DOI: 10.15439/2023F7295
ISSN 2300-5963 ACSIS, Vol. 35

Filtering Decision Rules Driven by Sequential Forward and Backward Selection of Attributes: An Illustrative Example in Stylometric Domain

Beata Zielosko*, Urszula Stańczyk† Kamil Jabloński*
*University of Silesia in Katowice, Institute of Computer Science, Będzińska 39, 41-200 Sosnowiec, Poland
Email: beata.zielosko@us.edu.pl, kjablonski1@us.edu.pl
†Silesian University of Technology, Department of Graphics, Computer Vision and Digital Systems
Akademicka 2A, 44-100 Gliwice, Poland, Email: urszula.stanczyk@polsl.pl

Abstract*—The paper presents investigations concerning the decision rule filtering process controlled by the estimated relevance of available attributes. In the conducted study, two search directions were used, sequential forward selection and sequential backward elimination. The steps of sequential search were governed by three rankings obtained for variables, all related to characteristics of data and rules that can be induced, as follows, (i) a ranking based on the weighting factor referring to the occurrence of attributes in generated decision reducts, (ii) the OneR ranking exploiting short rule properties, and (iii) the proposed ranking defined through the operation of greedy algorithm for rule induction. The three rankings were confronted and compared from the perspective of their usefulness for the selection of rules performed in the two directions and with two strategies for rule selection. The resulting sets of rules were analysed with respect to the properties of the constituent decision rules and from the point of performance for all constructed rule-based classifiers. Substantial experiments were carried out in the stylometric domain, treating the task of authorship attribution as classification. The results obtained indicate that for all three rankings and search paths it was possible to obtain a noticeable reduction of attributes while at least maintaining the power of inducers, at the same time improving characteristics of rule sets.

I. Introduction

ONE OF the main goals of data mining is the extraction of useful knowledge from large amounts of data or phenomena described by a high number of attributes. An important element of this process is the determination and selection of the most important attributes related to the described phenomenon [1]. The objective of this step, called feature selection, is to differentiate relevant variables from the entire set of features, while at the same time preserving the descriptive and representative qualities of the original set of attributes [2].

Feature selection can be accomplished by selecting a minimal subset of features that enables obtaining at least the same performance of a classifier as for the entire set of attributes [3]. In this case, feature subset selection requires assessing the quality of each discovered feature subset. Another way of proceeding is to construct a ranking of features based on a specific criterion. Then the variables are ordered from the most to the least important, and the top k features are selected based on a predefined threshold. Feature ranking is also known as feature weighting and involves evaluating individual attributes by assigning weights to them based on their relevance.

A technique used to search space of variables during the attribute selection process is an important factor. Since the problem of locating an optimal subset of features, taking into account all possible variable subsets, is NP-hard, greedy techniques, such as forward selection and backward elimination, are often used instead of exhaustive search. Forward selection begins with an empty set, which is gradually expanded by adding one feature (or a group of features) at a time until specific criteria are met. Sequential backward elimination involves starting with all attributes and progressively discarding them. Depending on the adopted criterion, added or rejected attributes can correspond to the highest positions in the ranking, or, they can be the lowest ranking elements.

One of the disadvantages of sequential selection is that interactions among features are not closely studied and dependencies can be missed when only one path of selection is investigated [4]. This problem can be remedied to some extent by varying the feature selection approach through patterns discovered in the data, such as decision rules, and discarding them only when they are dependant entirely on rejected variables, while keeping under consideration those that refer also to at least one attribute that is contained in the retained set. With this kind of processing, interactions among variables have more influence on the properties of recalled sets of rules.

The aim of the research presented in the paper and its contribution is the investigation and comparison of three influential factors, as follows: (i) two search strategies, i.e. sequential forward selection vs. sequential backward elimination, applied not directly to the variables in the dataset but through the filtering decision rules process, (ii) two approaches to rule selection, i.e., retaining rules that contain conditions only on the variables still in considerations vs. keeping the rules that include conditions on at least one of the attributes contained in the studied set, (iii) three ranking mechanisms, the OneR available in WEKA workbench [5], and two proposed, exploiting the properties of data and patterns discovered in them. One of those referred to the defined weighting factor, which takes into account the number of reducts in which a given attribute occurs and the cardinalities of these reducts [6], while the other was based on the properties of the greedy algorithm for the induction of decision rules, the number of occurrences in the rules and their support.

All experiments were performed on two datasets from the stylometry domain. The writing styles of the considered writers were learnt from available texts through the analysis of quantitative linguistic descriptors and advanced processing. To prevent bias on the observations, the datasets were prepared for the task of binary authorship attribution with balanced classes. The performance for induced rule-based classifiers was estimated with the help of test sets, over which the classification accuracy was averaged.

The results obtained allowed to conclude that all search paths led to increased performance for reduced sets of features while improving the characteristics of constructed rule sets. Backward elimination with keeping the rules referring to any attributes in the considered set allowed for reduction of more attributes than forward selection with limiting conditions in rules only to still present variables. The three investigated rankings produced close maximal predictions but for different numbers of attributes and rules. Greedy ranking held its ground when pitted against the other two, it even led to the one case of perfect recognition. These observations proved the merits of the described research works and again validated the methodology for ranking-driven rule selection.

The structure of the paper is organised as follows. Section II presents background information related to feature selection and induction of decision rules. Section III provides a description of stylometric analysis of texts, as the application domain. Section IV contains the explanation for the experiments performed and comments on the results obtained. Conclusions and future research plans are given in Section V.

II. Background information

In this section, aspects related to feature selection and decision rules are provided. Search strategies were described in the context of feature selection, and the main approaches for induction of rules were presented. Finally, the processing steps of rule filtering driven by feature selection were given.

A. Feature selection

During recent years, due to increasing demands for dimensionality reduction, extensive efforts in feature selection research have been made. It can be realised as a stage of data mining, related to data pre-processing, and then it affects such elements as visualisation, learning algorithms, and performance of classifiers. The main task of feature selection is to remove irrelevant or redundant variables so that their elimination from the set of attributes will not affect the performance of the learning algorithms [7]. The process of feature selection allows for data reduction and lowering of storage requirements. Furthermore, since the goal is to find the most relevant variables, it is possible to strive to improve data quality by enhancing data mining algorithms, that is, reducing learning time and improving predictive capabilities.

A feature selection procedure can be considered to contain three stages: (i) search for potential subsets of variables, (ii) evaluation of the subset of attributes based on some criteria, and (iii) setting the stop condition for the search. The final stage is closely linked to the initial one, as the search is repeated iteratively until the stopping criterion is met.

Due to the large search space, feature selection is also perceived as a combinatorial problem—for a dataset with N attributes, the search space is 2^N. Searching for an optimal subset of features taking into account all possible variable subsets is NP-hard problem [8]. An exhaustive search can be performed only if the number of attributes is relatively small. Instead, greedy [9] or meta-heuristics [10] approaches can be used.

