Construction of Features Ranking— Global Approach

Beata Zielosko
University of Silesia in Katowice/Institute of Computer Science
Sosnowiec, Poland *beata.zielosko@us.edu.pl*

Urszula Stańczyk
Silesian University of Technology
Department of Computer Graphics Vision and Digital Systems
Gliwice, Poland *urszula.stanczyk@polsl.pl*

Kamil Jabłoński
University of Silesia in Katowice/Faculty of Science and Technology
Sosnowiec, Poland *kjablonski1@us.edu.pl*

Abstract

The paper presents a research methodology focused on generating a global attribute ranking, based on discrete variants of datasets, transformed by multiple algorithms. The approach enables to accumulate information on feature importance from such local sources and, when it is represented in the form of a global ranking, identify the features that are the most relevant for decision-making processes. The research procedure was validated by experiments, in which the rankings were used to control filtering decision rules induced by the classic rough set approach. In the vast majority of cases it was possible to obtain noticeable attribute reduction, and predictions improved or comparable with the results obtained for local variants of the data.

Keywords: Ranking, Greedy Algorithm, Decision Rule Filtering, Stylometry, Distributed Data.

1. Introduction

Technological development and distributed architecture of information systems force extracting knowledge from data varying in structure and localised in distributed or centralised sources [7]. In successful system development, feature selection and ranking construction are key aspects, especially when the process is based on data analysis and machine learning. The main objectives of feature selection include facilitating decision-making processes, enhancing data understanding, and improving the accuracy of classifiers based on the selected most relevant features. The importance of attributes can be determined through their ranking [5].

The learnt knowledge should be easily accessible for classification, but also presented in some understandable form. Decision rules satisfy both requirements [11]. Support and length are important factors in determining rule quality. Support allows mapping important patterns present in the data, while shorter rules decrease the likelihood of model overfitting and enhance generality. Minimising lengths and maximising supports of decision rules belong to NP-hard problems, but the previous research [3] showed that, with some natural assumptions on the class NP, there exists the greedy algorithm for induction of decision rules, which is not far from the best approximate polynomial algorithms for minimising the length of decision rules.

Algorithms for induction of decision rules often require discrete features, especially within the framework of rough set theory [9]. Therefore, when the input space is continuous, discretisation it needed to transform continuous attribute values into categorical form [10]. The choice of a particular method is not trivial due to the diversity of algorithms and the specificity of data.

The rationale for processing data occurring in different forms, the characteristics and advantages of decision rules, and the properties of the greedy algorithm were the motivation for the novel methodology proposed. The diversity of discretisation algorithms and a lack of clear guidelines as to which discretisation algorithm is most appropriate for a given dataset, leads to such an approach where the input data are discretised by multiple algorithms, and the resulting variants of transformed data are treated as local data sources. For each data variant, based on decision rules induced by the greedy algorithm, a local attribute ranking is constructed. The global ranking accumulates information taking into account the specifics of all local data sources, and recalls from them a kind of common knowledge that is valid for all considered data variants.

The effectiveness of both ranking types is compared by employing them to control the process of decision rule filtering, performed for all distributed local data sources corresponding to discrete variants of the input data. The order of appearance of attributes in rankings and the classification models created on this basis allow for the discovery of the main patterns hidden in the data. The advantage of the proposed approach to the global ranking definition is its universality, as local rankings can be created based on various algorithms. In this work, local rankings are created based on the properties of the greedy algorithm for induction of decision rules.

In the investigations, the proposed methodology was applied to datasets from the stylometry domain and the authorship attribution task, treated as classification [12]. For all constructed rule-based classifiers the accuracy was established, and characteristics of the rule sets recalled were analysed. The promising results obtained, with multiple cases of vastly reduced rule sets offering at least the same but also improved performance, validated the research framework on the one hand, while on the other indicate the merit of further studies.

The paper consists of five sections. Section 2 presents the background and the proposed methodology for construction of global and local rankings. Section 3 is devoted to experimental setup. The experimental results are analysed in Section 4 and conclusions presented in Section 5.

2. Data Mining and Methodology for Multi-Attribute Weighting

The proposed methodology involves the construction of local and global rankings, which are used to guide the filtering decision rules process. The sets of retrieved rules result in the construction of classifiers. The processing stages are described below.

2.1. Properties of Greedy Algorithm

The greedy algorithm works sequentially. It is applied to each row of a dataset with the assigned decision. In each step, an attribute is selected that discerns the maximum number of rows that are labelled with different decisions. Based on the selected features, the decision rule is constructed.

Previous research on induction of decision rules [8] showed that, during the rule construction the greedy algorithm in each step selects an attribute that separates at least 50% of rows with a different label. In this work, the application of the greedy algorithm was extended through the proposed methodology, where the property related to the selection of attributes with good separation of rows from different decision classes was used for construction of attribute rankings.

2.2. Decision Rule Filtering

Decision rules allow for the construction of classifiers, but also allow for discovery of patterns occurring in the data and constitute an intuitive form of knowledge representation. There are many algorithms for induction of decision rules [7]. In the investigations, decision rules were induced by the greedy algorithm with aim of local ranking construction, and the exhaustive algorithm was used to build classification models and verify the rankings.

The exhaustive algorithm finds decision rules with minimum lengths [11]. When all rules on examples are found, cardinality of a rule set can be relatively high. The greedy algorithm is derived from the set cover problem and hence infers the minimal number of rules that is sufficient to correctly classify all cases from the train set, so it can find much smaller rule sets.

The length of a rule, equal to the number of descriptors in the premise, reflects its descriptiveness. Shorter rules are more universal, which promotes understanding. The support of a rule reports the number of instances in the training data that are consistent with it. Higher supports correspond to the frequently appearing patterns. These properties can be used in rule selection.

