Title: Reduct-based ranking of attributes

Author: Beata Zielosko, Urszula Stańczyk

Citation style: Zielosko Beata, Stańczyk Urszula. (2020). Reduct-based ranking of attributes. "Procedia Computer Science" (Vol. 176 (2020), s. 2576-2585), DOI: 10.1016/j.procs.2020.09.315

Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 176 (2020) 2576–2585
Procedia Computer Science
www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Reduct-based ranking of attributes

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

[a] Institute of Computer Science, University of Silesia in Katowice, Będzińska 39, 41-200 Sosnowiec, Poland
[b] Department of Graphics, Computer Vision and Digital Systems, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Akademicka 2A, 44-100 Gliwice, Poland

Abstract

The paper is dedicated to the area of feature selection, in particular a notion of attribute rankings that allow to estimate importance of variables. In the research presented for ranking construction a new weighting factor was defined, based on relative reducts. A reduct constitutes an embedded mechanism of feature selection, specific to rough set theory. The proposed factor takes into account the number of reducts in which a given attribute exists, as well as lengths of reducts. Two approaches for reduct generation were employed and compared, with search executed by a genetic algorithm. To validate the usefulness of the reduct-based rankings in the process of feature reduction, for gradually decreasing subsets of attributes, selected through rankings, sets of decision rules were induced in classical rough set approach. The performance of all rule classifiers was evaluated, and experimental results showed that the proposed rankings led to at least the same, or even increased classification accuracy for reduced sets of features than in the case of operating on the entire set of condition attributes. The experiments were performed on datasets from stylometry domain, with treating authorship attribution as a classification task, and stylometric descriptors as characteristic features defining writing styles.

Keywords: feature reduction; reduct; ranking of attributes; rough sets; decision rules; classification; stylometry

1. Introduction

Feature selection domain has been extensively studied by many researchers over the last years [13, 17, 29]. It has various applications in plenty of areas connected with data mining, machine learning, and knowledge representation [31]. The aim of feature selection is to distinguish relevant attributes among all available elements in the initial set, while maintaining descriptive and representative properties of the original features space. Reduction of irrelevant or redundant variables can be realised by a search for a minimum subset of features that satisfy some level of classification accuracy, or by using a ranking of attributes. In the former case, an algorithm can automatically determine the number of selected features. In the latter case, an algorithm retrieves k top-ranking features that meet some criterion.

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

10.1016/j.procs.2020.09.315

Generally feature selection approaches are grouped into three main categories, known as filter, wrapper, and embedded [10]. Filters are entirely independent on learning algorithms, wrappers exploit some evaluation methods based on learners. In the case of an embedded approach, evaluation mechanism is built-in directly into a learning phase.

In the research described, a filter method was combined with an embedded one, inherent to the rough sets theory (RST) [15, 23]. RST offers some tools for knowledge discovery from data, known as reducts [9, 30], which are based on computing the most relevant sets of features. From the classification point of view, a reduct can be interpreted as such minimal subset of condition attributes that has the same classification power as the entire set of features.

Methods for reducts construction constitute one of the main directions for studies in the rough sets theory [19, 20]. Unfortunately, the problem of finding a minimal reduct is NP-hard. It is also known that a potential number of all reducts that can be found for a given dataset with k attributes is equal to $N(k) = \binom{k}{k/2}$. These facts are the reasons for high computational costs and complexity, brought by a task of all reducts construction. They also give a motivation for inventing such heuristics, which allow to compute many reducts in some acceptable time. As a consequence, in the literature many types of reducts and algorithms for their generation can be found [21, 35].

In the research presented in the paper, so-called decision reducts were constructed by a genetic algorithm [33]. Obtained reducts contained not only different attributes, but also different numbers of attributes. Based on the constructed groups of reducts, a new weighting factor for features was proposed. This factor takes into account the number of reducts in which a given attribute exists, and lengths of these reducts. The scores calculated for all attributes were used to construct their rankings. Then, feature selection, driven by obtained rankings, was executed. For gradually decreasing subsets of selected attributes, decision rules were induced. For all resulting rule-based classifiers their performance was evaluated, and classification accuracy compared to the case of operating on the entire set of features.

Two ways for reduct generation were compared: (i) construction of many groups with a small number of reducts by running the algorithm multiple times independently, and (ii) construction of one group with a high number of reducts by running the algorithm one time. For ranking calculation three approaches were tested: (i) based on each small group of reducts, (ii) relating to a large group of reducts, and (iii) with combining small groups of reducts into a large one for which scores were calculated. The influence of a ranking construction mode on classification accuracy of resulting rule-based classifiers was observed.

The experiments were performed on datasets from stylometry domain [1, 16]. Induced classifiers were applied in the task of binary authorship attribution for one pair of male and one pair of female writers. Authors were recognised through their writing styles, defined with the help of quantitative linguistic descriptors.

The structure of the paper is organised as follows. Section 2 is devoted to the rough set theory, the concept of reducts and algorithms for their construction. Section 3 provides the brief description of stylometric analysis of texts, as the application domain. Section 4 addresses the area of feature selection and the task of construction of attribute rankings. Section 5 presents the preformed experiments and comments to the obtained results. Conclusions and future research plans are given in Section 6.