To select a subset of variables from the input data, different search strategies can also be applied, including genetic algorithms, evolutionary computation techniques, heuristic search algorithms, and various hybrid strategies. Among greedy techniques, the sequential search performed as forward selection and backward elimination can be distinguished [11]. The sequential backward elimination method starts with all the variables, and then gradually features are removed from the set, either one by one, or in groups. In each step, the eliminated variable or variables contribute the least to the criterion function. Forward selection starts with the empty set to which sequentially features are added, again either one at a time or in groups, until certain criteria are met.

Both search strategies are heuristic and cannot guarantee the optimality of the selected features. Among the alternatives to these approaches, floating, branch-and-bound, and randomised can be mentioned [12]. Random search methods, for example, genetic algorithms, add some randomness to the search procedure to help escape from a local optimum. In certain cases, especially when dealing with high-dimensional datasets, an individual search is performed. Such methods evaluate each feature individually based on a specific criterion or condition. The branch-and-bound algorithm finds the optimal feature subset if the criterion function used is monotonic [3]. Floating search methods prevent the situation where the variable is deleted in backward elimination, and then it cannot be reselected, and also when a feature is added in forward selection and cannot be deleted once it was selected [11].

B. Ranking construction

Feature selection can be performed in two different ways, by selecting a subset of attributes or by creating a ranking of variables [13]. In the latter case, the variables are ordered according to the adopted criterion or evaluation function from the most important to the least important and the top k attributes are selected from the ranking, with k being some pre-selected threshold number. Feature ranking plays an important role in directing the search process in different machine learning tasks, especially when an exhaustive search is computationally unfeasible and a heuristic search approach is necessary. It determines the order in which the variables are explored by the algorithms within the feature space.

Feature ranking methods use different measures, for example, based on similarity score, statistics, information theory, or on some functions of the classifier's outputs [1]. Traditional ranking approaches evaluate variables without incorporating any learning algorithm. This category typically consists of filter-based feature selection methods, such as referring to information gain, correlation, or Relief algorithm. However, there are also some studies on wrapper techniques, which involve methods such as recursive feature elimination [14], and the classifier-aided feature ranking approach [15].

In the paper, three ranking mechanisms were studied, related to the properties of the data, and discovered patterns in the form of decision reducts and decision rules. One ranking was based on the defined weighting factor calculated through reducts, another was related to the OneR algorithm, and the third ranking was proposed by the authors and based on the properties of the greedy algorithm for rule induction. All the rankings obtained were used as filters for sets of induced rules.

1) Ranking of attributes based on reducts: Reduct is one of the key notions in rough sets theory [16] and refers to feature selection performed within the framework of rough sets. There are many definitions of a reduct because they deal with different criteria related to the selection of attributes and computing the most relevant sets of variables, for example, decision and local reducts for decision tables, reducts for information systems, reducts based on the generalised decision, or fuzzy decision reducts.

A reduct can be defined as a minimal set of attributes that preserves the degree of dependency of the entire set of attributes. Taking into account the performance, the reduct is such a minimal subset of attributes that has the same classification power as the complete set of available attributes [17].

The problem of calculating reducts is NP-hard, therefore, different heuristic approaches are used for its construction, for example finding reducts through sampling data from a decision table [18], heuristics based on discernibility matrix [19], greedy algorithms [9], Boolean reasoning, and many others [20]. In the investigation presented in the paper, the genetic algorithm [21], implemented in the Rough Sets Exploration System (RSES) [22], was used to construct the reducts. It is a binary genetic algorithm where every binary individual encodes one subset of attributes that is a potential reduct. The fitness function of a subset R has the form:

$$F(R) = \frac{n - L_R}{n} + \frac{2C_R}{m^2 - m}, \quad (1)$$

where n is the length of bit strings equal to a number of attributes, and m gives a number of objects. L_R denotes a number of "1"-s in the subset R, and C_R denotes the number of object pairs (with different decision values) discerned by the attribute subset R. Calculating C_R is the most time-consuming operation. It is accelerated by the "distinction table", a binary matrix of size $(n{+}1)\times(m^2{-}m)/2$. Each column corresponds to one attribute (the last column corresponds to the decision), and each row corresponds to one pair of different objects. The value "1" denotes an attribute with a different value on the pair of objects. Finding a reduct means finding the minimal subset of columns that cover the matrix.

The described genetic algorithm allows to generate a satisfactorily high number of reducts in relatively short time. The resulting reducts may contain different attributes and may also have different cardinalities. For the set of induced reducts, the weighting factor for features was proposed that takes into account the number of reducts in which a given attribute exists, and cardinalities of these reducts [6],

$$W_F(G_{Red}, a) = \sum_{i=k_{min}}^{k_{max}} \frac{card\,(RED(G_{Red}, a, i))}{card\,(G_{Red}) \cdot i}, \quad (2)$$

where k_{min} and k_{max} are respectively the minimal and the maximal reduct cardinalities detected for the group G_{Red}. $RED(G_{Red}, a)$ denotes the set of all reducts from the group G_{Red} that include the attribute a, and $RED(G_{Red}, a, k)$ is the set of reducts of length k that contain the attribute a. Then $card\,(RED(G_{Red}, a, k))$ returns for the group G_{Red} the number of reducts with specific length equal to k that contain the given attribute a. The values of W_F range from 0 (the attribute a is not included in any of the reducts in this group) to $1/k_{min}$, when the attribute is included in all the reducts and all the reducts have the same cardinality (then $k_{min} = k_{max}$).

A higher value of the weighting factor presented indicates that the attribute appears in more reducts with lower cardinalities, and low values of W_F are obtained for attributes that are included in fewer reducts containing more variables. All attributes included in a group can be ordered by the scores calculated for them, and a ranking is obtained as a result.

The described weighting factor promotes reducts with a small number of attributes. This way of reasoning follows from the fact that in a situation where we have two reducts and one of them has a smaller number of attributes, according to the definition of a reduct, this smaller number of attributes is sufficient to protect the performance of the system. Moreover, it complies with the Minimum Description Length principle [23]: "the best hypothesis for a given set of data is the one that leads to the largest compression of data". Additionally, reducts with smaller numbers of attributes are preferred from a knowledge representation perspective.

2) OneR algorithm: The OneR (One Rule) algorithm is a simple classification algorithm that is used in the field of machine learning. Its purpose is to select the most conclusive feature from all available features in the dataset, in order to create a simple classification model. This is done by calculating the number of occurrences of particular class labels for each value of a given attribute in the dataset. After this process, the OneR algorithm selects the feature for which the value is the most discriminating in the context of predicting class labels. In practice, for the selected feature, a single condition is created in a decision rule that is used to classify new instances. The algorithm generates one rule per unique attribute value of the selected best feature.

The main strength of the OneR algorithm is its ability to select the most relevant feature in the context of class prediction [24]. Although the OneR algorithm is simple and does not take into account interdependencies between features, it often allows to obtain satisfactory classification accuracy. In addition, this algorithm tends to choose the value of attribute that occurs the most frequently, and in this way it allows to ignore noise existing in the data. OneR is also called one-level decision tree algorithm. It selects attributes from a dataset one by one and generates a different set of rules based on the error rate from the training set. Finally, it chooses the attribute that offers rules with minimum error [25].

