all heuristics, for both data sets are listed in Table 4, and denoted as F-AllH (for female writer data set) and M-AllH (for male).

It can be observed that only few attributes held the same positions in both rankings, few were placed closely, while others differed significantly. These findings reflect the fact that writing styles of authors of opposite gender typically display distinctive variations, which is the reason for not including them in the same data set.

These two rankings of attributes were employed for the purpose of feature elimination. From the complete set the features were rejected one by one, starting at the lowest ranking positions and then going up the ranking. For each reduced set of attributes new rule sets were inferred and the performance of rule classifiers evaluated with test sets, as shown in the next section.

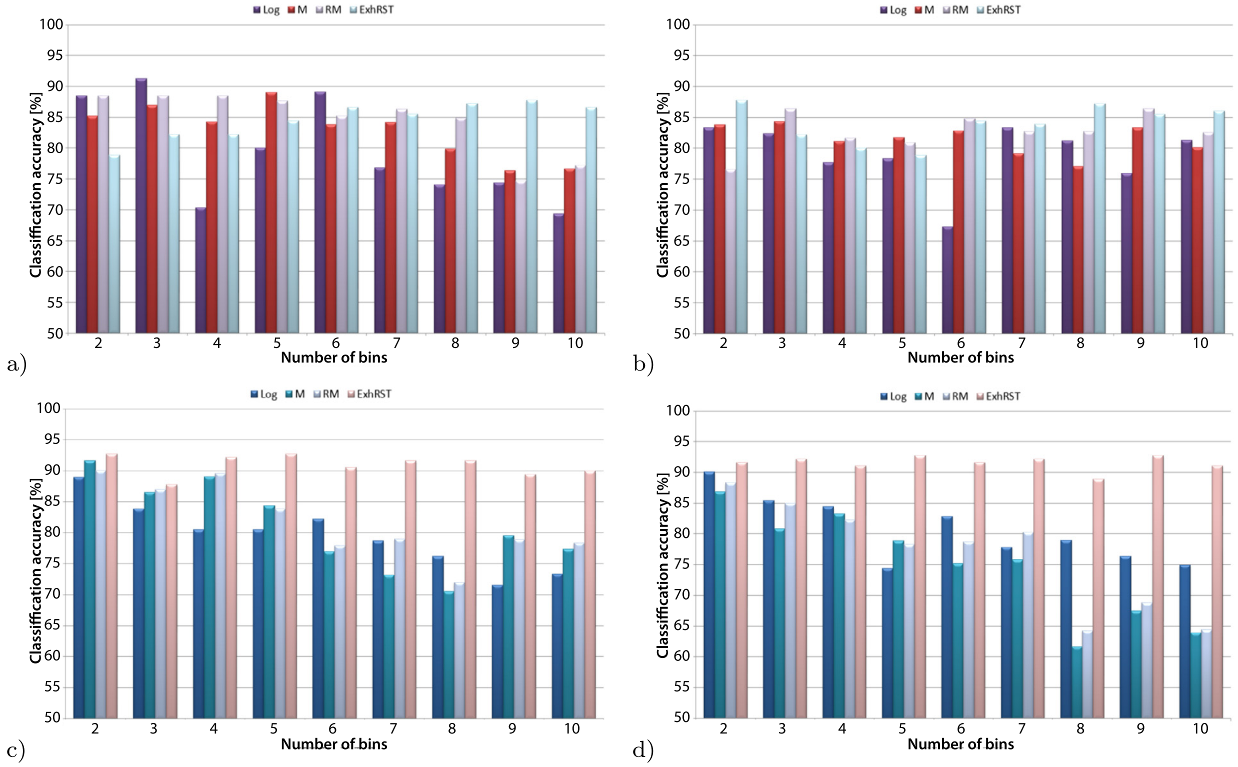

Fig. 2. Performance of rule classifiers for sets discretised with a) and b) unsupervised equal width binning (Duw) approach, c) and d) unsupervised equal frequency binning (Duf), on the left for female writer data set, on the right for male writer data set.