2. Rough sets and reducts

Rough set theory was proposed by Z. Pawlak as a way of dealing with inconsistent and uncertain data. It also supports reduction of dimensionality based on data dependencies [23]. To deal with uncertain concepts a dataset is partitioned into some indiscernible (equivalent) classes, and imprecise concepts are approximated based on these partitions. Objects characterised by the same values of attributes are indiscernible from the point of view of the available knowledge. An *indiscernibility* relation is defined for a subset of attributes $B \subseteq A$ and the set of objects U:

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

A class of indiscernible objects forms a granule (atom) of knowledge about the universe U. In the rough sets theory the granules of indiscernible objects are considered instead of particular objects.

The main structure for data representation is an *information system* or, a special case of information system called a *decision table*. More formally, an information system is defined as $S = (U, A)$, where U is a non-empty, finite set of objects, and A is a non-empty, finite set of attributes, i.e., for every $a \in A$, $a_i : U \rightarrow V_{a_i}$, where V_{a_i} is the set of values of attribute a_i. In the case of a decision table $S = (U, A \bigcup \{d\})$, a set A is a set of *condition attributes*, and $d \notin A$ is a distinguished attribute called a *decision* or a class label, with values $V_d = \{d_1, \ldots, d_{|V_d|}\}$.

In RST feature selection methods are mainly based on the notion of a reduct. Various kinds of reducts are defined, for example: reducts for information systems, decision and local reducts for decision tables, decision and local reducts based on the generalised decision, fuzzy decision reducts [5], etc. Also, different ways leading to reduct construction are used: fuzzy-rough approach supported with measures based on attribute rankings [14], brute-force approach that is applicable to tables with a relatively small number of attributes, Boolean reasoning [22], genetic algorithms [32], many versions of greedy algorithms [2], and others [21].

In the paper, a *decision reduct* is studied, which can be interpreted as such minimal subset of condition attributes that is sufficient to discern any objects in a decision table with different class labels. Formally, a decision reduct is the set of attributes $B \subset A$ such that satisfies the condition:

$$IND(B) \subset IND(d), \tag{2}$$

and discarding any of the attributes included in B violates this condition. The intersection of all reducts is called a *core*. Therefore, the core is the subset of most important attributes. However, the core can be empty (in particular when many reducts are found), and in this case it could be possible to partition the set with all reducts into a few smaller groups with non-empty intersections. When a reduct consists of k attributes, k is called the reduct length.

Based on the reduct including k attributes, it is possible to infer a set of decision rules, for each object x_i from a decision table S. Decision rules take the form: $(a_{i_1} = v_1) \wedge \ldots \wedge (a_{i_k} = v_k) \rightarrow d$, where $v_i \in V_{a_i}$. When decision rules are based on reducts, they are induced from reduced sets of attributes. Taking into account this way of decision rules construction, it becomes evident that the reduct length can be considered as an important factor in the knowledge discovery process, and also for knowledge representation. Short reducts lead to obtaining effective decision algorithms, which allow for prediction of class labels for objects in a test set, using only a reduced set of attributes contained in the reduct, instead of the entire set of features. Constructed in this way decision rules are simple and short, so they are more preferred from the point of view of understanding and interpretation by experts. Also the time required for the process of classification with such rules can be reduced.

In the experiments described in the paper, in the search for reducts the genetic algorithm [33] was used, implemented in Rough Sets Exploration System (RSES) system [3]. This algorithm allows for construction of a sufficiently high number of short reducts in a reasonable time. It is based on a binary genetic algorithm, and there are used classical binary operators such as mutation and crossover, and the “roulette wheel” selection algorithm. With “individuals” represented by bit strings, the fitness function of a subset R (chromosome) has the form:

$$F(R) = \frac{n - L_R}{n} + \frac{2C_R}{m^2 - m}, \tag{3}$$

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, which represents a potential reduct—a subset of attributes. C_R indicates the number of pairs of objects (with different decisions) discerned by the subset of attributes from R.

The process of computation required for C_R has been accelerated by an additional structure, called “distinction table”. It is a binary matrix of the size $(n + 1) \times (m^2 - m)/2$. Each column corresponds to one attribute (the last column corresponds to the decision), and each row of the matrix corresponds to one pair of different objects. If a condition attribute has different values for a pair of objects, then at the intersection of the column and the row the value 1 exists. To find a reduct means to find the minimal subset of columns covering the matrix.

With the described general scheme of genetic algorithm for reduct construction used as a basis, other procedures for generation of short reducts have been developed as well [32]. For example, a greedy algorithm where testing, if an obtained subset of attributes is a reduct, depends on the order of attributes, and in this case the genetic algorithm is used to find this required order of features.

3. Stylometric analysis and characteristic features

A text can be characterised not only by what it is about, but also by how it is written, that is by a writing style [1]. Linguistic elements used, such as individual characters, parts of speech, punctuation marks, observations on layout and formatting, help in stylometric analysis, and lead to author characterisation, comparison and recognition. Individuality of writing styles makes for a high variety in markers that can be employed in a task of authorship attribution [16]. Once discriminating style descriptors are established, text samples of unknown or questioned authorship can be reliably attributed by comparing them with labelled examples. This approach means treating recognition of authorship as a supervised learning task [12, 28], with stylometric markers used as characteristic features defining objects.