3) Ranking of attributes based on greedy algorithm properties: In the research, the authors propose a ranking mechanism exploiting the properties of the greedy algorithm for the induction of decision rules [26]. Such an algorithm constructs a decision rule for each row of a decision table. In each iteration, attributes are selected to form the conditions of the rules. The selected attribute separates the maximum number of rows from a set of rows with a different class label, so a decision table is divided into sub-tables as dictated by given attribute and corresponding value. The partitioning of a table is completed when all rows in the sub-table, corresponding to the selected attribute, have the same class labels.

As shown in previous research [27], given certain assumptions about the NP class, the greedy algorithm used to induce decision rules produces results that are not far from the best approximate polynomial algorithms for minimising the length of the rules, which is important for knowledge representation. Short rules can be considered as more general so they allow to reflect patterns hidden in the data and prevent overfitting, which is important for the classification process.

During research focused on the greedy algorithm, it was observed that in the majority of cases, when constructing decision rules, the greedy algorithm at each iteration selects an attribute that separates at least 50% of the remaining rows with different decisions.

The proposed ranking was based on the attributes contained in the decision rules, the percentage of separated rows with decisions different from the decision attached to a given rule, and the support of the rule. The latter element is an important factor in assessing the quality of decision rules. In order to construct the ranking, the decision rules were induced by the greedy algorithm and duplicate rules were removed from the entire set of rules. Then, for each attribute, the number of its occurrences in the rules was determined, assigning the highest positions in the ranking to the attributes with the highest number of occurrences. If the number of occurrences was the same for several attributes, then the percentage of rows separated by the given attribute was taken into account. The third factor that played a role in determining the score for each attribute was the support of the rule in which the attribute appeared, which led to the assignment of higher positions in the ranking to attributes from the rules with higher support.

C. Decision rules

Decision rules belong to popular forms used for data representation. They are induced from datasets very often presented as a decision table $T = (U, A \bigcup \{d\})$ [16], where U is a non-empty, finite set of objects, $A = \{a_1, \dots, a_m\}$ is a set of condition 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, and $d \notin A$ is a distinguished attribute called a decision, with values $V_d = \{d_1, \dots, d_{|V_d|}\}$. The decision rules take the form:

$$(a_{i_1} = v_1) \wedge \dots \wedge (a_{i_k} = v_k) \to d = v_d,$$

where $a_{i_1}, \dots, a_{i_k} \in \{a_1, \dots, a_m\}$, $v_i \in V_{a_i}$, and $v_d \in V_d$. Pairs $(a_{i_1} = v_1)$ are called descriptors or conditions. The number of conditions in a premise part of a rule is its length. Short rules are preferred from the point of view of knowledge representation and with regard to the MDL principle. They are easier to understand and interpret. When assessing the quality of decision rules, support is another important factor. It is a number of such objects from the decision table whose attribute values satisfy the premise part of the rule, and they have the same decision as the one attached to the rule. This measure allows to discover major patterns present in the data.

There are a wide variety of approaches for induction of decision rules. Among the exact ones, Boolean reasoning and extensions of dynamic programming should be mentioned [28]. The construction of decision rules with maximum support or minimum length is considered an NP-hard problem, so different heuristics are used. They are based on modifications of exact approaches, different kinds of greedy algorithms, methods relying on sequential covering, genetic algorithms, and many others. In the rough set theory, the popular approach is also induction of rules based on a reduct. Then each rule has length equal to the cardinality of the reduct, and each object from a decision table has assigned values corresponding to condition attributes included only in this reduct.

Apart from using decision rules as a form of knowledge representation, they are very often used as classifiers. In this situation, the rule filtering process can be treated as a method of pruning the rule set to fine-tune the classifier by reducing the number of rules. The use of filtering rules in the framework of the feature selection process often leads to improved classification accuracy.

In the experiments performed, the decision rules were induced by the exhaustive algorithm implemented in the RSES system. It constructs all minimal decision rules, i.e. rules with minimal numbers of descriptors (pairs attribute = value) in their premise parts. Then, they were filtered sequentially, according to the search strategy added or removed, driven by the studied rankings of attributes.

III. Stylometric data

A writing style is an individual characteristic, based to some extent on social and cultural background, education, lifetime experiences, elements that are learnt, but also on personal linguistic preferences and habits. To obtain a definition of an authorial profile, access to some representative samples of writing is needed. Comparative analysis and stylometric data mining lead to the discovery of patterns specific to writers and the construction of approximating descriptions that can be applied to text samples of unknown or unconfirmed authorship to find the closest match. This way of carrying out the authorship attribution task means solving a classification problem [29], therefore, a dataset to be prepared needs to include some training and test samples, all relying on a set of selected efficient style-markers [30].

Stylometric descriptors that work best refer to common language elements as they are used almost subconsciously, so they are less prone to forgery or imitation. Lexical and syntactic markers are often employed for the task [31]. They provide quantitative characteristics through frequency of occurrence for function words and punctuation marks, which results in real-valued features. In the experiments reported, the set of markers contained 24 elements with values calculated over text samples obtained by partitioning long novels by four acclaimed writers into smaller chunks. The authors studied, Edith Wharton, Mary Johnston, Jack London, and James Oliver Curwood, were paired according to gender [32], in order to form two datasets with binary authorship attribution.

The division of long texts into smaller parts resulted in imposing a specific stratification of the input space [33]. To avoid bias when evaluating the performance of a classifier, the datasets (the male writer dataset and the female writer dataset) prepared included one train set and two test sets. The samples contained in sets of different types were based on separate novels. With binary classification, balanced data and the same importance of all classes, classification accuracy was used as a measure of performance, providing information on the average portion of correctly attributed text samples from test sets.

Among popular data mining approaches, those that involve induction of decision rules belong to the most advantageous. They not only enable assigning authors to samples, but also enhance understanding of the stylometric domain by providing an inside view on linguistic patterns detected for authors by the transparent form of discovered rules. Short rules, with a few conditions in their premises, are preferred over long rules [24]. The former are more general, while the latter with their too detailed definitions can cause over-fitting.

The datasets were discretised with the Fayyad and Irani algorithm [34]. It is one of the top-down supervised methods, which starts with assigning one large interval to represent in the discrete domain all the values of a transformed variable. Then, referring to the MDL principle and calculation of entropy [23], candidates for cut-points are evaluated to discover which are most supportive to distinction of classes. If further partitioning is disadvantageous to entropy, the processing stops. As a consequence, it is possible that some variables are removed from consideration in the discrete domain when they have a single categorical representation. In the experiments, for the female writer dataset 20 out of the total of 24 features received more than a single bin, and for the male writer dataset the set of attributes was reduced to 22.