The filtering rules process can also be controlled through the attributes included in descriptors. Rule filtering driven by the attribute ranking allows to select such rules that rely on the most relevant features and helps to recall such sets of decision rules that provide classification quality comparable to or better than that of the entire set of decision rules, while at the same time improving rule sets characteristics, such as number of rules, average lengths and supports.

In the proposed research methodology, two types of rankings are developed, local for individual variants of discrete data, and global. Both enable building classifiers optimised from the point of view of the number of rules, ensuring lower storage requirements. The resulting rule sets provide a mapping of global knowledge and patterns found in local data sources.

2.3. Steps of the Proposed Methodology

The proposed methodology consists of: (i) initial data preparation, (ii) data discretisation, (iii) induction of decision rules by the greedy and exhaustive algorithms, (iv) construction of local attribute rankings, (v) construction of the global attribute ranking, (vi) application of both local and global rankings to filtering rules on local sources, from the sets inferred by the exhaustive algorithm, (vii) construction of rule-based classifiers based on filtered rules and analysis of rule sets characteristics, and (viii) evaluation and comparison of the results obtained for the global ranking with the results obtained for local data sources.

Data preparation involves constructing datasets in the chosen application domain. All sets represent instances of binary classification with balanced decision classes, and the attributes have continuous values. In the next step, all datasets are discretised using selected supervised and unsupervised methods [4]. As many discretisation algorithms are used, that many variants of data are obtained, each representing one local data source. From each discrete data variant, the decision rules are induced with two algorithms: the greedy algorithm described in Sec. 2.1, and the exhaustive algorithm implemented in the Rough Set Exploration System (RSES) [1]. Then, based on the investigation of characteristics of decision rules induced by the greedy algorithm and its properties, local rankings of attributes are constructed for each variant of data.

Unsupervised discretisation involves the input parameter specifying the number of intervals to be constructed. This number can be varied, so it is highly probable that local sources, corresponding to multiple versions of data discretised by unsupervised transformations, will be predominant. To avoid it, the global ranking is built in two stages. In the first step within the framework of supervised and unsupervised discretisation methods, for each a single representative ranking is formed, and then based on these two orderings, the global ranking is defined.

From each discrete data variant that serve as local sources, decision rules are induced by the exhaustive algorithm. The sets of inferred rules are subjected to the filtering process, controlled by rankings of attributes. For all sources, the global and the corresponding local rankings are applied, so for each source, two filtering processes are executed. Filtering is carried out sequentially forward. Starting with the top ranking positions, such rules are recalled from the entire set that include conditions only on the considered attributes. The set of features is gradually expanded by adding less significant attributes. At the beginning of the process, the set of recalled rules is empty, and at the end it contains all induced rules.

Based on the sets of filtered rules, the classifiers are constructed and tested, and their characteristics recorded. These include evaluation of performance, rule lengths and supports, numbers of retrieved rules, and numbers of attributes. These characteristics are compared between the ones obtained for the global ranking and those resulting from local ones.

2.4. Construction of Local and Global Attribute Rankings

Let M denote the number of discrete data variants corresponding to local sources, grouped into S discretisation approaches, each in s_i versions, so $M = \sum_{i=1}^{S} s_i$. For all data variants, the greedy algorithm infers sets of decision rules, from which duplicates are removed. For all M sets of rules, for each of the N attributes in a dataset, the number of its occurrences in the induced rules is counted. In the case of multiple attributes with the same values, the maximum percentage of separated rows with a different decision is calculated. Then, the maximum support of decision rules in which the feature appears is taken into account. The attributes are sorted descendingly by the described weights. As a result of such processing, M local rankings $LRank_i$ are created for the features. Not necessarily all attributes appear in all rankings, so they can contain varying numbers of features. An attribute that appears in more rankings is considered more important than an attribute that appears in fewer rankings.

The global ranking is created in multiple stages. In the first step, for each attribute a and each local ranking $LRank_i$, elementary weight is determined as the attribute's ranking position $Pos(LRank_i, a)$ divided by the number of attributes in this ranking $Nr(LRank_i)$,

$$w_a^{LRank_i} = \frac{Pos(LRank_i, a)}{Nr(LRank_i)}. \tag{1}$$

The smaller the weight, the more important the attribute is considered and the higher it is ranked. The minimum is equal to $1/Nr(LRank_i)$ when the attribute in question takes the top of this ranking, and the maximum is 1, when the attribute occupies the bottom position.

Then, within the framework of each of the S discretisation approaches, for all s_i versions of this approach, for each a attribute, the number of its occurrences k_a in the s_i local rankings is determined. The attributes are grouped by the values of k, and within each category the attribute weights are added, leading to the accumulated weight $W_a^{approach_k}$ relative to both k and the specific discretisation approach. Ordering of attributes relative to k ensures that the sum of attribute weights is sorted within a group of attributes with the same number of occurrences. To create the global ranking W_a^{global}, the weights of attributes $W_a^{approach}$ are summed over S discretisation approaches and then sorted within groups with the same number of occurrences,

$$W_a^{global_k} = \sum_{j=1}^{S} \left(W_a^{approach_{j_k}} \right) = \sum_{j=1}^{S} \left(\sum_{i=1}^{s_{j_k}} w_a^{LRank_i} \right). \tag{2}$$

Therefore, also W_a^{global} is relative to k. The features are arranged in descending order.

3. Experimental Setup

The section provides details of the stages included in the proposed methodology. It describes the application domain and its transformations, decision rule induction, and ranking construction.

3.1. Input Datasets and Discretisation

In the investigations, the application domain was the stylometric analysis of texts, with authorship attribution considered as a supervised machine learning task [12]. The authors are recognised based on their writing styles, defined through quantitative stylometric features. In the research, the set of 24 attributes was selected, based on the frequencies of occurrence for popular function words and punctuation marks, which made for the continuous input domain.

Two datasets were constructed, each for a binary authorship attribution task with balanced classes. The authors were well-known writers, Mary Johnston and Edith Wharton (Female writer dataset), and Jack London and James Curwood (Male writer dataset). Each dataset (female and male) included a single train set (200 samples) and two test sets (each 90 samples).