Unfortunately, domain knowledge is insufficient to provide a ready set of attributes that would be universal enough to employ for any given problem [24]. Instead, many different such sets are used in research, and quite often data mining approaches are applied for selection of these features, which are most relevant for compared writers. With this attitude, in the data pre-processing stage input datasets are constructed containing relatively high numbers of attributes, which are next analysed and some of them discarded through feature reduction methods [15].

Authors of the same gender most often display some common traits, which makes gender recognition as one of the tasks included among author characterisation problems. Which is why authors of opposite sex should not be compared against each other in quest for establishing authorship. Also, text samples that are used as a base for calculation of values for characteristic features should be of comparable size, representative of writing styles, and sufficiently numerous for all authors. Calculations often involve obtaining frequencies of occurrence for selected descriptors, which makes them continuous. With real-valued attributes the problem of knowledge discovery in continuous space needs to be considered, or some approaches to discretisation employed [28].

In the research works presented in this paper, binary authorship attribution was performed, for one pair of female writers (Edith Wharton and Many Johnston) and one pair of male writers (Jack London and James Curwood). Their novels were divided into three groups, one learning, and two test sets. The works were partitioned into smaller text chunks, 200 for the training sets, and 90 per each test set. Over all text samples in the learning sets, frequencies of occurrence for a hundred most frequently employed words in English language and standard punctuation marks were found, giving a set with lexical and syntactic characteristic features.

In the next step of initial pre-processing, several rankings for attributes were calculated within WEKA environment [11], and they included Gain Ratio, Information Gain, χ^2, and Symmetrical Uncertainty. These features that were found as irrelevant at least once, were excluded from further experiments, and only those which, for both datasets, were always assigned a score higher than zero, were included in tests. They were as follows:

after almost any around before but by during how never on same such until that then there though whether within what who ; ,

Such prepared datasets were next discretised by supervised Fayyad and Irani algorithm [7], which made them ready for processing by classical rough set approach.

4. Feature selection and ranking of attributes

Tasks realised in the feature selection field are important from the point of view of: knowledge discovery (classification), and knowledge representation [25]. In the former case, we aim to improve accuracies of the constructed classifiers, as well as accelerate computation time, and decrease complexity of calculations. In the case of knowledge representation, we would like to obtain a data model that is not complicated i.e., allows to understand easily, and represent knowledge stored in the data with simplicity [6, 8].

The three main categories of feature selection methods, filter, wrapper, and embedded, differ in the way of construction of a set of variables relevant to classification, by how the choice of a particular attribute can be performed. This procedure can be realised as an external process with respect to a classification system employed in a pattern recognition task, it can depend on classification accuracy of such system, or it can engage some mechanism built-in in the system [27]. These general approaches can be combined together, which leads to hybrid solutions.

A ranking of attributes allows to determine importance of variables. Usually there are used some statistical measures [18], machine learning techniques, or specialised algorithms that provide some scores and based on them attributes are ordered [26]. A ranking order reflects discovered significance of elements. Most often the descending ordering is employed, where the most important variables take positions at the top of a ranking, and the least relevant have assigned the lowest positions at the bottom. The popular statistical measures are based on entropy [4], and Relief or OneR are examples of algorithms that calculate scores for attributes. Typically, construction of an attribute ranking is considered as a part of initial data pre-processing stage, and it is realised independently on classification systems.

Reducts, obtained in the framework of rough sets theory, can be considered as a feature selection method, which allows to find the most relevant variables [34]. For the initial set of available features many reducts can be found, with different lengths, so the question is: which of them should be selected? When any two reducts are considered, both must ensure the same classification power (w.r.t. decision table) as the entire set of available attributes, as from the classification point of view, it is a characteristic of all reducts. However, if one reduct contains fewer features than the other, these features can be regarded as more important than others, since their lower number is sufficient to protect the performance of the system. This line of reasoning leads to assigning higher significance to shorter reducts and attributes included in them. Shorter reducts are also preferable from the point of view of knowledge representation.

In this study, for the input datasets several sets of reducts were obtained through the independent runs of the genetic algorithm. Based on these groups, contained in them reducts, reducts characteristics, and attributes included in reducts, a weighting factor for features was proposed and used for construction of attribute rankings.

Let G_{Red} denote a set of reducts Red_l, then $card(G_{Red})$ gives the number of reducts included in the group G_{Red}. Let $RED(G_{Red}, a)$ be the set of all reducts from the group G_{Red} that include a attribute, and $RED(G_{Red}, a, k)$ the set of reducts with length k including a attribute. Then $card\,(RED(G_{Red}, a, k))$ returns for the group G_{Red} the number of reducts with specific length equal k that contain the given attribute a. The weighting factor is defined as:

$$W_F(G_{Red}, a) = \sum_{i=k_{min}}^{k_{max}} \frac{card\,(RED(G_{Red}, a, i))}{card(G_{Red}) \cdot i}, \tag{4}$$

where k_{min} and k_{max} are respectively the minimal and the maximal reduct length for the group G_{Red}. The values of W_F range from 0 (the attribute a is included in none of the reducts in this group), to $1/k_{min}$ when the attribute is included in all reducts, and all reducts have one and the same length. Thus a higher value of the weighting factor 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 including more variables. The calculated scores assigned to attributes led to their descending orders, which were used as rankings of features, each ranking based on a certain group of reducts generated.