5. Experiments on feature selection

Feature selection processes typically aim at reduction of the number of attributes, while at the same time at least preserving the power of employed classifiers. It means that effectiveness of feature reduction can be analysed when the performance for the entire set of features is known.

5.1. Evaluation of performance for rule classifiers for all features

Before starting feature reduction process, to provide a reference point, the performance of rule classifiers was evaluated for the complete set of considered features for all discrete versions of data sets. All rule sets were applied to test samples for their classification, using standard voting in case of conflicts [65], that is with assigning each rule as many votes as its support. The performance was evaluated for the three selected heuristics (Log, M, RM), and the rule sets induced by exhaustive algorithm implemented in RSES system (denoted as ExhRST). The results are shown in Fig. 2. For all tables and charts the results present classification accuracy, understood as the number of correctly attributed instances to the total number of samples, averaged over two test sets.

In the plots categories correspond to numbers of bins defined for all variables, ranging from 2 to 10, and series to decision algorithms (either heuristic or ExhRST). For equal frequency binning rough set approach clearly outperformed all heuristics, giving the best results regardless of the number of intervals considered. For this discretisation method the performance of heuristic based rule sets rather decreased with the increasing numbers of bins. For equal width binning ExhRST decision algorithms were better in some cases, but not all, despite the advantage of access to so many induced rules. Heuristics fared with varying degree of success, there was no clear visible trend.

Evaluation of performance of rule classifiers in case of supervised discretisation applied to data sets is shown in Table 5. There is also included standard deviation calculated based on samples.

For female writers rough set approach returned rules that were unmatched in classification for both discretisation methods, while for male writers that was true only for Kononenko method. For heuristic M for female writers the results were surprising low, but then for male writers and Fayyad discretisation the same heuristic was the best. Thus it can be concluded that induced rule sets delivered to some degree on their promises, as estimated by rule characteristics previously analysed.

Table 5
Performance of rule classifiers [%] (± standard deviation based on samples) observed for data sets discretised by supervised methods.

Supervised discretisation by	Decision algorithm			
	Log	M	RM	ExhRST
	Female writer data set			
Fayyad (DsF)	87.62 ± 2.22	51.11 ± 1.57	52.22 ± 3.14	**96.11** ± 2.36
Kononenko (DsK)	88.14 ± 0.02	67.11 ± 24.35	67.78 ± 25.14	**97.23** ± 0.78
	Male writer data set			
Fayyad (DsF)	69.01 ± 23.96	**93.89** ± 0.78	92.78 ± 0.78	88.89 ± 0.00
Kononenko (DsK)	65.00 ± 18.22	85.47 ± 14.27	78.12 ± 24.67	**90.00** ± 4.71

Table 6
Performance of rule classifiers [%] (± standard deviation based on samples) observed in the process of feature elimination for the three selected versions of discrete training sets.