IV. Performed experiments

The experimental process of the research works consisted of the following stages:

- Preparation of two datasets (female writers and male writers), which included discretisation by Fayyad and Irani algorithm applied to all condition attributes;
- Construction of three rankings of attributes:
  - *Reducts*—based on reducts and the proposed weighting factor;
    Using a genetic algorithm implemented in the RSES system, one group of 150 reducts was generated. Obtained reducts consisted of different attributes from the whole set of available features and had different cardinalities. The weighting factor defined in Eq. (2) took into account all these elements and returned scores for the variables. The ordering of attributes by their scores resulted in the ranking.
  - *OneR*—based on the OneR algorithm implemented in WEKA software [5];
  - *Greedy*—based on the properties of the greedy algorithm for induction of decision rules, that is, the number of rules in which a given attribute occurs, the percentage of separated rows with different decisions and the support of decision rules.
- Induction of decision rules by exhaustive algorithm, for the input datasets;
- Filtration of sets of rules accordingly to sequential forward selection and sequential backward elimination driven by attributes included in a given ranking;
- Evaluation of performance for rule-based classifiers with test sets;
- Assessment of the quality of rule sets from the point of view of knowledge representation, i.e., taking into account the number of rules, average length and average support;
- Comparative study of results, for two search directions, two rule selection strategies, and three rankings.

Details of all steps are provided below, along with comments on the results obtained.

A. Rankings

For the female and male writer datasets, the rankings obtained were presented in rows of Table I (where the letters F and M indicate the female and male writer datasets, respectively). The row *Position* denotes the position of the given attribute in a ranking, and 1 is considered the highest ranking position, assigned to the most important feature.

For the female writer dataset, in the case of the ranking constructed through reducts and the OneR algorithm, the entire set of attributes was used, with the exception of attr14, attr16, attr18, and attr21. It resulted from the situation that these attributes had only 1 bin allocated by the supervised discretisation process. For the male writer dataset, there was a similar situation, i.e. instead of 24 attributes, only 22 were used to create the ranking since attr11 and attr14 were the

Table I
RANKINGS OF ATTRIBUTES FOR FEMALE AND MALE WRITERS

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Reducts-F	attr23	attr1	attr22	attr17	attr12	attr3	attr2	attr10	attr9	attr5	attr6	attr8	attr4	attr13	attr7	attr20	attr11	attr15	attr0	attr19		
Reducts-M	attr23	attr6	attr3	attr7	attr5	attr20	attr16	attr1	attr15	attr8	attr9	attr17	attr22	attr4	attr12	attr0	attr18	attr21	attr13	attr10	attr2	attr19
OneR-F	attr23	attr1	attr17	attr22	attr20	attr2	attr13	attr6	attr8	attr4	attr9	attr12	attr3	attr7	attr10	attr15	attr11	attr0	attr19	attr5		
OneR-M	attr23	attr17	attr3	attr1	attr6	attr16	attr13	attr0	attr18	attr9	attr7	attr2	attr8	attr22	attr12	attr5	attr4	attr20	attr21	attr15	attr19	attr10
Greedy-F	attr1	attr23	attr2	attr17	attr3	attr13	attr11	attr22	attr10	attr8	attr6	attr0	attr19	attr5	attr7	attr4						
Greedy-M	attr23	attr1	attr3	attr0	attr21	attr10	attr16	attr18	attr8	attr2	attr22	attr7	attr6	attr12	attr19	attr15	attr9					

attributes to which only 1 bin was assigned in the Fayyad and Irani discretisation process.

From the ranking created based on the accuracy of the greedy algorithm, in addition to the attributes with a single categorical representation, other variables were also excluded because they did not appear in the induced decision rules. Therefore, these rankings were shorter and contained 16 attributes for the female dataset and 17 for the male dataset.

It is worth noting that for the male writer dataset, all three rankings assigned the highest position to the same attribute: attr23. In the case of the female set, this attribute was ranked second only in the ranking related to the greedy algorithm. Furthermore, the features disregarded by the greedy ranking (attr9, attr12, attr15, and attr20 for female writers, and attr4, attr5, attr13, attr17, attr20 for male writers) were not recognised as irrelevant or close to irrelevant by other rankings, for example, for the male writers attr17 was found as the second ranking for the OneR algorithm.

B. Strategies employed in decision rule filtering

Forward selection was performed by sequentially filtering and increasing the set of decision rules. Starting with the highest ranking attribute, from the entire set of rules those were selected that contained conditions (in their premises) relating only to this attribute. Then, in the second step, a subset of recalled attributes was extended to the top two positions, and such rules were selected that relied only on these two variables as conditions. Next, three top ranking features were studied, and so on. In each step of the sequential search, the conditions in the rules were limited only to the currently selected subset. The forward rule filtering process continued until all available features and rules were included in the set considered.

The backward elimination was achieved by sequentially decreasing the set of decision rules. Starting with the attribute in the lowest ranking position, those rules were selected from the entire set of rules, which contained in their premises the condition referring to this very attribute. If a rule included some other attributes that worked as conditions, then that rule was not removed from the set of rules. The second step of backward reduction meant rejection of rules with conditions limited to the two lowest ranking variables, and so on, until the set of rules was exhausted. The difference between the two strategies involved is shown in the illustrative small example.

Let us assume a set of five condition attributes, for simplicity ranked as follows, where 1 is considered the top ranking position, and 5 the bottom of the ranking:

Position 1: attr1
Position 2: attr2
Position 3: attr3
Position 4: attr4
Position 5: attr5

The set of rules, subject to filtering driven by ranking, consists of eight elements.

Rule 1: with condition on attr1
Rule 2: with conditions on attr2 and attr3
Rule 3: with conditions on attr1 and attr5
Rule 4: with conditions on attr1 and attr4
Rule 5: with condition on attr3
Rule 6: with conditions on attr3 and attr5
Rule 7: with conditions on attr2 and attr5
Rule 8: with condition on attr4

For backward elimination, the processing starts with all rules included in the recalled set. Then, for the filtering steps, the resulting sets are as follows.

Step 1: Rejected attributes: attr5, recalled rules: all rules
Step 2: Rejected attributes: attr5, attr4, recalled rules: 1, 2, 3, 4, 5, 6, 7
Step 3: Rejected attributes: attr5, attr4, attr3, recalled rules: 1, 2, 3, 4, 7
Step 4: Rejected attributes: attr5, attr4, attr3, attr2, recalled rules: 1, 3, 4
Step 5: Rejected attributes: all, recalled rules: no rules

For forward selection at the starting point, the set of recalled rules is empty. It is next gradually expanded as listed below.

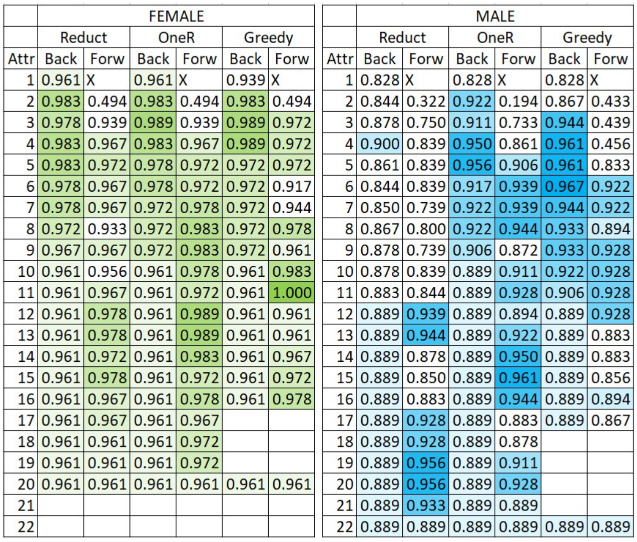
Step 1: Selected attributes: attr1, recalled rules: 1
Step 2: Selected attributes: attr1, attr2, recalled rules: 1
Step 3: Selected attributes: attr1, attr2, attr3, recalled rules: 1, 2, 5
Step 4: Selected attributes: attr1, attr2, attr3, attr4, recalled rules: 1, 2, 4, 5, 8
Step 5: Selected attributes: all, recalled rules: all rules

The process of rule filtering carried out for the greedy rankings was slightly different than for the other two rankings, because the rule sets induced by the exhaustive algorithm included rules with conditions on such features that were absent in the greedy ranking. Therefore, the first step of rule elimination was to remove rules containing only attributes that did not appear in these rankings, while the last step for forward selection was to add these rules.