To prepare datasets for processing, all sets were independently discretised using various procedures. In the research, selected representatives of both supervised and unsupervised algorithms were employed. The supervised Fayyad and Irani [2] and Kononenko [6] methods are considered superior, as they condition the construction of categorical representations on recognised classes. They are non-parametric and returned single data variants, denoted dsF and dsK.

Unsupervised procedures focus only on transformed domains. In equal width binning (duw) a specified number of intervals of the same width is defined, while for equal frequency binning (duf) such bins are formed that ensure the representation of the same number of original datapoints in each. For both, 9 data variants were obtained by varying the number of intervals to be constructed from 2 (duw2 or duf2) to 10 (duw10 or duf10). The total number of discrete data variants was 20 per dataset, and that ensured as many local data sources were represented.

3.2. Induction of Decision Rules

Each local data source, corresponding to one of data variants obtained from discretisation, was subjected to induction of decision rules with two algorithms. The rules inferred by the greedy algorithm (described in Sec. 2.1) were exploited in the process of local and global ranking construction. The rules returned by the exhaustive algorithm implemented in the RSES system, were treated as new sources of discovered knowledge and to verify the efficiency of rankings. The characteristics of the rule sets, treated as reference points, are given in Table 1. They include the number of rules NoR, the average rule length AvgL, and the average rule support AvgS.

Table 1. The characteristics of rule sets inferred from all local sources with the exhaustive algorithm

Discretisation	Female dataset			Male dataset			Discretisation	Female dataset			Male dataset		
method	AvgL	AvgS	NoR	AvgL	AvgS	NoR	method	AvgL	AvgS	NoR	AvgL	AvgS	NoR
dsF	4.81	6.62	4121	5.07	5.87	15283	dsK	5.33	5.39	10190	5.05	5.52	20815
duf2	5.70	3.71	103645	5.65	3.57	138910	duw2	5.48	6.52	2094	4.63	7.52	1509
duf3	4.46	2.08	122527	4.37	2.04	135696	duw3	5.57	3.65	26025	5.70	3.32	32447
duf4	3.82	1.84	81723	3.75	1.80	96327	duw4	4.98	2.63	46480	4.95	2.86	47574
duf5	3.46	1.67	68994	3.38	1.66	76240	duw5	4.51	2.12	67054	4.55	2.28	79561
duf6	3.24	1.56	58327	3.20	1.53	65184	duw6	4.13	1.97	75888	4.17	1.99	77033
duf7	3.09	1.52	49026	3.07	1.50	55184	duw7	3.90	1.80	70152	4.00	1.88	75733
duf8	2.99	1.48	42490	2.96	1.48	47511	duw8	3.69	1.75	60422	3.73	1.77	72675
duf9	2.90	1.45	37750	2.87	1.46	42278	duw9	3.53	1.68	59332	3.53	1.67	68722
duf10	2.80	1.45	34155	2.78	1.44	38670	duw10	3.42	1.63	54187	3.41	1.63	61920

For each data variant obtained from the Male set, the number of rules was greater than for the data variants of the Female set, with the exception of the local duw2 source. The smallest average rule lengths for both datasets were obtained for higher numbers of bins used for unsupervised methods, while the highest average rule supports resulted from either supervised processing or unsupervised discretisation with small bin numbers.

The rule filtering procedure was applied for each local rule set, the process driven by attributes selected from a ranking. To all rule sets, the constructed global ranking was applied and the corresponding local ranking. The sets of recalled rules were used as the basis for classifier construction and their characteristics were compared with the reference points.

3.3. Rankings

For the Female and Male datasets, together 40 local rankings were constructed. Table 2 presents the global ranking (W^{global}) and two generalised rankings for the supervised (W^{ds}) and unsupervised (W^{du}) approaches, obtained in the first stage of the global ranking construction for each dataset. In the global ranking, all available attributes are present. For supervised methods, this situation does not occur for 7 attributes from the Female dataset and for 6 attributes from the Male dataset. Subsequent positions contain different attributes for both genders of writers. The attributes in the first position in W^{global} and W^{ds} are the same, for both the Female and Male datasets, similar in the case of W^{du} ranking.

Table 2. Global ranking and generalised rankings obtained for supervised and unsupervised methods