Nr of attributes	Female writer data set			Male writer data set		
	Duw02	DsF	DsK	Duw02	DsF	DsK
2	87.78 ± 6.29	88.89 ± 4.71	88.89 ± 4.71	0.00 ± 0.00	58.89 ± 12.57	58.89 ± 12.57
3	87.78 ± 6.29	93.89 ± 0.78	93.89 ± 0.78	81.11 ± 1.57	50.00 ± 0.00	65.00 ± 22.78
4	87.78 ± 6.29	85.00 ± 3.93	89.45 ± 2.35	82.22 ± 1.57	88.33 ± 5.50	73.94 ± 18.00
5	87.78 ± 6.29	85.00 ± 3.93	91.11 ± 4.71	85.00 ± 2.36	88.33 ± 5.50	75.68 ± 21.82
6	87.78 ± 6.29	89.45 ± 2.35	91.11 ± 4.71	84.44 ± 0.00	91.11 ± 0.00	79.75 ± 19.21
7	88.33 ± 5.50	93.89 ± 2.36	95.56 ± 4.72	87.78 ± 4.72	90.56 ± 2.35	81.17 ± 15.63
8	88.89 ± 4.71	97.78 ± 1.57	97.78 ± 1.57	87.78 ± 3.14	93.89 ± 0.78	88.61 ± 6.68
9	**90.00** ± 4.71	97.23 ± 0.78	97.23 ± 0.78	82.23 ± 6.29	88.89 ± 3.14	87.22 ± 5.50
10	**90.00** ± 4.71	97.23 ± 2.35	97.23 ± 2.35	82.23 ± 6.29	92.22 ± 1.57	89.45 ± 0.78
11	85.56 ± 0.00	**98.34** ± 2.35	98.34 ± 2.35	82.23 ± 6.29	91.12 ± 6.29	90.00 ± 3.14
12	86.67 ± 1.57	**98.34** ± 2.35	98.34 ± 2.35	82.78 ± 5.50	93.89 ± 5.50	93.33 ± 1.57
13	87.23 ± 2.35	97.78 ± 6.29	97.23 ± 2.35	84.45 ± 4.72	**95.00** ± 5.50	93.34 ± 3.15
14	87.23 ± 2.35	**98.34** ± 0.78	97.78 ± 1.57	82.78 ± 7.07	92.23 ± 7.86	92.23 ± 6.29
15	86.67 ± 1.57	**98.34** ± 0.78	97.78 ± 1.57	83.34 ± 6.29	91.67 ± 7.07	91.12 ± 6.29
16	86.11 ± 2.36	97.78 ± 1.57	97.23 ± 2.35	83.89 ± 7.07	93.89 ± 7.07	93.34 ± 6.29
17	85.56 ± 4.72	97.78 ± 0.00	98.34 ± 0.78	84.45 ± 7.86	**95.00** ± 5.50	93.34 ± 6.29
18	85.00 ± 7.07	96.67 ± 1.57	97.23 ± 2.35	84.45 ± 7.86	94.45 ± 4.72	**95.00** ± 3.93
19	85.00 ± 7.07	96.67 ± 1.57	97.78 ± 1.57	**89.45** ± 3.92	94.45 ± 4.72	**95.00** ± 2.36
20	85.00 ± 8.64	96.67 ± 1.57	97.23 ± 2.35	**89.45** ± 3.92	94.45 ± 4.72	**95.00** ± 2.36
21	87.22 ± 5.50	96.67 ± 1.57	97.23 ± 2.35	88.33 ± 5.50	93.34 ± 3.15	93.89 ± 0.78
22	82.78 ± 8.64	97.23 ± 0.78	**98.89** ± 0.00	86.67 ± 7.86	92.22 ± 1.57	92.22 ± 3.14
23	83.34 ± 11.00	97.23 ± 0.78	**98.89** ± 0.00	87.22 ± 7.07	88.34 ± 2.35	90.56 ± 3.92
24	78.89 ± 12.57	96.11 ± 2.36	97.23 ± 0.78	87.78 ± 7.86	88.89 ± 0.00	90.00 ± 4.71

5.2. Evaluation of performance for rule classifiers for feature elimination

As previously explained, for the observations on feature reduction, three variants of discrete data sets were pre-selected, based on the promising characteristics of rule sets induced in rough set approach, that is with the best ratio of average support to average rule length. These variants corresponded to unsupervised equal width binning with just two bins (Duw02) for all attributes, supervised Fayyad approach (DsF), and supervised Kononenko algorithm (DsK).

For these chosen discrete data sets feature elimination was executed by removing a single element at a time from the set of available attributes, and the choice of variable was dictated by its position in a ranking constructed based on characteristics of rule sets inferred by heuristics (given as F-AllH and M-AllH in Table 4, for female and male writer data set respectively).

The processing started with the complete set of attributes and was continued till a single feature was left. However, in case of just one or two attributes in a decision table it can turn out that the table becomes contradictory, and no rules can be found to classify training samples. That is the reason for classification accuracy equal zero and as this happened for all tested sets with single variables working as only available features, this case was excluded from the results given in Table 6. There are also listed values of standard deviation based on samples. The bottom row of the table shows classification results of rule classifiers inferred with complete sets of features, provided as a reference point for comparison.