5. Performed experiments

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

- search for reducts by the genetic algorithm—15 groups of 10 reducts, and one group of 150 reducts;
- construction of rankings of features based on attributes contained in reducts and calculated factor W_F;
- induction of decision rules for attributes selected through rankings, with gradually decreasing numbers of features considered;
- evaluation of performance for rule classifiers with test sets;
- comparative study of results—multiple small groups of reducts vs. one larger group.

5.1. Calculation of reducts and construction of attribute rankings

For both female and male datasets, by running independently the genetic algorithm implemented in RSES system [3] (with default settings), 15 small groups of reducts, and a single larger group of reducts, were found. The small groups were denoted with indices from 1 to 15, and all had the same cardinality of 10. One larger group was labelled with 0, and its cardinality was 150. Fig. 1 shows minimum, average and maximum length of reducts for each group.

Generated reducts contained different attributes from the whole set of 24 features, and had different lengths. For female writer dataset the length of reducts for 15 groups was in the range from 4 to 7, for one group of 150 reducts

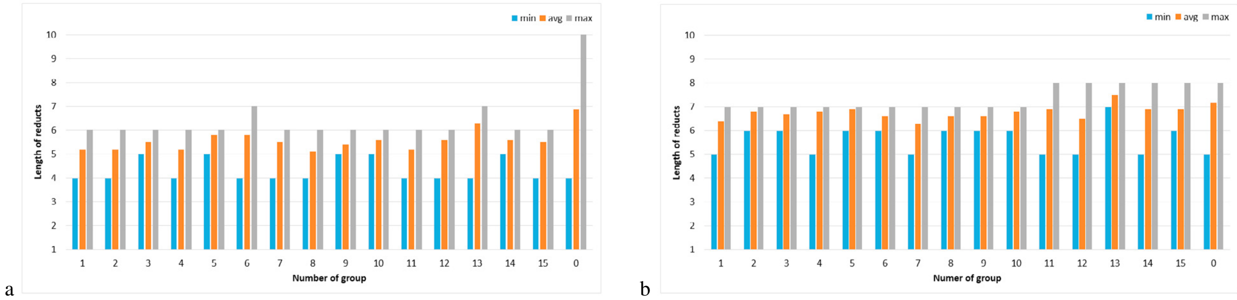


Figure 1. Minimum, average and maximum length of reducts generated for 15 groups of 10 reducts and one group of 150 redcuts for: (a) female writers, (b) male writers.

the maximum length was 10. For male writers the length of reducts for 15 groups, as well as for one group of 150 reducts, was in the range from 5 to 8. It can be observed that for female writers the variety, with respect to the detected minimum and maximum lengths among groups, was relatively low (most of groups had the same values), while for male writer dataset more differences could be found.

Based on the obtained groups of reducts, the weighting factor W_F for features was computed for all groups, and the corresponding rankings of attributes were constructed. Table 1 presents the rankings of features for both datasets, obtained for: 15 small groups of reducts treated independently (denoted by G i, with $i = 1 \dots 15$), for one large group (denoted as G 0), and based on combining all 15 small groups of reducts together into one large group (named G A). As small groups were found independently on each other, it is possible that some of reducts were generated more than once, and then included in more than one group. With this combining groups into one, these repeated reducts were not removed, instead they were treated as any other elements of the group, thus its cardinality was considered as 150.

For all rankings presented for male and female datasets, it can be observed that some of attributes were consistently considered as of some high rank (for example comma), whereas others were often placed at the bottom, and still others took different positions. Yet no two rankings were found the same, and this was true also for G 0 and G A rankings, even though both operated on relatively larger groups of reducts. Also, for small groups of reducts, significant differences with respect to numbers of attributes can be detected.

5.2. Induction of decision rules and evaluation of performance for rule classifiers

For features contained in each ranking a new decision table was created containing only attributes from this ranking, and for such decision table, decision rules were induced using exhaustive algorithm implemented in RSES system. This algorithm constructs all rules with a minimal number of pairs *attribute* = *value* in a premise part of a rule.

Next, following the ordering of features imposed by the ranking, the attributes placed at the lowest ranking positions were sequentially discarded one by one from the decision table. For each decision table reduced in this way, decision rules were induced again. The process of removing attributes was completed when the obtained decision table was contradictory, i.e., there existed at least two objects with the same values of condition attributes but different decisions.

In the next step, accuracy of the constructed rule-based classifiers was calculated as a number of correctly classified objects from a test set, divided by the number of all objects in this test set. In case of conflicts a standard voting strategy was employed, with weighting votes of rules by their supports. For evaluation two test sets were used, and then averages calculated. For the entire set of 24 features, which was considered as a reference point, the averaged classification accuracy was the same for both female and male writer datasets, and it was equal to **96.1**%.