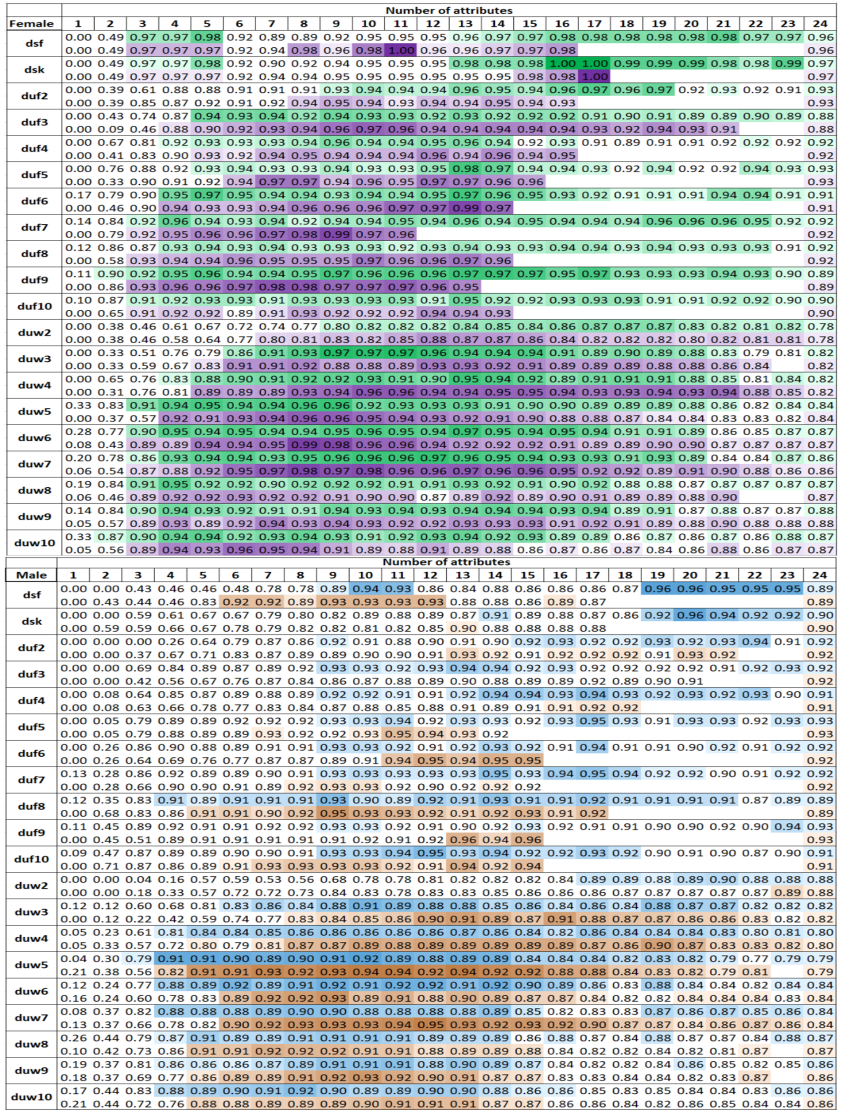
Ranking	Position in a ranking																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
	Female dataset																							
W^{ds}	a1	a23	a2	a17	a3	a13	a11	a10	a8	a6	a22	a0	a19	a5	a7	a4	a16							
W^{du}	a0	a1	a23	a2	a17	a13	a20	a6	a19	a3	a14	a18	a11	a22	a10	a8	a9	a12	a4	a7	a5	a21	a16	a15
W^{global}	a1	a23	a2	a17	a13	a3	a0	a6	a11	a10	a19	a8	a22	a5	a7	a4	a16	a20	a14	a18	a9	a12	a21	a15
	Male dataset																							
W^{ds}	a1	a23	a3	a0	a2	a16	a18	a21	a10	a7	a8	a22	a6	a19	a12	a15	a11	a9						
W^{du}	a0	a1	a2	a23	a3	a6	a17	a18	a16	a13	a19	a21	a8	a5	a20	a14	a7	a9	a11	a12	a4	a10	a15	a22
W^{global}	a1	a0	a23	a3	a2	a16	a18	a21	a6	a8	a19	a7	a10	a22	a12	a11	a9	a15	a17	a13	a5	a20	a14	a4

4. Obtained Results

The experimental results include the characteristics of induced and filtered sets of rules. The results are provided for all stages of rule filtering based on constructed local and global rankings.

4.1. Trends in Performance in Rule Filtering

For the rule-based classifiers obtained by attribute-driven filtering, the accuracy was obtained by labelling samples from the test sets. The results are shown in Figure 1, for the global (green and blue colours of columns), and for a local (violet and brown colours of columns) rankings, across all variants of discretised datasets, the left half for the Female and the right for the Male dataset. The cell with the number 24 and all subsequent in the right-hand-side contains the classification accuracy for the full set of attributes and provides a reference point for the results obtained. The coloured cells indicate all cases where the classification accuracy exceeded the reference point. The intensity of the cell's colour depends on how much the accuracy was improved.