For each discretisation method the best classification result is indicated in the table in bold. For both data sets discretised by unsupervised equal width binning with just two bins for all attributes the performance observed was generally the lowest from the three discretisation approaches presented, while the other two algorithms resulted in relatively close classification results.

On the other hand, for female writers for Duw02 versions of discrete sets the observed improvement in classification accuracy was the highest, what is more, all but one of reduced subsets of attributes offered some increase in correct predictions. For Kononenko discretisation the percentage of correctly attributed samples was the highest (98.89%) of the three methods, but the increase was correspondingly the lowest, and happened after elimination of just either one or two

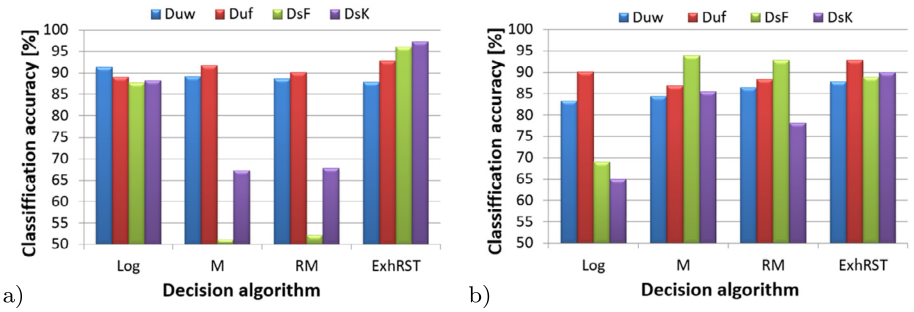

Fig. 3. Best performance of all rule classifiers for sets discretised with all tested approaches, for a) female writer data set, b) for male writer data set.

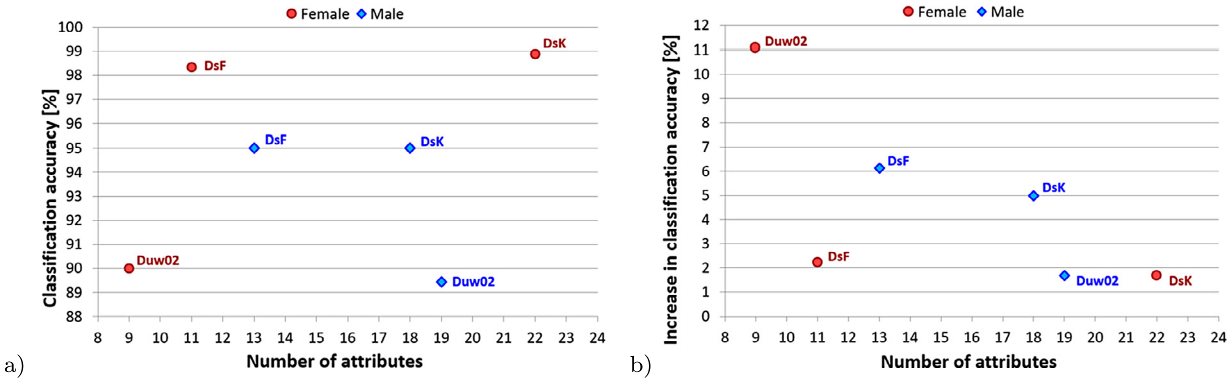

Fig. 4. Points in the optimisation space for rule sets generated in the process of feature elimination based on obtained rankings for both data sets.

features. Fayyad algorithm placed in the middle, with recognition ratio close to the best, but allowing to reduce two thirds of available features.

Male writer data set generally proved to be more difficult in recognition, with lower percentage of correct classifications than female writer data set. Both supervised discretisation methods achieved the same maximum in performance at 95%, but again for Kononenko fewer features (6 out of 24) could be reduced than for Fayyad approach (11 out of 24).