Figs. 2 and 3 present classification accuracy obtained during ranking-based reduction of attributes for both datasets, for 15 groups of 10 reducts and one group of 150 reducts denoted as Group 0. The graphical visualisation shown in Fig. 2 enables to observe some general trends in performance, whereas Fig. 3 provides detailed information and indicates by the coloured cells all cases where the classification accuracy exceeded the reference point. The intensity of cell colour depends on how much the accuracy was improved.

It can be observed that generally better results were obtained for female writer dataset than for the male one. In the former case, for all groups of reducts, with the exception of G 4, there exist numbers of attributes that gave the same

Table 1. Rankings of attributes for 15 groups of reducts (G *i*), one group with 150 reducts (G 0), and one group combined from 15 groups (G A)

Female writer dataset

G 1	G 2	G 3	G 4	G 5	G 6	G 7	G 8	G 9	G 10	G 11	G 12	G 13	G 14	G 15	G 0	G A
on	,	,	,	,	,	on	,	,	,	,	,	,	,	on	,	,
,	on	on	until	until	on	,	on	until	until	on	on	on	on	,	on	on
who	who	until	on	any	who	until	who	by	same	who	who	;	by	who	;	until
same	until	but	same	same	by	by	until	on	by	until	by	never	until	by	until	who
until	but	what	who	on	;	same	same	same	;	then	after	any	;	before	same	;
after	same	;	;	what	after	;	;	;	any	;	until	such	then	;	by	by
by	by	by	by	then	before	then	by	what	on	but	;	before	any	then	but	same
then	what	never	what	;	until	never	around	who	what	any	then	what	after	same	after	then
any	then	who	any	who	then	what	but	but	who	same	such	how	how	around	any	what
;	such	same	there	there	how	who	that	there	then	by	what	then	same	there	who	after
how	;	how	that	after	but	there	then	then	after	before	same	same	what	after	there	any
but	after	then	but	but	around	how	there	after	there	after	never	until	such	but	then	but
such	there	any	such	around	same	any	what	that	but		that	but	never	that	what	before
what	never	after	how	how	what	but	before	before	that		but	there	but	until	such	never
		such	then	before	such	that	after	never	such		before	who	there	how	how	there
		before	after		whether	after	never				any	after	who		around	such
		around			never	before	such								never	how
															before	that
															that	around
															whether	whether

Male writer dataset

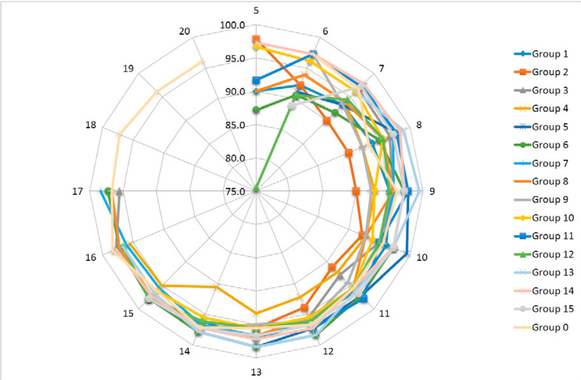
G 1	G 2	G 3	G 4	G 5	G 6	G 7	G 8	G 9	G 10	G 11	G 12	G 13	G 14	G 15	G 0	G A
,	,	then	,	that	,	,	,	,	,	,	by	,	,	,	,	,
though	who	,	by	,	how	on	though	how	there	though	,	around	how	there	there	by
by	how	by	until	around	there	around	how	who	around	there	there	who	on	by	by	how
on	though	how	on	on	any	until	then	though	who	by	how	though	that	then	how	though
around	before	there	though	until	by	how	there	by	by	on	then	until	by	though	who	there
how	on	such	then	any	such	though	who	that	until	how	though	such	around	whether	around	around
there	there	same	;	then	on	by	around	almost	then	who	on	how	;	until	though	on
who	any	that	that	by	who	almost	by	around	such	before	that	same	any	but	on	who
until	around	;	there	almost	though	who	after	before	how	within	such	before	though	any	before	then
;	by	on	any	such	;	that	same	within	;	around	within	that	there	around	then	until
any	within	though	how	within	then	such	before	until	on	until	until	then	then	same	any	that
almost	almost	who	around	how	almost	;	whether	there	within	that	around	within	before	that	until	any
then	then	around	such	but	around	then	until	on	but	almost	any	whether	what	how	;	such
that	;	after	what	who	that	same	that	after	same	what	whether	there	almost	before	what	;
before	until	any	but	though	until	within	almost	any		whether	almost	almost	whether	who	same	almost
within	same	whether	almost	whether	after					then		by	same	after	that	before
such		until	who	there	same					same		on	such	such	almost	within
			within							;		what		on	within	same
										after		;		within	such	whether
										such		any			after	but
												but			but	after
															whether	what

or higher classification accuracy than the reference value. The maximum value equal to 99.5% was detected for 10 attributes in Group 5, and 9 attributes in G 13. In fact, Group 13 shows higher values than 96.1 at each step of reducing attributes contained in the ranking. Moreover, in the case of Groups 2, 3, 10, 13, 14 and 0, the smallest numbers of attributes, i.e., 5, 6 or 7, allow to exceed the classification accuracy found for the whole set of 24 features. It should also be noted that for all groups with the exception of G 3 and 4, the accuracy observed for all attributes contained in each ranking (before some of them were discarded), was the same or grater than the reference point.