C. Performance of rule-based classifiers

For all rule-based classifiers obtained in the decision rule filtering process, performance was evaluated with test sets. Fig. 1 presents the average classification accuracy obtained.

Decision rules induced by the exhaustive algorithm were selected through the backward elimination (column Back) and forward search (column Forw) strategies, with the conditions for recalling rules governed by the three rankings (groups of columns, Reduct, OneR, Greedy). The results shown in the bottom row provide the reference point, because they correspond to the case where the entire sets of attributes and rules were taken into account. The X mark denotes the situation where no rules were included in the set of recalled rules. The coloured cells indicate where the classification accuracy exceeded the reference point. The intensity of cell colour depends on how much the accuracy was improved. For each step of the rule filtering process, the columns Attr indicate the ranking position considered (which for forward search corresponds to the number of variables taken into account).



	FEMALE					
	Reduct		OneR		Greedy	
Attr	Back	Forw	Back	Forw	Back	Forw
1	0.961	X	0.961	X	0.939	X
2	0.983	0.494	0.983	0.494	0.983	0.494
3	0.978	0.939	0.989	0.939	0.989	0.972
4	0.983	0.967	0.983	0.967	0.989	0.972
5	0.983	0.972	0.978	0.972	0.972	0.972
6	0.978	0.967	0.978	0.972	0.972	0.917
7	0.978	0.967	0.972	0.978	0.972	0.944
8	0.972	0.933	0.972	0.983	0.972	0.978
9	0.967	0.967	0.972	0.983	0.972	0.961
10	0.961	0.956	0.961	0.978	0.961	0.983
11	0.961	0.967	0.961	0.972	0.961	1.000
12	0.961	0.978	0.961	0.989	0.961	0.961
13	0.961	0.978	0.961	0.989	0.961	0.961
14	0.961	0.972	0.961	0.983	0.961	0.967
15	0.961	0.978	0.961	0.972	0.961	0.972
16	0.961	0.967	0.961	0.978	0.961	0.978
17	0.961	0.967	0.961	0.967		
18	0.961	0.961	0.961	0.972		
19	0.961	0.961	0.961	0.972		
20	0.961	0.961	0.961	0.961	0.961	0.961
21						
22						

	MALE					
	Reduct		OneR		Greedy	
Attr	Back	Forw	Back	Forw	Back	Forw
1	0.828	X	0.828	X	0.828	X
2	0.844	0.322	0.922	0.194	0.867	0.433
3	0.878	0.750	0.911	0.733	0.944	0.439
4	0.900	0.839	0.950	0.861	0.961	0.456
5	0.861	0.839	0.956	0.906	0.961	0.833
6	0.844	0.839	0.917	0.939	0.967	0.922
7	0.850	0.739	0.922	0.939	0.944	0.922
8	0.867	0.800	0.922	0.944	0.933	0.894
9	0.878	0.739	0.906	0.872	0.933	0.928
10	0.878	0.839	0.889	0.911	0.922	0.928
11	0.883	0.844	0.889	0.928	0.906	0.928
12	0.889	0.939	0.889	0.894	0.889	0.928
13	0.889	0.944	0.889	0.922	0.889	0.883
14	0.889	0.878	0.889	0.950	0.889	0.883
15	0.889	0.850	0.889	0.961	0.889	0.856
16	0.889	0.883	0.889	0.944	0.889	0.894
17	0.889	0.928	0.889	0.883	0.889	0.867
18	0.889	0.928	0.889	0.878		
19	0.889	0.956	0.889	0.911		
20	0.889	0.956	0.889	0.928		
21	0.889	0.933	0.889	0.889		
22	0.889	0.889	0.889	0.889	0.889	0.889

Figure 1. Accuracy of rule-based classifiers, for female and male writers

As can be observed in Fig. 1, in the case of the Greedy column and the backward elimination strategy, rejecting rules only with attributes not included in the ranking resulted in the same classification accuracy as for the entire set of attributes. For forward selection, the results for the last step of selecting features included in the rankings differed slightly from the ones given in the bottom row, after adding the rules with attributes that did not appear in this ranking. For female writers, a small improvement was noted, and for male writers, a small decrease was visible.

When the backward elimination strategy was combined with the Reduct and OneR rankings and applied to the female writer dataset, it should be noted that for all ranking positions considered the classification accuracy was always at least at the reference level, even in the last step of filtering for the attribute in the first position in the rankings. The highest value of the classification accuracy of 0.989 was obtained for the Greedy and OneR rankings, and was related to the third position in these rankings. For ranking based on reducts this value was slightly smaller (0.983) and happened in processing of the fourth position in the ranking.

In the case of forward selection executed for the female writer dataset, the highest possible classification quality equal to 1.0 existed for 11 attributes placed at top positions in the Greedy ranking. For the OneR algorithm the maximum was equal to 0.989 and for the Reduct ranking 0.978, and both were detected when the twelfth positions were processed.

For the male writer dataset for the top position in the three rankings, backward elimination obviously returned the same results. The highest classification accuracy of 0.967 was obtained for the Greedy ranking related to the sixth top position in the ranking. It was also the highest improvement noted for this dataset. Apart from the top two ranking positions, for all the rest of filtering steps, the classification was either the same or improved over the reference point. For the OneR ranker, the best performance (0.956) referenced the fifth top ranking position. With the exception of the top ranking position, for the OneR ranking in the entire rule filtering path, the reported performance was at least as good as for the entire sets of rules and attributes considered. The ranking based on reducts brought the worst results among the three rankings, however, even here they were still detected cases of maintaining or increasing performance for the reduced sets of rules.

In the forward search applied to the male dataset, the OneR ranking was most advantageous: for the fifteenth ranking position the maximal classification accuracy 0.961 was recorded. The second best level of predictions (0.956) was obtained for the nineteenth position of the Reduct ranking. The Greedy algorithm came last with the highest accuracy of 0.928, however, it resulted from processing the ninth ranking position, so more decision rules and features were discarded than for the other two cases.

D. *Characteristic of rule-based models*

The entire process of rule filtering driven by rankings involved two search directions, two strategies for rule selection, and three rankings. For all the sets and subsets of decision rules constructed, their characteristics were observed, as shown in Table II. These observations included the number of rules (NoR column), average rule length (Len column), and average rule support (Supp column). The column Attr points to the ranking position considered.

As could be expected, analysis of the rule sets showed that as the number of rules in the set decreased, their average lengths tended to decrease, and the average supports increased. This was particularly evident in the rows at the bottom or close to the bottom of the tables. The average values relating to the shortest rules with the highest support were marked in bold.