5.3. Summary of experiments leading to feature selection

A part of the experiments performed was dedicated to observations on the influence of the selected discretisation method applied to data sets on characteristics of attributes and rule sets inferred. For all four discretisation approaches used in research, Fig. 3 displays the best performance of rule classifiers inferred by selected heuristics and rough sets. The series correspond to discretisation method, while categories are defined by decision algorithms. This presentation enables to study the power of each classifier from the perspective of discretisation, which can be used for the choice of the method that is best suited.

For female writer data set there were observed surprisingly low results for supervised discretisation for all three greedy heuristics, always outperformed by decision algorithms induced from train sets subjected to unsupervised discretisation, with the extreme minimum of barely 50% for Fayyad method for both M and RM heuristics. Only for rough set approach supervised discretisation resulted in better performance of induced rule sets than for unsupervised binning. On the other hand, for male writer data set for both M and RM heuristics Fayyad method brought best classification accuracy. In all cases but one, unsupervised equal frequency binning gave better results than unsupervised equal width binning. Generally, both these unsupervised discretisation methods gave better results than could be expected based on criticism they so often receive.

Typical goals of feature selection include operation on reduced numbers of attributes while at least maintaining the original power of the studied classifiers. These aims were achieved for both data sets and all three selected discretisation algorithms. The cases of best performance for fewest attributes are shown in Fig. 4, which allows once again to consider the optimisation space with two dimensions. The plot on the left displays directly classification accuracy as obtained in the best case, on the right as a relative increase with respect to the performance for the decision rules induced for the entire set of features.

In the left chart it can be observed that for female writer data set there is no single Pareto point. The highest classification accuracy of 98.89% in supervised Kononenko discretisation approach occurs for the fewest reduced features — just two. Unsupervised equal width binning brings the highest reduction of features (9 left out of 24), but with the lowest number of correct predictions (90%) out of these three maxima. On the other hand, in the plot on the right for female writer data set a single Pareto point is visible: the very same Duw02 caused the highest increase in performance (11.11%) in relation to the rule set induced for the entire set of available attributes, for the highest number of reduced features.

For male writer data set in both charts single Pareto points can be observed for Fayyad supervised discretisation. The best performance was 95%, obtained for 13 out of 24 features, which corresponded to the increase of 6.11% when compared with results for the set of all attributes.

It is also important to note that for both data sets, for all three approaches to discretisation of these sets, it can be observed that reduction of features while following their ranking resulted in cases of improved performance. The numbers of induced rules were many times smaller than in the algorithms induced for the complete set of features, which is also an advantage of the proposed methodology and proves its merit.

6. Conclusions

The paper shows results from research conducted as a case study in stylometric domain and a task of authorship attribution, with main considerations dedicated to induction of rule sets for various versions of discretised input data sets and various induction algorithms. In the proposed new methodology greedy heuristics were employed for the purpose of gathering information on available features. The discovered knowledge was then stored in the form of generated rules, pointed by rule characteristics such as length and support. By exploiting these characteristics and observations of Pareto points heuristics were limited to the chosen three, and based on the frequency of usage of attributes as conditions in induced rules, rankings of features were constructed.

The obtained rankings were next employed as a way of governing feature reduction for rough set approach. Using exhaustive algorithm for rule induction, firstly the rules were inferred for the complete set of features, and then for their gradually decreased numbers. In processing selected versions of discrete input sets were used. The performance for all rule classifiers was evaluated with test sets, and experimental results showed cases of increased recognition for reduced sets of features and noticeably lower numbers of generated rules, which validated the proposed methodology.

CRediT authorship contribution statement

U. Stańczyk: Conceptualization, Investigation, Methodology, Validation, Writing - original draft, Writing - review & editing. **B. Zielosko:** Conceptualization, Investigation, Methodology, Software, Validation, Writing - original draft, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research presented was performed at the Silesian University of Technology, Gliwice, within the statutory project RAU6/2020, and at the University of Silesia in Katowice. Literary works used in experiments are available thanks to Project Gutenberg (www.gutenberg.org).