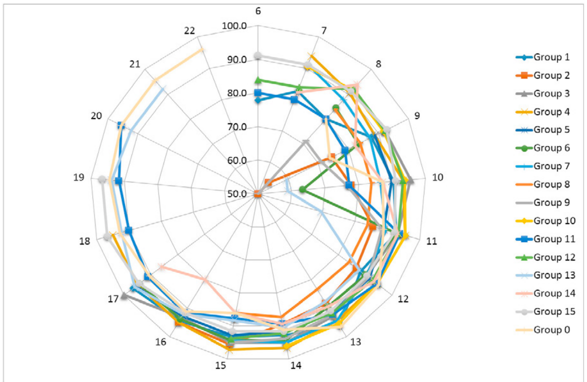
In the case of male dataset, only for 7 groups out of all 16 tested, it was possible to find some numbers of attributes, for which the classification accuracy was the same or improved when compared to operation on 24 features. On the other hand, the maximum for this dataset was equal to 100% of correct recognition, and it was obtained for 17 attributes in Group 3. For all attributes from rankings still included in corresponding decision tables, the same or enhanced classification accuracy (with respect to the reference point) was found for Groups 1, 3, 10, 15 and 0.

5.3. Comparative analysis of performance

In the research there was also studied the influence of the way in which groups of reducts were obtained on the constructed rankings, and resulting from them reduction of attributes. The accuracy of rule classifiers observed for a



a



b

Figure 2. Performance of rule classifiers [%] observed in the process of feature reduction, for: (a) female writer dataset, (b) male writer dataset

	number of attributes															
	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Group 1	90.0	92.2	93.3	93.9	95.0	96.2	96.2	96.2	97.3	96.1						
Group 2	97.8	92.3	90.0	90.0	90.0	92.3	91.1	93.9	95.6	97.2						
Group 3			96.7	96.7	95.6	95.0	92.8	95.6	96.7	96.7	97.3	97.8	95.6			
Group 4	90.0	93.9	93.9	92.3	92.8	92.8	92.3	92.3	93.4	90.6	95.0	95.6				
Group 5		91.1	93.3	97.8	97.8	99.5	97.3	97.3	98.4	97.8	97.8					
Group 6	87.2	90.6	91.7	95.0	97.3	97.3	97.3	98.4	98.4	97.8	97.8	97.3	97.3			
Group 7	91.7	97.3	97.8	97.3	95.6	95.6	95.6	96.1	96.7	96.7	95.6	96.1	98.4			
Group 8	90.0	93.9	93.9	95.6	95.6	92.8	95.6	95.6	96.7	97.3	97.3	97.3	96.7			
Group 9	91.7	97.3	93.4	92.3	92.3	92.8	94.5	96.2	95.0	97.3	97.3					
Group 10	96.7	96.1	96.1	95.6	92.8	93.9	95.6	95.6	95.6	95.6	96.1					
Group 11	91.7	97.3	97.3	97.8	97.8	95.0	97.8	97.3								
Group 12	75.6	90.6	94.4	95.6	95.0	95.0	95.6	96.2	95.6	96.2	97.8	96.7				
Group 13			96.2	98.9	99.5	97.3	96.2	98.4	98.4	97.8	96.7	96.7				
Group 14	97.3	97.3	97.8	98.4	97.3	97.3	95.6	97.3	97.3	97.3	97.3	97.8				
Group 15		88.9	97.3	97.3	97.3	97.3	96.7	96.7	96.7	97.3	97.8					
Group 0		97.3	96.7	93.3	96.1	95.6	95.6	96.7	95.6	97.2	96.1	98.4	96.7	97.3	96.1	96.1

a

	number of attributes																
	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Group 1	77.8	82.8	80.0	87.8	90.6	92.8	88.3	96.1	92.8	93.4	94.5	96.7					
Group 2		50.0	54.5	74.5	77.8	85.6	86.7	88.3	89.4	95.6	95.0						
Group 3			92.2	92.2	95.6	93.3	90.6	92.8	95.0	95.0	93.4	100.0					
Group 4		93.9	90.6	88.9	93.4	95.6	94.5	94.5	96.7	97.2	95.0	95.0	95.0				
Group 5				88.9	90.0	88.9	91.7	91.7	91.7	92.8	92.8	95.6					
Group 6			84.5	83.3	63.3	92.8	90.0	89.5	94.5	94.5	93.9	94.5					
Group 7		91.1	87.8	87.2	86.7	89.5	90.6	94.5	95.0	94.5							
Group 8			83.9	83.9	83.9	83.9	83.9	83.4	87.3	86.1							
Group 9	50.0	50.0	71.2	71.2	76.7	88.4	91.1	92.2	93.9	95.0							
Group 10		90.6	91.7	91.7	93.9	95.6	94.5	95.6	96.1								
Group 11	80.0	80.0	80.0	78.9	77.3	93.9	94.5	92.2	89.4	87.8	91.7	91.1	90.0	91.7	95.6		
Group 12	83.9	83.9	92.2	92.2	93.3	93.4	92.8	91.7	92.2	93.9							
Group 13				59.5	58.9	69.5	87.8	87.8	90.6	88.4	92.3	96.1	92.2	93.9	92.2	92.2	
Group 14		82.3	93.9	82.3	87.2	93.3	93.3	89.5	89.5	86.1	80.0	86.1					
Group 15	91.1	91.1	91.1	92.8	91.1	92.2	90.6	91.1	91.7	91.7	91.7	94.5	96.7	96.7			
Group 0			80.0	73.9	87.2	88.9	95.0	96.7	91.1	86.1	91.1	90.6	92.8	94.5	95.6	95.6	96.1