In the case of forward selection, the differences regarding the number of rules, their length, and support were more visible than in the case of backward elimination. It was due to the nature of how the strategies employed for rule selection in each case worked, as they were not the same.

With forward as a search direction, the processing started with the empty set of rules and then, gradually, in each step some recalled rules were added. These rules could include

Table II
CHARACTERISTICS OF RULE SETS WITH FILTERING DRIVEN BY RANKINGS

	Reducts - Female						OneR - Female						Greedy - Female					
	Back			Forw			Back			Forw			Back			Forw		
Attr	NoR	Len	Supp	NoR	Len	Supp	NoR	Len	Supp	NoR	Len	Supp	NoR	Len	Supp	NoR	Len	Supp
20	4121	4.8	6.6	4121	4.8	6.6	4121	4.8	6.6	4121	4.8	6.6	4121	4.8	6.6	4121	4.8	6.6
19	4120	4.8	6.6	3830	4.7	6.8	4120	4.8	6.6	3935	4.8	6.4						
18	4119	4.8	6.6	3660	4.7	6.7	4119	4.8	6.6	3644	4.7	6.6						
17	4119	4.8	6.6	2557	4.5	7.6	4118	4.8	6.6	3491	4.7	6.5						
16	4118	4.8	6.6	2071	4.4	8.1	4118	4.8	6.6	2804	4.5	6.9	4117	4.8	6.6	805	4.0	10.2
15	4117	4.8	6.6	1608	4.4	8.1	4117	4.8	6.6	1957	4.4	7.8	4113	4.8	6.6	548	3.8	11.2
14	4112	4.8	6.6	1204	4.2	8.7	4116	4.8	6.6	1412	4.2	8.5	4101	4.8	6.6	414	3.7	11.7
13	4107	4.8	6.6	902	4.0	9.2	4113	4.8	6.6	1097	4.0	8.8	4097	4.8	6.6	387	3.6	11.2
12	4097	4.8	6.6	624	3.9	9.6	4109	4.8	6.6	688	3.8	10.1	4093	4.8	6.6	369	3.6	11.4
11	4077	4.8	6.6	369	3.6	11.0	4102	4.8	6.6	442	3.5	11.3	4091	4.8	6.6	346	3.6	11.0
10	4042	4.8	6.5	232	3.4	12.8	4080	4.8	6.5	276	3.3	13.2	4051	4.8	6.6	206	3.3	13.1
9	4033	4.8	6.5	215	3.3	12.4	4027	4.8	6.5	182	3.1	14.5	3985	4.8	6.5	127	2.9	15.2
8	3957	4.8	6.5	132	3.0	14.5	3914	4.8	6.5	114	2.9	16.5	3890	4.8	6.5	99	2.7	15.6
7	3860	4.9	6.5	97	2.8	15.4	3705	4.8	6.5	73	2.6	18.3	3719	4.9	6.3	68	2.5	15.8
6	3675	4.9	6.6	47	2.7	22.3	3445	4.8	6.6	56	2.6	19.0	3563	4.9	6.3	55	2.5	16.6
5	3365	4.9	6.5	30	2.5	27.5	3015	4.8	6.5	26	2.4	29.0	3252	4.8	6.3	42	2.5	16.8
4	2651	4.8	6.8	18	2.3	32.1	2651	4.8	6.8	18	2.3	32.1	2656	4.8	6.3	27	2.4	18.1
3	2351	4.8	6.6	10	2.2	**37.6**	1932	4.7	6.6	9	2.1	**36.4**	2382	4.8	6.0	16	2.3	16.5
2	1548	4.6	6.6	4	**2.0**	34.0	1548	4.6	6.6	4	**2.0**	34.0	1548	4.6	6.6	4	**2.0**	**34.0**
1	378	**3.9**	**7.2**	0	0.0	0.0	378	**3.9**	**7.2**	0	0.0	0.0	1224	4.8	6.3	0	0.0	0.0

	Reducts - Male						OneR - Male						Greedy - Male					
	Back			Forw			Back			Forw			Back			Forw		
Attr	NoR	Len	Supp	NoR	Len	Supp	NoR	Len	Supp	NoR	Len	Supp	NoR	Len	Supp	NoR	Len	Supp
22	15283	5.1	5.9	15283	5.1	5.9	15283	5.1	5.9	15283	5.1	5.9	15283	5.1	5.9	15283	5.1	5.9
21	15283	5.1	5.9	11850	5.0	6.2	15283	5.1	5.9	12597	5.0	6.1						
20	15282	5.1	5.9	10067	4.9	6.2	15283	5.1	5.9	9835	4.9	6.3						
19	15282	5.1	5.9	8520	4.8	6.4	15283	5.1	5.9	7565	4.8	6.8						
18	15281	5.1	5.9	6155	4.7	6.7	15282	5.1	5.9	6115	4.8	7.2						
17	15278	5.1	5.9	4780	4.6	7.2	15272	5.1	5.9	4224	4.6	7.8	15270	5.1	5.9	2574	4.6	7.6
16	15271	5.1	5.9	3777	4.6	7.1	15256	5.1	5.9	3094	4.6	8.3	15246	5.1	5.9	1768	4.4	8.0
15	15256	5.1	5.9	2537	4.4	7.7	15229	5.1	5.9	2086	4.4	9.3	15212	5.1	5.9	1270	4.2	9.1
14	15231	5.1	5.9	1625	4.2	8.5	15184	5.1	5.9	1247	4.3	10.6	15150	5.1	**5.9**	902	4.1	10.0
13	15188	5.1	5.9	1125	4.2	8.8	15129	5.1	5.9	931	4.2	11.4	15035	5.1	5.9	554	3.9	11.2
12	15133	5.1	5.8	789	4.1	9.9	15023	5.1	5.9	609	4.0	12.5	14826	5.1	5.8	333	3.7	11.1
11	14972	5.1	5.8	486	4.0	10.1	14974	5.1	5.9	517	3.9	12.6	14549	5.1	5.8	210	3.5	12.3
10	14791	5.1	5.8	307	3.8	11.8	14717	5.1	5.9	359	3.8	13.3	14074	5.1	5.8	157	3.5	14.5
9	14484	5.1	5.8	199	3.6	12.2	14403	5.1	5.8	221	3.6	15.4	13871	5.1	5.8	132	3.3	14.0
8	14118	5.1	5.8	121	3.5	15.3	14104	5.1	5.8	147	3.7	15.7	13188	5.1	5.7	87	3.1	14.2
7	13454	5.1	5.7	62	3.3	17.1	13562	5.1	5.8	69	3.3	21.5	12563	5.2	5.7	51	3.0	17.1
6	12952	5.1	5.7	33	2.9	19.9	12543	5.1	**5.9**	42	3.2	22.3	11791	5.2	5.5	30	2.8	17.9
5	11807	5.2	5.7	18	2.6	28.2	11695	5.1	5.8	21	2.8	24.9	10844	5.2	5.6	23	2.9	18.2
4	9876	5.1	5.7	11	2.5	36.2	10043	5.1	5.6	9	2.1	37.9	9887	5.2	5.6	11	2.5	30.4
3	7492	5.1	**6.0**	5	2.2	50.8	7782	5.1	5.7	4	2.3	47.8	7485	5.2	5.4	7	2.0	**39.0**
2	5338	5.1	5.8	2	**2.0**	**61.5**	5522	5.1	5.5	1	**2.0**	**60.0**	5240	5.1	4.9	4	**1.8**	36.5
1	1596	**4.7**	4.3	0	0.0	0.0	1596	**4.7**	4.3	0	0.0	0.0	1596	**4.7**	4.3	0	0.0	0.0

conditions limited to the variables in the subset considered. In the first step only the top ranking attributes were taken into account, in the second step the top two were accepted, and so on. If a rule also contained conditions on other features (placed somewhere lower in the ranking), then it was not included in the recalled set. Therefore, always a ranking position that was processed directly gave the number of variables studied, and a subset of features present in the rules was explicitly visible.