References

[1] Z. Pawlak, Rough sets and intelligent data analysis, Inf. Sci. 147 (2002) 1–12.
[2] Z. Pawlak, A. Skowron, Rudiments of rough sets, Inf. Sci. 177 (1) (2007) 3–27.
[3] A. An, N. Cercone, Rule quality measures improve the accuracy of rule induction: an experimental approach, in: Z.W. Raś, S. Ohsuga (Eds.), Foundations of Intelligent Systems, ISMIS 2000, in: Lecture Notes in Computer Science, vol. 1932, Springer, 2000, pp. 119–129.
[4] L. Wróbel, M. Sikora, M. Michalak, Rule quality measures settings in classification, regression and survival rule induction — an empirical approach, Fundam. Inform. 149 (2016) 419–449.
[5] H.S. Nguyen, Approximate Boolean reasoning: foundations and applications in data mining, in: J.F. Peters, A. Skowron (Eds.), Transactions on Rough Sets V, in: Lecture Notes in Computer Science, vol. 4100, Springer, 2006, pp. 334–506.
[6] Z. Pawlak, A. Skowron, Rough sets and Boolean reasoning, Inf. Sci. 177 (1) (2007) 41–73.
[7] T. Amin, I. Chikalov, M. Moshkov, B. Zielosko, Dynamic programming approach to optimization of approximate decision rules, Inf. Sci. 119 (2013) 403–418.
[8] T. Amin, I. Chikalov, M. Moshkov, B. Zielosko, Relationships between length and coverage of decision rules, Fundam. Inform. 129 (1–2) (2014) 1–13.
[9] B. Zielosko, Application of dynamic programming approach to optimization of association rules relative to coverage and length, Fundam. Inform. 148 (1–2) (2016) 87–105.
[10] J. Błaszczyński, R. Słowiński, M. Szeląg, Sequential covering rule induction algorithm for variable consistency rough set approaches, Inf. Sci. 181 (5) (2011) 987–1002.
[11] P. Clark, T. Niblett, The CN2 induction algorithm, Mach. Learn. 3 (4) (1989) 261–283.
[12] M. Sikora, L. Wróbel, A. Gudyś Guider, A guided separate-and-conquer rule learning in classification, regression, and survival settings, Knowl.-Based Syst. 173 (2019) 1–14.