b

Figure 3. Accuracy of rule classifiers [%] observed in the process of feature reduction, for: (a) female writer dataset, (b) male writer dataset

ranking corresponding to one group of 150 reducts (column G 0 in Table 1) was compared with a ranking built by combining all 15 small groups of 10 reducts each, into a single group (column G A in Table 1). In both cases the number of reducts was the same, but different ranks for attributes were found. Fig. 4 presents the performance of rule classifiers in feature reduction for both ways of ranking construction.

In the case of female writer dataset the values of classification accuracy were close with a slight ascendancy for G A ranking. In the case of male writer dataset, the difference in accuracy of rule-based classifiers was more noticeable and higher values were obtained for feature selection driven by G A ranking. For both datasets, the classification accuracy found at the beginning of feature reduction process, obtained for all attributes included in rankings, was the same.

The results of experiments indicate that the way of independently running multiple times the genetic algorithm for construction of a small number of reducts leads to better classification accuracy than in the case of running the genetic algorithm one time for construction of a larger group of reducts. Thus several smaller subsets of reducts provide better approximation of knowledge about considered attributes than just one set of reducts, even when it is of higher

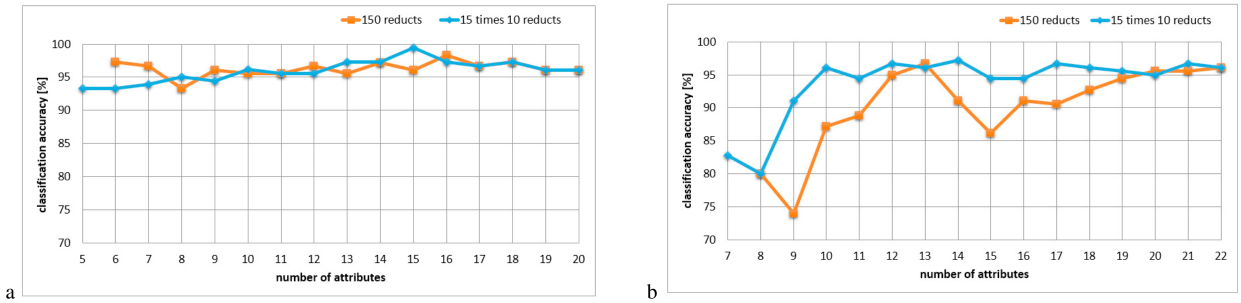


Figure 4. Performance of rule classifiers [%] obtained for ranking-based selection of attributes, with rankings formed by one group of 150 reducts, and 15 groups combined into one, for: (a) female writer dataset, (b) male writer dataset

cardinality. It is known that such heuristic algorithms are used in various problems and situations, where it is difficult to find an exact solution in some acceptable time. However, the solutions obtained by such algorithms are often sub-optimal and they cannot avoid local optima. In this case, increasing a number of runs for genetic algorithm increased also the chances to avoid some local optima.

The results from the performed experiments show the merit of the proposed idea of constructed rankings for attributes, based on decision reducts and the weighting factor W_F.

6. Conclusions

A ranking of attributes is one of feature selection approaches. When it is based on some mechanism inherent to data mining methods, it can be an example of embedded solutions to feature reduction problem. The paper presents research results dedicated to such case, with the application of decision reducts, which is the fundamental concept in rough set processing. In the executed experiments reducts were used in the process of knowledge discovery regarding attributes, which led to their ordering, depending on relative frequencies of occurrence in reducts with specific lengths.

Weighting of features resulted in their rankings, which were next exploited for feature selection. For gradually decreasing subsets of attributes, chosen by discarding less relevant elements from each ranking, decision rules were generated in the classical rough set approach, with exhaustive algorithm for rule induction. These inferred rule sets were then used in classification of test samples, with standard voting as conflict resolution strategy. The performance of thus constructed rule classifiers was compared against each other, and with the case of set of rules inferred from the input datasets with the entire set of condition attributes, which was used as the reference point.

The experiments show that while using rankings based on reducts and the proposed weighting factor W_F, to drive the process of feature reduction, it is possible to find several cases with at least the same (but also increased), classification accuracy with discarding a significant portion of available condition attributes. The results indicate that rankings constructed by fusion of information on attributes from several small groups of reducts, found in independent runs of the genetic algorithm, worked better than the one based on single larger set of reducts, generated in one run, despite the fact that in both cases the overall number of considered reducts was the same.

In the future research instead of just subsets of reducts, there will be studied all relative reducts that can be found through exhaustive algorithm for the input datasets, as well as different weighting functions for attributes.