	Number of attributes																							
Female	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
dsf	0.00	0.49	0.97	0.97	0.98	0.92	0.89	0.89	0.92	0.95	0.95	0.95	0.96	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.97	0.97	0.96
	0.00	0.49	0.97	0.97	0.97	0.92	0.94	0.98	0.96	0.98	1.00	0.96	0.96	0.97	0.97	0.98								0.96
dsk	0.00	0.49	0.97	0.97	0.98	0.92	0.90	0.92	0.94	0.95	0.95	0.95	0.98	0.98	0.98	1.00	1.00	0.99	0.99	0.99	0.98	0.98	0.99	0.97
	0.00	0.49	0.97	0.97	0.97	0.92	0.94	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.98	0.98	1.00							0.97
duf2	0.00	0.39	0.61	0.88	0.88	0.91	0.91	0.91	0.93	0.94	0.94	0.94	0.96	0.95	0.94	0.96	0.97	0.96	0.97	0.92	0.93	0.92	0.91	0.93
	0.00	0.39	0.85	0.87	0.92	0.91	0.92	0.94	0.95	0.94	0.93	0.94	0.94	0.95	0.94	0.93								0.93
duf3	0.00	0.43	0.74	0.87	0.94	0.93	0.94	0.92	0.94	0.93	0.93	0.92	0.93	0.92	0.92	0.92	0.91	0.90	0.91	0.89	0.89	0.90	0.89	0.88
	0.00	0.09	0.46	0.88	0.90	0.92	0.93	0.94	0.96	0.97	0.96	0.94	0.94	0.94	0.94	0.94	0.93	0.92	0.94	0.93	0.91			0.88
duf4	0.00	0.67	0.81	0.92	0.93	0.93	0.93	0.94	0.96	0.94	0.94	0.95	0.96	0.94	0.92	0.93	0.91	0.89	0.91	0.91	0.92	0.92	0.92	0.92
	0.00	0.41	0.83	0.90	0.93	0.92	0.94	0.95	0.94	0.94	0.94	0.96	0.94	0.96	0.94	0.95								0.92
duf5	0.00	0.76	0.88	0.92	0.93	0.94	0.93	0.93	0.94	0.93	0.93	0.95	0.98	0.97	0.94	0.94	0.93	0.92	0.94	0.92	0.92	0.94	0.93	0.93
	0.00	0.33	0.90	0.91	0.92	0.94	0.97	0.97	0.94	0.96	0.95	0.97	0.97	0.96	0.96									0.93
duf6	0.17	0.79	0.90	0.95	0.97	0.95	0.94	0.94	0.93	0.94	0.94	0.95	0.97	0.96	0.95	0.93	0.92	0.91	0.91	0.91	0.94	0.94	0.91	0.91
	0.00	0.46	0.90	0.94	0.93	0.93	0.94	0.96	0.96	0.96	0.97	0.97	0.99	0.97										0.91
duf7	0.14	0.84	0.92	0.96	0.94	0.93	0.94	0.92	0.94	0.94	0.95	0.94	0.96	0.94	0.95	0.94	0.94	0.94	0.96	0.96	0.96	0.95	0.92	0.92
	0.00	0.79	0.92	0.95	0.96	0.96	0.97	0.98	0.99	0.97	0.96													0.92
duf8	0.12	0.86	0.87	0.93	0.94	0.93	0.94	0.93	0.93	0.93	0.92	0.93	0.94	0.93	0.93	0.94	0.94	0.93	0.94	0.93	0.93	0.93	0.91	0.92
	0.00	0.58	0.93	0.94	0.94	0.96	0.95	0.95	0.95	0.97	0.96	0.96	0.97	0.96										0.92
duf9	0.11	0.90	0.92	0.95	0.96	0.94	0.94	0.95	0.97	0.96	0.96	0.96	0.97	0.97	0.97	0.95	0.97	0.93	0.93	0.93	0.94	0.93	0.90	0.89
	0.00	0.86	0.93	0.96	0.96	0.97	0.98	0.98	0.97	0.97	0.97	0.96	0.95											0.89
duf10	0.10	0.87	0.91	0.92	0.93	0.93	0.91	0.93	0.93	0.93	0.93	0.91	0.95	0.92	0.92	0.93	0.93	0.93	0.91	0.91	0.92	0.92	0.90	0.90
	0.00	0.65	0.91	0.92	0.92	0.89	0.91	0.93	0.92	0.92	0.92	0.94	0.94	0.93										0.90
duw2	0.00	0.38	0.46	0.61	0.67	0.72	0.74	0.77	0.80	0.82	0.82	0.82	0.84	0.85	0.84	0.86	0.87	0.87	0.87	0.83	0.82	0.81	0.82	0.78
	0.00	0.38	0.46	0.58	0.64	0.77	0.80	0.81	0.83	0.82	0.85	0.88	0.87	0.87	0.86	0.84	0.82	0.82	0.82	0.80	0.82	0.81	0.81	0.78
duw3	0.00	0.33	0.51	0.76	0.79	0.86	0.91	0.93	0.97	0.97	0.97	0.96	0.94	0.94	0.94	0.91	0.89	0.90	0.89	0.88	0.83	0.79	0.81	0.82
	0.00	0.33	0.59	0.67	0.83	0.91	0.91	0.92	0.88	0.88	0.89	0.93	0.93	0.92	0.91	0.89	0.89	0.89	0.88	0.88	0.86	0.84		0.82
duw4	0.00	0.65	0.76	0.83	0.88	0.90	0.91	0.92	0.92	0.93	0.91	0.90	0.95	0.94	0.92	0.89	0.91	0.91	0.91	0.88	0.85	0.81	0.84	0.82
	0.00	0.31	0.76	0.81	0.89	0.89	0.89	0.93	0.94	0.96	0.96	0.94	0.94	0.95	0.95	0.94	0.93	0.93	0.94	0.93	0.94	0.88	0.85	0.82
duw5	0.33	0.83	0.91	0.94	0.95	0.94	0.94	0.96	0.96	0.92	0.93	0.93	0.93	0.91	0.90	0.90	0.89	0.89	0.89	0.88	0.86	0.82	0.84	0.84
	0.00	0.37	0.57	0.92	0.91	0.93	0.94	0.96	0.96	0.95	0.94	0.93	0.92	0.91	0.90	0.88	0.88	0.87	0.84	0.84	0.83	0.83	0.82	0.84
duw6	0.28	0.77	0.90	0.95	0.94	0.95	0.94	0.94	0.95	0.96	0.95	0.94	0.97	0.95	0.94	0.95	0.94	0.91	0.91	0.89	0.86	0.85	0.87	0.87
	0.08	0.43	0.89	0.89	0.94	0.94	0.95	0.99	0.98	0.96	0.96	0.94	0.92	0.92	0.92	0.91	0.89	0.89	0.90	0.90	0.87	0.87	0.87	0.87
duw7	0.20	0.78	0.86	0.93	0.94	0.94	0.93	0.95	0.96	0.96	0.96	0.97	0.96	0.95	0.94	0.93	0.93	0.91	0.93	0.89	0.84	0.84	0.87	0.86
	0.06	0.54	0.87	0.88	0.92	0.95	0.97	0.98	0.97	0.98	0.96	0.96	0.97	0.96	0.96	0.95	0.92	0.92	0.89	0.91	0.90	0.88	0.86	0.86
duw8	0.19	0.84	0.91	0.95	0.92	0.92	0.90	0.92	0.92	0.92	0.91	0.91	0.93	0.92	0.91	0.90	0.92	0.88	0.88	0.87	0.87	0.87	0.87	0.87
	0.06	0.46	0.89	0.92	0.92	0.93	0.92	0.92	0.91	0.90	0.90	0.87	0.89	0.92	0.89	0.90	0.91	0.89	0.89	0.88	0.90			0.87
duw9	0.14	0.84	0.90	0.94	0.93	0.92	0.91	0.91	0.94	0.93	0.94	0.93	0.94	0.94	0.94	0.93	0.94	0.89	0.91	0.87	0.88	0.87	0.87	0.88
	0.05	0.57	0.89	0.93	0.89	0.92	0.94	0.93	0.94	0.93	0.92	0.92	0.93	0.93	0.93	0.91	0.92	0.91	0.89	0.88	0.90	0.88	0.88	0.88
duw10	0.33	0.87	0.90	0.94	0.94	0.92	0.93	0.94	0.93	0.91	0.92	0.93	0.94	0.92	0.93	0.89	0.89	0.86	0.87	0.86	0.87	0.86	0.88	0.87
	0.05	0.56	0.89	0.94	0.93	0.96	0.95	0.94	0.91	0.89	0.88	0.91	0.89	0.88	0.86	0.87	0.86	0.87	0.84	0.86	0.88	0.86	0.87	0.87