[13] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc., 1993.
[14] M. Azad, B. Zielosko, M. Moshkov, I. Chikalov, Decision rules, trees and tests for tables with many-valued decisions-comparative study, in: J. Watada, L.C. Jain, R.J. Howlett, N. Mukai, K. Asakura (Eds.), Proceedings of the 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, KES 2013, in: Procedia Computer Science, vol. 22, Elsevier, 2013, pp. 87–94.
[15] J. Ang, K. Tan, A. Mamun, An evolutionary memetic algorithm for rule extraction, Expert Syst. Appl. 37 (2) (2010) 1302–1315.
[16] D. Ślęzak, J. Wróblewski, Order based genetic algorithms for the search of approximate entropy reducts, in: G. Wang, Q. Liu, Y. Yao, A. Skowron (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, RSFDGrC 2003, in: Lecture Notes in Computer Science, vol. 2639, Springer, 2003, pp. 308–311.
[17] M.J. Moshkov, M. Piliszczuk, B. Zielosko, On construction of partial reducts and irreducible partial decision rules, Fundam. Inform. 75 (1–4) (2007) 357–374.
[18] U. Stańczyk, B. Zielosko, K. Żabiński, Application of greedy heuristics for feature characterisation and selection: a case study in stylometric domain, in: H. Nguyen, Q. Ha, T. Li, M. Przybyla-Kasperek (Eds.), Rough Sets, IJCRS 2018, in: Lecture Notes in Computer Science, vol. 11103, Springer, Quy Nhon, Vietnam, 2018, pp. 350–362.
[19] I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh (Eds.), Feature Extraction: Foundations and Applications, Studies in Fuzziness and Soft Computing, vol. 207, Physica-Verlag, Springer, 2006.
[20] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[21] S. Argamon, K. Burns, S. Dubnov (Eds.), The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning, Springer, Berlin, 2010.
[22] E. Stamatatos, A survey of modern authorship attribution methods, J. Am. Soc. Inf. Sci. Technol. 60 (3) (2009) 538–556.
[23] J. Dougherty, R. Kohavi, M. Sahami, Supervised and unsupervised discretization of continuous features, in: Machine Learning Proceedings 1995: Proceedings of the 12th International Conference on Machine Learning, Elsevier, 1995, pp. 194–202.
[24] H. Liu, H. Motoda, Computational Methods of Feature Selection, Data Mining and Knowledge Discovery, Chapman & Hall/CRC, 2007.
[25] A. Janusz, D. Ślęzak, Rough set methods for attribute clustering and selection, Appl. Artif. Intell. 28 (3) (2014) 220–242.
[26] R. Jensen, Q. Shen, Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches, IEEE Press Series on Computational Intelligence, Wiley-IEEE Press, 2008.
[27] U. Stańczyk, Ranking of characteristic features in combined wrapper approaches to selection, Neural Comput. Appl. 26 (2) (2015) 329–344.
[28] R. Kohavi, G. John, Wrappers for feature subset selection, Artif. Intell. 97 (1) (1997) 273–324.
[29] U. Stańczyk, Weighting of attributes in an embedded rough approach, in: A. Gruca, T. Czachórski, S. Kozielski (Eds.), Man-Machine Interactions 3, in: Advances in Intelligent and Soft Computing, vol. 242, Springer-Verlag, Berlin, Germany, 2013, pp. 475–483.
[30] U. Stańczyk, Selection of decision rules based on attribute ranking, J. Intell. Fuzzy Syst. 29 (2) (2015) 899–915.
[31] X. Jia, L. Shang, B. Zhou, Y. Yao, Generalized attribute reduct in rough set theory, Knowl.-Based Syst. 91 (2016) 204–218.
[32] M. Grzegorowski, D. Ślęzak, On resilient feature selection: computational foundations of r-C-reducts, Inf. Sci. 499 (2019) 25–44.
[33] H. Ge, L. Li, Y. Xu, C. Yang, Quick general reduction algorithms for inconsistent decision tables, Int. J. Approx. Reason. 82 (2017) 56–80.
[34] J. Liang, F. Wang, C. Dang, Y. Qian, An efficient rough feature selection algorithm with a multi-granulation view, Int. J. Approx. Reason. 53 (6) (2012) 912–926.
[35] M.S. Raza, U. Qamar, Feature selection using rough set-based direct dependency calculation by avoiding the positive region, Int. J. Approx. Reason. 92 (2018) 175–197.
[36] Y. Yang, D. Chen, H. Wang, E. Tsang, D. Zhang, Fuzzy rough set based incremental attribute reduction from dynamic data with sample arriving, Fuzzy Sets Syst. 312 (2017) 66–86, theme: Fuzzy Rough Sets.
[37] Y. Yang, D. Chen, H. Wang, Active sample selection based incremental algorithm for attribute reduction with rough sets, IEEE Trans. Fuzzy Syst. 25 (4) (2017) 825–838.
[38] J. Liang, F. Wang, C. Dang, Y. Qian, A group incremental approach to feature selection applying rough set technique, IEEE Trans. Knowl. Data Eng. 26 (2) (2014) 294–308.