The strategy applied in the backward elimination of decision rules started with the entire set of rules and then the groups of rules were gradually excluded, taking into account conditions on attributes from the lowest positions in a ranking. In this case, the assumption was that the eliminated rules should contain only attributes considered and discarded so far in the ranking. If a rule also included conditions on other features that were higher ranking, then such a rule was kept in the remaining set. This processing resulted in operation on higher numbers of rules for the same ranking position than when compared to the strategy applied in the forward selection. In fact, for each ranking position a set of rules recalled by forward selection was a subset of rules retained by backward elimination. It was especially striking in the case of the Greedy ranking and the number of rules obtained as characteristics for the constructed rule sets.

The advantage of this strategy was visible in the classification results, in particular for the female data set, where for almost every position of the ranking, the accuracy of rule-based classifiers was at least as good as the reference level considered for all variables from the set. Thus, this direction and the filtering rule strategy contributed to enhancing the power of the classifier. The drawback of such processing lies in keeping in considerations the higher numbers of attributes, and a lack of clear specification of their subset taken into account in each step. If there was a rule referring to all features, such rule would be kept to the very end, to the last step of filtering process, despite its significant length that indicates too close definition to be of any practical use, as the probability of exactly the same detailed pattern among test samples is low.

When the characteristics of the obtained rule sets were analysed, the classification accuracy of the constructed classifier was treated with extra attention. To provide the general look on the rule filtering process, over the entire run of feature and rule selection, the average performance was calculated, and the corresponding standard deviation (per sample), as shown in Table III. For both datasets and all rankings, backward elimination always brought better results than forward selection when combined with their strategies for rule selection. When the averaged performance is compared with the reference points of accuracy for the entire set of attributes available, it is clear that the backward search resulted in the improvement for all rankings for female writer dataset, while for male writers that was true for the OneR and Greedy rankings. For female writers the highest average classification accuracy was obtained for the Reduct-based ranking and for male writers for Greedy ranking. On the other hand, for female writers standard deviation reflected almost only direction and not ranking, yet for male writers the highest (but still rather small, only fractional) values were obtained for Greedy ranking.

Table III
SUMMARY OF OBTAINED ACCURACY OF RULE-BASED CLASSIFIERS

	Ranking and search direction					
	Reduct		OneR		Greedy	
	Back	Forw	Back	Forw	Back	Forw
	Female					
Average	0.989	0.939	0.968	0.949	0.968	0.937
St.dev.	0.01	0.11	0.01	0.11	0.01	0.12
	Male					
Average	0.877	0.840	0.899	0.870	0.910	0.817
St.dev.	0.02	0.14	0.03	0.16	0.04	0.18

The rule characteristics can be treated as dimensions in an optimisation space. Among them, the performance and the ranking position for which undiminished performance was reported could also be included. Such a summarising look was given in Table IV, where the lowest values are preferred for: ranking position, number of rules, average rule length. The highest values are preferred for: classification accuracy and average rule support. No overall Pareto points were detected, but for each dimension, some maxima and minima can be observed, or groups of characteristics could be analysed. When the same values of the observed criterion were recorded more than once, the occurrence for the highest ranking position was selected as the best, since it corresponded to the most extensive reduction, as long as it happened while the observed performance was not lower than the reference point.

As the forward selection strategy quantitatively enlarged the set of rules, it can be noted that these were moderately short rules with relatively large supports, and the number of such rules was small. For the female writer dataset, the highest value of average rule support was 37.6 for an average length of 2.2 and the number of rules equal to 10. For the male writer dataset, the highest value of average rule support was 61.5 for an average length of 2.2 and the number of rules equal to 2.

Table IV
SUMMARY OF OBTAINED BEST RESULTS

Optimality criterion	Female	Male
	Other characteristics	
Acc: **1.0** - F 0.967 - M	Ranking: Greedy, Pos: 11 Direction: Forw NrR: 346, AvgL: 3.6, AvgS: 11.0	Ranking: Greedy, Pos: 6 Direction: Back NrR: 11791, AvgL: 5.2, AvgS: 5.5
NrR: **16** - F 21 - M	Ranking: Greedy, Pos: 3 Direction: Forw Acc: 0.973, AvgL: 2.3, AvgS: 16.5	Ranking: OneR, Pos: 5 Direction: Forw Acc: 0.906, AvgL: 2.8, AvgS: 24.9
AvgL: **2.3** - F 2.8 - M	Ranking: Greedy, Pos: 3 Direction: Forw Acc: 0.972, NoR: 16, AvgS: 16.5	Ranking: OneR, Pos: 5 Direction: Forw Acc: 0.906, NoR: 21, AvgS: 24.9
AvgS: **32.1** - F 24.9 - M	Ranking: Reducts, OneR, Pos: 4 Direction: Forw Acc: 0.967, NoR: 18, AvgL: 2.3	Ranking: OneR, Pos:5 Direction: Forw Acc: 0.906, NoR: 21, AvgL: 2.8
Position: **1** - F 2 - M	Ranking: Reducts, OneR Direction: Back, Acc: 0.961 NoR: 378, AvgL: 3.9, AvgS: 7.2	Ranking: OneR Direction: Back, Acc: 0.922 NoR: 5522, AvgL: 5.1, AvgS: 5.5

In the case of the backward elimination strategy, the cut in the number of rules was generally smaller than for the forward search. The smallest reduction occurred for the Greedy ranking, for the female set. For the male set, the number of rules corresponding to the attribute in the highest ranking position was the same for all rankings, similarly the average rule length. Furthermore, for this dataset, the number of rules decreased about 10 times under this search strategy. For the Reduct and OneR rankings and female writers it was even greater.

The experiments carried out with varying search directions and strategies for rule selection enabled studying the effectiveness of the three rankings in the rule filtering process. The proposed Greedy ranking held its ground against the other two, leading to noticeably improved predictions for rule sets of decreased cardinalities, which is evidenced by the fact how often it led to the best results given in Table IV, and which clearly illustrates its merits.

V. CONCLUSIONS

The paper provides an illustrative example for the proposed research methodology dedicated to decision rule filtering governed by attribute rankings. The process of rule selection was executed with sequential backward reduction, where an entire set of induced rules is available at the beginning and then some elements from this set are discarded; and with sequential forward search, where the processing starts with the empty set to which recalled elements are added gradually. Along with two search directions, two strategies for rule selection were used, one with recalling rules including conditions only on variables from the currently considered subset, and the other with finding rules dependent on at least one of the attributes in the studied set.