	Number of attributes																							
Male	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
dsf	0.00	0.00	0.43	0.46	0.46	0.48	0.78	0.78	0.89	0.94	0.93	0.86	0.84	0.88	0.86	0.86	0.86	0.87	0.96	0.96	0.95	0.95	0.95	0.89
	0.00	0.43	0.44	0.46	0.83	0.92	0.92	0.89	0.93	0.93	0.93	0.93	0.88	0.88	0.86	0.89	0.87							0.89
dsk	0.00	0.00	0.59	0.61	0.67	0.67	0.79	0.80	0.82	0.89	0.88	0.89	0.87	0.91	0.89	0.88	0.87	0.86	0.92	0.96	0.94	0.92	0.92	0.90
	0.00	0.59	0.59	0.66	0.67	0.78	0.79	0.82	0.82	0.81	0.82	0.85	0.90	0.88	0.88	0.88	0.88							0.90
duf2	0.00	0.00	0.00	0.26	0.64	0.79	0.87	0.86	0.92	0.91	0.88	0.90	0.91	0.90	0.92	0.93	0.92	0.92	0.93	0.92	0.93	0.94	0.91	0.92
	0.00	0.00	0.37	0.67	0.71	0.83	0.87	0.89	0.89	0.90	0.90	0.91	0.93	0.92	0.91	0.92	0.92	0.92	0.91	0.93	0.92			0.92
duf3	0.00	0.00	0.69	0.84	0.89	0.87	0.89	0.92	0.93	0.93	0.92	0.93	0.94	0.94	0.92	0.93	0.92	0.92	0.92	0.92	0.91	0.92	0.93	0.92
	0.00	0.00	0.42	0.56	0.67	0.76	0.87	0.84	0.86	0.87	0.88	0.89	0.90	0.88	0.89	0.89	0.92	0.89	0.90	0.91				0.92
duf4	0.00	0.08	0.64	0.85	0.87	0.89	0.88	0.89	0.92	0.92	0.91	0.91	0.92	0.94	0.94	0.93	0.94	0.93	0.92	0.93	0.92	0.93	0.90	0.91
	0.00	0.08	0.63	0.66	0.78	0.77	0.83	0.84	0.87	0.88	0.85	0.88	0.91	0.89	0.91	0.91	0.92	0.92						0.91
duf5	0.00	0.05	0.79	0.89	0.89	0.92	0.92	0.92	0.93	0.93	0.94	0.92	0.93	0.93	0.92	0.93	0.95	0.93	0.91	0.93	0.93	0.92	0.93	0.93
	0.00	0.05	0.79	0.88	0.89	0.89	0.93	0.92	0.92	0.93	0.95	0.94	0.93	0.92										0.93
duf6	0.00	0.26	0.86	0.90	0.88	0.89	0.91	0.91	0.93	0.93	0.92	0.91	0.92	0.93	0.92	0.91	0.94	0.91	0.91	0.90	0.92	0.91	0.92	0.92
	0.00	0.26	0.64	0.69	0.76	0.77	0.87	0.87	0.89	0.91	0.94	0.95	0.94	0.95	0.95									0.92
duf7	0.13	0.28	0.86	0.92	0.89	0.89	0.90	0.91	0.93	0.93	0.93	0.93	0.93	0.95	0.93	0.94	0.95	0.94	0.92	0.92	0.90	0.91	0.92	0.92
	0.00	0.28	0.66	0.90	0.90	0.91	0.89	0.92	0.93	0.93	0.92	0.90	0.92	0.92	0.92									0.92
duf8	0.12	0.35	0.83	0.91	0.89	0.91	0.91	0.91	0.93	0.90	0.89	0.92	0.91	0.93	0.91	0.91	0.92	0.91	0.91	0.91	0.91	0.87	0.89	0.89
	0.00	0.68	0.83	0.86	0.91	0.91	0.90	0.92	0.95	0.93	0.93	0.92	0.91	0.92	0.93	0.91	0.92							0.89
duf9	0.11	0.45	0.89	0.92	0.91	0.91	0.92	0.92	0.93	0.93	0.92	0.91	0.90	0.92	0.93	0.92	0.91	0.91	0.90	0.90	0.92	0.90	0.94	0.93
	0.00	0.45	0.51	0.89	0.91	0.91	0.91	0.91	0.91	0.92	0.91	0.92	0.96	0.94	0.96									0.93
duf10	0.09	0.47	0.87	0.89	0.89	0.90	0.90	0.91	0.93	0.93	0.94	0.95	0.93	0.94	0.92	0.92	0.93	0.92	0.90	0.91	0.90	0.87	0.90	0.91
	0.00	0.71	0.87	0.86	0.89	0.91	0.93	0.93	0.93	0.93	0.92	0.91	0.94	0.92	0.94									0.91
duw2	0.00	0.00	0.04	0.16	0.57	0.59	0.53	0.56	0.68	0.78	0.78	0.81	0.82	0.82	0.82	0.84	0.89	0.89	0.88	0.89	0.90	0.88	0.88	0.88
	0.00	0.00	0.18	0.33	0.57	0.72	0.72	0.73	0.84	0.83	0.78	0.83	0.83	0.85	0.86	0.86	0.86	0.87	0.87	0.87	0.87	0.87	0.89	0.88
duw3	0.12	0.12	0.60	0.68	0.81	0.83	0.86	0.84	0.88	0.91	0.89	0.88	0.88	0.85	0.86	0.84	0.86	0.84	0.88	0.87	0.87	0.82	0.82	0.82
	0.00	0.12	0.22	0.42	0.59	0.74	0.77	0.83	0.84	0.85	0.86	0.90	0.91	0.89	0.87	0.91	0.88	0.87	0.87	0.86	0.86	0.83	0.82	0.82
duw4	0.05	0.23	0.61	0.81	0.84	0.84	0.85	0.86	0.86	0.86	0.86	0.86	0.87	0.86	0.84	0.82	0.86	0.84	0.84	0.84	0.83	0.80	0.81	0.80
	0.05	0.33	0.57	0.72	0.80	0.79	0.81	0.87	0.87	0.89	0.88	0.89	0.89	0.89	0.89	0.89	0.87	0.86	0.90	0.87	0.83	0.83	0.82	0.80
duw5	0.04	0.30	0.79	0.91	0.91	0.90	0.89	0.90	0.91	0.92	0.89	0.88	0.89	0.89	0.84	0.84	0.84	0.82	0.83	0.82	0.79	0.77	0.79	0.79
	0.21	0.38	0.56	0.82	0.91	0.91	0.93	0.92	0.93	0.94	0.94	0.92	0.94	0.92	0.92	0.88	0.88	0.84	0.83	0.82	0.79	0.81		0.79
duw6	0.12	0.24	0.77	0.88	0.89	0.92	0.89	0.91	0.92	0.91	0.92	0.92	0.91	0.92	0.90	0.89	0.86	0.83	0.88	0.84	0.84	0.82	0.84	0.84
	0.16	0.24	0.60	0.78	0.83	0.89	0.92	0.92	0.93	0.89	0.91	0.88	0.90	0.89	0.87	0.87	0.84	0.82	0.82	0.84	0.84	0.84	0.83	0.84
duw7	0.08	0.37	0.82	0.88	0.88	0.88	0.89	0.90	0.90	0.88	0.88	0.88	0.88	0.89	0.85	0.82	0.83	0.83	0.87	0.86	0.87	0.85	0.86	0.84
	0.13	0.37	0.66	0.78	0.82	0.90	0.92	0.93	0.93	0.93	0.94	0.95	0.93	0.92	0.93	0.92	0.90	0.87	0.87	0.84	0.86	0.87	0.86	0.84
duw8	0.26	0.44	0.79	0.87	0.91	0.89	0.89	0.91	0.91	0.91	0.91	0.89	0.89	0.89	0.86	0.88	0.87	0.84	0.88	0.87	0.87	0.84	0.88	0.87
	0.10	0.42	0.73	0.86	0.91	0.91	0.92	0.92	0.92	0.91	0.91	0.88	0.89	0.89	0.88	0.84	0.82	0.82	0.84	0.82	0.81	0.87		0.87
duw9	0.19	0.37	0.81	0.86	0.86	0.86	0.87	0.89	0.91	0.91	0.91	0.88	0.90	0.89	0.87	0.84	0.82	0.82	0.84	0.86	0.85	0.82	0.85	0.86
	0.18	0.37	0.69	0.77	0.86	0.89	0.89	0.91	0.92	0.93	0.92	0.90	0.91	0.87	0.87	0.83	0.83	0.84	0.84	0.82	0.83	0.87		0.86
duw10	0.17	0.44	0.83	0.88	0.89	0.90	0.91	0.92	0.90	0.89	0.89	0.90	0.90	0.88	0.86	0.86	0.85	0.83	0.85	0.84	0.84	0.83	0.86	0.86
	0.21	0.44	0.72	0.76	0.88	0.88	0.89	0.89	0.89	0.90	0.91	0.91	0.91	0.87	0.87	0.86	0.86	0.84	0.82	0.86	0.85	0.84	0.84	0.86