Acknowledgements

The research works described in the paper were performed at the University of Silesia in Katowice, Institute of Computer Science, and at the Silesian University of Technology, Gliwice, within the statutory project (RAU-6, 2020).

References

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

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

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

[4] Biesiada, J., Duch, W., Kachel, A., Pałucha, S., 2005. Feature ranking methods based on information entropy with Parzen windows, in: Proceedings of International Conference on Research in Electrotechnology and Applied Informatics, Katowice, Poland. pp. 109–119.

[5] Cornelis, C., Martín, G.H., Jensen, R., Ślęzak, D., 2008. Feature selection with fuzzy decision reducts, in: Wang, G., Li, T., Grzymala-Busse, J.W., Miao, D., Skowron, A., Yao, Y. (Eds.), Rough Sets and Knowledge Technology, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 284–291.

[6] Dash, M., Liu, H., 1997. Feature selection for classification. Intelligent Data Analysis 1, 131–156.

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

[8] Ferreira, A., Figueiredo, M., 2012. Efficient feature selection filters for high-dimensional data. Pattern Recognition Letters 33, 1794–1804.

[9] Grzegorowski, M., Ślęzak, D., 2019. On resilient feature selection: Computational foundations of r-c-reducts. Information Sciences 499, 25 – 44.

[10] Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. (Eds.), 2006. Feature Extraction: Foundations and Applications. volume 207 of *Studies in Fuzziness and Soft Computing*. Physica-Verlag, Springer.

[11] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I., 2009. The WEKA data mining software: An update. SIGKDD Explorations 11, 10–18.

[12] Han, J., Kamber, M., Pei, J., 2011. Data Mining: Concepts and Techniques. Morgan Kaufmann.

[13] Janusz, A., Ślęzak, D., 2014. Rough set methods for attribute clustering and selection. Applied Artificial Intelligence 28, 220–242.

[14] Jensen, R., Shen, Q., 2007. Fuzzy-rough sets assisted attribute selection. IEEE Transactions on Fuzzy Systems 15, 73 – 89.

[15] Jensen, R., Shen, Q., 2008. Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches. IEEE Press Series on Computational Intelligence, Wiley-IEEE Press.

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

[17] Liu, H., Motoda, H., 2007. Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series). Chapman & Hall/CRC.

[18] Mansoori, E., 2013. Using statistical measures for feature ranking. International Journal of Pattern Recognition and Artifficial Intelligence 27, 1350003–14.

[19] Moshkov, M.J., Piliszczuk, M., Zielosko, B., 2007. On construction of partial reducts and irreducible partial decision rules. Fundamenta Informaticae 75, 357–374.

[20] Nguyen, H.S., 2003. On the decision table with maximal number of reducts. Electronic Notes in Theoretical Computer Science 82, 198 – 205. International Workshop on Rough Sets in Knowledge Discovery and Soft Computing (Satellite Event for ETAPS 2003).

[21] Nguyen, H.S., Ślęzak, D., 1999. Approximate reducts and association rules - correspondence and complexity results, in: RSFDGrC 1999. Springer. volume 1711 of *LNCS*, pp. 137–145.

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

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

[24] Pearl, L., Steyvers, M., 2012. Detecting authorship deception: a supervised machine learning approach using author writeprints. Literary and Linguistic Computing 27, 183–196.

[25] Reif, M., Shafait, F., 2014. Efficient feature size reduction via predictive forward selection. Pattern Recognition 47, 1664–1673.

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

[27] Stańczyk, U., 2013. Weighting of attributes in an embedded rough approach, in: Gruca, A., Czachórski, T., Kozielski, S. (Eds.), Man-Machine Interactions 3. Springer-Verlag, Berlin, Germany. volume 242 of *AISC*, pp. 475–483.

[28] Stańczyk, U., Zielosko, B., 2019. On approaches to discretisation of stylometric data and conflict resolution in decision making, in: Rudas, I.J., Csirik, J., Toro, C., Botzheim, J., Howlett, R.J., Jain, L.C. (Eds.), Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES-2019, Budapest, Hungary, 4-6 September 2019. Elsevier. volume 159 of *Procedia Computer Science*, pp. 1811–1820.

[29] Stańczyk, U., Zielosko, B., Jain, L.C. (Eds.), 2018. Advances in Feature Selection for Data and Pattern Recognition. volume 138 of *Intelligent Systems Reference Library*. Springer.

[30] Widz, S., Stawicki, S., 2016. Generalized majority decision reducts, in: Ganzha, M., Maciaszek, L.A., Paprzycki, M. (Eds.), Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016, Gdańsk, Poland, September 11-14, 2016, IEEE. pp. 165–174.

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

[32] Wróblewski, J., 1996. Theoretical foundations of order-based genetic algorithms. Fundamenta Informaticae 28, 423–430.

[33] Wróblewski, J., 1998. Covering with reducts - a fast algorithm for rule generation, in: Polkowski, L., Skowron, A. (Eds.), Rough Sets and Current Trends in Computing, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 402–407.

[34] Yu, L., Liu, H., 2004. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research 5, 1205–1224.

[35] Zielosko, B., Piliszczuk, M., 2008. Greedy algorithm for attribute reduction. Fundamenta Informaticae 85, 549–561.