In the investigations, three rankings of attributes were employed. The proposed ranking based on the percentage of separated rows and the properties of the greedy algorithm was confronted with the previously defined ranking referring to decision reducts, and the OneR ranker available in the popular WEKA environment. For the three rankings, the selection of rules was performed in the two directions, and the resulting rule sets were analysed with respect to the properties of constituent decision rules, such as their numbers, average length, and average support, but also from the point of evaluation of performance for all constructed rule-based classifiers when applied for labelling of samples from test sets.

The results from the experiments indicate that for all three rankings and search paths it was possible to obtain a noticeable reduction of attributes while at least maintaining the power of inducers, at the same time improving characteristics of rule sets. The special focus on Greedy ranking enabled to discover that it not only led to discarding some variables from the available sets, treating them as irrelevant, but also proved effective for rule filtering.

Future research will include application of the Greedy ranking in the feature selection process for other types of inducers, with different mathematical backgrounds and modes of operation. Also, the influence of discretisation step will be studied, as one of the factors greatly influencing representation of data and the patterns present in it.

Acknowledgment

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

References

[1] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, 2017.

[2] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Eds., *Feature Extraction: Foundations and Applications*, ser. Studies in Fuzziness and Soft Computing. Physica-Verlag, Springer, 2006, vol. 207.

[3] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1, pp. 245–271, 1997.

[4] U. Stańczyk, "Weighting of features by sequential selection," in *Feature Selection for Data and Pattern Recognition*, ser. Studies in Computational Intelligence, U. Stańczyk and L. Jain, Eds. Berlin, Germany: Springer-Verlag, 2015, vol. 584, pp. 71–90.

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

[6] B. Zielosko and U. Stańczyk, "Reduct-based ranking of attributes," in *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES-2020, Virtual Event, 16-18 September 2020*, ser. Procedia Computer Science, M. Cristani, C. Toro, C. Zanni-Merk, R. J. Howlett, and L. C. Jain, Eds., vol. 176. Elsevier, 2020, pp. 2576–2585.

[7] H. Liu and H. Motoda, *Computational Methods of Feature Selection*. CRC Press, 2007.

[8] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, no. 1, pp. 237–260, 1998.

[9] B. Zielosko and M. Piliszczuk, "Greedy algorithm for attribute reduction," *Fundam. Informaticae*, vol. 85, no. 1-4, pp. 549–561, 2008.

[10] M. M. Mafarja and S. Mirjalili, "Hybrid whale optimization algorithm with simulated annealing for feature selection," *Neurocomputing*, vol. 260, pp. 302–312, 2017.

[11] P. Pudil, J. Novovièová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[13] U. Stańczyk, B. Zielosko, and L. C. Jain, "Advances in feature selection for data and pattern recognition: An introduction," in *Advances in Feature Selection for Data and Pattern Recognition*, ser. Intelligent Systems Reference Library, U. Stańczyk, B. Zielosko, and L. C. Jain, Eds. Springer, 2018, vol. 138, pp. 1–9.

[14] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[15] W. Altidor, T. M. Khoshgoftaar, and J. V. Hulse, "An empirical study on wrapper-based feature ranking," in *2009 21st IEEE International Conference on Tools with Artificial Intelligence*, 2009, pp. 75–82.

[16] Z. Pawlak and A. Skowron, "Rough sets and boolean reasoning," *Information Sciences*, vol. 177, no. 1, pp. 41–73, 2007.

[17] A. Janusz and D. Ślęzak, "Utilization of attribute clustering methods for scalable computation of reducts from high-dimensional data," in *2012 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2012, pp. 295–302.

[18] Y. Yang, D. Chen, H. Wang, E. C. Tsang, and D. Zhang, "Fuzzy rough set based incremental attribute reduction from dynamic data with sample arriving," *Fuzzy Sets and Systems*, vol. 312, pp. 66–86, 2017.

[19] Y. Liu, L. Zheng, Y. Xiu, H. Yin, S. Zhao, X. Wang, H. Chen, and C. Li, "Discernibility matrix based incremental feature selection on fused decision tables," *International Journal of Approximate Reasoning*, vol. 118, pp. 1–26, 2020.

[20] J. Henzel, A. Janusz, M. Sikora, and D. Ślęzak, "On positive-correlation-promoting reducts," in *Rough Sets*, R. Bello, D. Miao, R. Falcon, M. Nakata, A. Rosete, and D. Ciucci, Eds. Springer International Publishing, 2020, pp. 213–221.

[21] J. Wróblewski, "Ensembles of classifiers based on approximate reducts," *Fundam. Informaticae*, vol. 47, no. 3–4, p. 351–360, 2001.

[22] J. Bazan and M. Szczuka, "The rough set exploration system," in *Transactions on Rough Sets III*, ser. Lecture Notes in Computer Science, J. F. Peters and A. Skowron, Eds. Berlin, Heidelberg: Springer, 2005, vol. 3400, pp. 37–56.

[23] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[24] R. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63–91, 1993.

[25] S. Ali and K. A. Smith, "On learning algorithm selection for classification," *Applied Soft Computing*, vol. 6, no. 2, pp. 119–138, 2006.

[26] M. J. Moshkov, M. Piliszczuk, and B. Zielosko, "Greedy algorithm for construction of partial association rules," *Fundam. Informaticae*, vol. 92, no. 3, pp. 259–277, 2009.

[27] ——, "On construction of partial reducts and irreducible partial decision rules," *Fundam. Informaticae*, vol. 75, no. 1-4, pp. 357–374, 2007.

[28] B. Zielosko, "Sequential optimization of γ-decision rules," in *Federated Conference on Computer Science and Information Systems - FedCSIS 2012, Wroclaw, Poland, 9-12 September 2012, Proceedings*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2012, pp. 339–346.

[29] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[30] M. Eder, "Style-markers in authorship attribution a cross-language study of the authorial fingerprint," *Studies in Polish Linguistics*, vol. 6, no. 1, pp. 99–114, 2011.

[31] H. Wu, Z. Zhang, and Q. Wu, "Exploring syntactic and semantic features for authorship attribution," *Applied Soft Computing*, vol. 111, p. 107815, 2021.

[32] S. G. Weidman and J. O'Sullivan, "The limits of distinctive words: Re-evaluating literature's gender marker debate," *Digital Scholarship in the Humanities*, vol. 33, pp. 374–390, 2018.

[33] U. Stańczyk and G. Baron, "On heterogeneity or sub-classes aspect in construction of stylometric input datasets," in *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES-2022, Verona, Italy, 7-9 September 2022*, ser. Procedia Computer Science, M. Cristani, C. Toro, C. Zanni-Merk, R. J. Howlett, and L. C. Jain, Eds. Elsevier, 2022, vol. 207, pp. 2526–2535.

[34] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuousvalued attributes for classification learning," in *13th International Joint Conference on Articial Intelligence*, vol. 2. Morgan Kaufmann Publishers, 1993, pp. 1022–1027.