Fig. 1. Classification accuracy of rule-based classifiers constructed in the rule filtering process based on the global and local attribute rankings obtained, for male and female writers.

For both global and local rankings and all discrete variants of the datasets, an increased classification accuracy can be observed for fewer than 24 attributes. The trend is more pronounced for the Female than for the Male dataset. For supervised discretisation methods and female writers, the global ranking provides comparable predictions with fewer attributes than the corresponding local rankings. For dsF, this accuracy is around 98% for a rule set built relying only on 5 attributes. For dsK, an accuracy of 100% occurs for a set of rules based on 16 attributes.

In the case of unsupervised methods, the differences in the results obtained using global and local rankings are small, and an increase in classification accuracy can be observed for a similar reduction in the number of attributes. For the Male dataset and supervised discretisation, the global ranking outperforms the classification accuracy for a reduced set of attributes, that is, 96% for the global ranking, while the reference values are 89% and 90%. In the case of equal width binning, greater performance improvements were observed for a reduced number of attributes than with equal frequency binning. For the duw5 dataset, the global ranking provides classification accuracy of 91% based just on 4 attributes, with the reference point equal to 79%.

4.2. Summary of Obtained Rule Sets Characteristics

A summary of the results based on the global ranking, constructed and verified for all discrete variants of the female and male datasets, is presented in Table 3. It displays the best classification accuracy obtained for the lowest number of features considered. The table lists a discrete variant of a dataset (left-most column), the lowest possible position of an attribute in a given ranking (*Pos*), characteristics of the rule set such as average length (*AvgL*), average support (*AvgS*), information on cardinality of a reduced set of rules (*NoR*), the number of rules for all attributes in considerations (*NAllR*), and classification accuracy (*Cl. Acc* [%]).