[39] Y. Liu, L. Zheng, Y. Xiu, H. Yin, S. Zhao, X. Wang, H. Chen, C. Li, Discernibility matrix based incremental feature selection on fused decision tables, Int. J. Approx. Reason. 118 (2020) 1–26.
[40] Y. Yao, Three-way granular computing, rough sets, and formal concept analysis, Int. J. Approx. Reason. 116 (2020) 106–125.
[41] Q. Wan, J. Li, L. Wei, T. Qian, Optimal granule level selection: a granule description accuracy viewpoint, Int. J. Approx. Reason. 116 (2020) 85–105.
[42] Y. Jing, T. Li, J. Huang, Y. Zhang, An incremental attribute reduction approach based on knowledge granularity under the attribute generalization, Int. J. Approx. Reason. 76 (2016) 80–95.
[43] A. Ferone, Feature selection based on composition of rough sets induced by feature granulation, Int. J. Approx. Reason. 101 (2018) 276–292.
[44] C. Wang, Y. Shi, X. Fan, M. Shao, Attribute reduction based on k-nearest neighborhood rough sets, Int. J. Approx. Reason. 106 (2019) 18–31.
[45] F. Pacheco, M. Cerrada, R. Sanchez, D. Cabrera, C. Li, J.V. de Oliveira, Attribute clustering using rough set theory for feature selection in fault severity classification of rotating machinery, Expert Syst. Appl. 71 (2017) 69–86.
[46] X. Wang, J. Yang, X. Teng, W. Xia, R. Jensen, Feature selection based on rough sets and particle swarm optimization, Pattern Recognit. Lett. 28 (4) (2007) 459–471.
[47] R. Jensen, Q. Shen, Finding rough set reducts with ant colony optimization, in: Proceedings of the 2003 UK Workshop on Computational Intelligence, 2003, pp. 15–22.
[48] Y. Chen, Q. Zhu, H. Xu, Finding rough set reducts with fish swarm algorithm, Knowl.-Based Syst. 81 (2015) 22–29.
[49] J. Bazan, M. Szczuka, The rough set exploration system, in: J.F. Peters, A. Skowron (Eds.), Transactions on Rough Sets III, in: Lecture Notes in Computer Science, vol. 3400, Springer, Berlin, Heidelberg, 2005, pp. 37–56.
[50] J. Bazan, H. Nguyen, S. Nguyen, P. Synak, J. Wróblewski, Rough set algorithms in classification problem, in: L. Polkowski, S. Tsumoto, T. Lin (Eds.), Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems, in: Studies in Fuzziness and Soft Computing, vol. 56, Physica, Heidelberg, 2000, pp. 49–88.
[51] T. Bonates, P.L. Hammer, A. Kogan, Maximum patterns in datasets, Discrete Appl. Math. 156 (6) (2008) 846–861.
[52] H.S. Nguyen, D. Ślęzak, Approximate reducts and association rules - correspondence and complexity results, in: RSFDGrC '99: Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, in: Lecture Notes in Computer Science, vol. 1711, Springer, 1999, pp. 137–145.
[53] U. Feige, A threshold of $\ln n$ for approximating set cover, J. ACM 45 (1998) 634–652, ACM, New York.
[54] M. Moshkov, B. Zielosko, Combinatorial Machine Learning - A Rough Set Approach, Studies in Computational Intelligence, vol. 360, Springer, 2011.
[55] F. Alsolami, T. Amin, M. Moshkov, B. Zielosko, K. Żabiński, Comparison of heuristics for optimization of association rules, Fundam. Inform. 166 (1) (2019) 1–14.
[56] M. Jockers, D. Witten, A comparative study of machine learning methods for authorship attribution, Lit. Linguist. Comput. 25 (2) (2010) 215–223.
[57] M. Koppel, J. Schler, S. Argamon, Computational methods in authorship attribution, J. Am. Soc. Inf. Sci. Technol. 60 (1) (2009) 9–26.
[58] M. Eder, Does size matter? Authorship attribution, small samples, big problem, Dig. Scholarship Humanit. 30 (2015) 167–182.

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

[60] S. Garcia, J. Luengo, J. Saez, V. Lopez, F. Herrera, A survey of discretization techniques: taxonomy and empirical analysis in supervised learning, IEEE Trans. Knowl. Data Eng. 25 (4) (2013) 734–750.

[61] U. Fayyad, K. Irani, Multi-interval discretization of continuous valued attributes for classification learning, in: Proceedings of the 13th International Joint Conference on Artificial Intelligence, vol. 2, Morgan Kaufmann Publishers, 1993, pp. 1022–1027.

[62] I. Kononenko, On biases in estimating multi-valued attributes, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI'95, vol. 2, Morgan Kaufmann Publishers Inc., 1995, pp. 1034–1040.

[63] J. Rissanen, Modeling by shortest data description, Automatica 14 (5) (1978) 465–471.

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

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