Table 3. The best performance and corresponding rule characteristics obtained for global ranking

Discretisation	Female dataset						Male dataset					
method	Pos	AvgL	AvgS	NoR	NAllR	Cl. Acc. [%]	Pos	AvgL	AvgS	NoR	NAllR	Cl. Acc [%]
dsF	5	6.62	2.39	38	4121	98.33	19	4.69	7.53	4027	15283	95.56
dsK	16	4.00	10.16	805	10190	100.00	20	4.76	6.56	8137	20815	95.56
duf2	17	5.16	5.46	9753	103645	97.22	22	5.53	3.89	78731	138910	93.89
duf3	9	3.62	5.00	882	122527	93.89	13	3.93	3.01	5330	135696	94.44
duf4	9	3.17	3.33	1320	81723	95.56	17	3.56	2.00	22841	96327	94.44
duf5	13	3.10	2.05	6391	68994	97.78	17	3.23	1.79	20876	76240	95.00
duf6	13	2.94	1.86	6398	58327	97.22	17	3.06	1.63	18959	65184	93.89
duf7	4	2.05	5.34	104	49026	96.11	14	2.85	1.68	8056	55184	95.00
duf8	5	2.12	3.80	222	42490	93.89	9	2.52	1.95	2021	47511	93.33
duf9	13	2.58	1.69	5372	37750	97.22	23	2.86	1.48	36330	42278	93.89
duf10	13	2.45	1.65	5228	34155	95.00	12	2.41	1.65	4328	38670	95.00
duw2	17	4.17	8.99	435	2094	86.67	21	4.18	8.64	895	1509	90.00
duw3	9	3.00	7.11	168	26025	97.22	10	3.69	5.69	426	32447	90.56
duw4	13	3.93	3.47	2390	46480	95.00	13	3.73	3.65	2324	47574	86.67
duw5	9	3.09	3.99	633	67054	96.11	10	3.29	3.43	1272	79561	91.67
duw6	13	3.51	2.32	4598	75888	96.67	9	3.07	3.05	956	77033	92.22
duw7	12	3.20	2.30	3336	70152	96.67	8	2.78	2.78	842	75733	90.00
duw8	4	1.93	6.36	84	60422	95.00	9	2.89	2.38	1493	72675	91.11
duw9	14	3.04	1.85	7296	59332	94.44	11	2.93	2.09	2728	68722	91.11
duw10	5	2.11	3.72	209	54187	94.44	8	2.74	2.21	1109	61920	91.67

For the dsK variant of the female writer dataset, the proposed global ranking provides the best classification accuracy (100%) among all considered cases. This rule set contains 16 of the 24 attributes, and the number of rules was reduced by about 90%. For the Male dataset, 8 was the smallest number of attributes in rule sets induced for duw7 and duw10 sources. These values exceeded the reference point and provided a reduction of 99% of the rule set. For the Female dsF dataset, the largest reduction was achieved for the rule set based on 5 attributes containing only 38 (from the total of 4121) rules with a performance 98.3%.

For all discrete variants of datasets, the proposed methodology for ranking construction and the mechanism of decision rule filtering resulted in increasing accuracy above the reference point for a decreased number of attributes. This observation confirms the efficiency of the novel mechanism for constructing global rankings based on distributed local knowledge.

5. Conclusions

Discretisation is an important issue that should be considered in the context of the use of various machine learning algorithms, which are not always adapted to work with continuous data. In the approach proposed in the paper, different variants of a discretised dataset are considered as distributed local sources. For such distributed data, local attribute rankings are constructed using the properties of the greedy algorithm. These local rankings are exploited for the construction of a novel global ranking, which can be seen as a form of general knowledge that also considers the unique characteristics of local variants of the data. Processing, in which global knowledge based on local sources is accumulated, is important for creating modern information systems.

In the experiments, all constructed rankings were involved in the procedure of filtering decision rules. The obtained promising results, with multiple cases of vastly reduced rule sets offering often improved performance, validated the proposed research framework. They also indicate the merit of further research work, which will concern other types of classifiers relying on the proposed global ranking, and other approaches for local ranking construction.

Acknowledgements

The research works presented were carried out at the Institute of Computer Science, University of Silesia in Katowice, Sosnowiec, Poland, and in the Department of Computer Graphics, Vision and Digital Systems (RAU-6, 2024), at the Silesian University of Technology, Gliwice, Poland.

References

1. Bazan, J., Szczuka, M.: The rough set exploration system. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets III, LNCS, vol. 3400, pp. 37–56. Springer (2005)
2. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuousvalued attributes for classification learning. In: 13th International Joint Conference on Articial Intelligence. vol. 2, pp. 1022–1027. Morgan Kaufmann Publishers (1993)
3. Feige, U.: A threshold of $\ln n$ for approximating set cover. In: Journal of the ACM (JACM), vol. 45, pp. 634–652. ACM New York (1998)
4. Garcia, S., Luengo, J., Saez, J., Lopez, V., Herrera, F.: A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering **25**(4), 734–750 (2013)
5. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
6. Kononenko, I.: On biases in estimating multi-valued attributes. In: 14th International Joint Conference on Articial Intelligence. pp. 1034–1040 (1995)
7. Moshkov, M., Zielosko, B., Tetteh, E.T., Glid, A.: Learning decision rules from sets of decision trees. In: Buchmann, R.A., et al. (eds.) Information Systems Development: Artificial Intelligence for Information Systems Development and Operations (ISD2022 Proceedings), Cluj-Napoca, Romania, 31.08-2.09.2022. Risoprint/AIS (2022)
8. Moshkov, M.J., Piliszczuk, M., Zielosko, B.: Greedy algorithm for construction of partial association rules. Fundamenta Informaticae **92**(3), 259–277 (2009)
9. Pawlak, Z., Skowron, A.: Rough sets and boolean reasoning. Information Sciences **177**(1), 41–73 (2007)
10. Stańczyk, U., Zielosko, B., Baron, G.: Significance of single-interval discrete attributes: Case study on two-level discretisation. Applied Sciences **14**(10) (2024)
11. Stepaniuk, J., Skowron, A.: Three-way approximation of decision granules based on the rough set approach. International Journal of Approximate Reasoning **155**, 1–16 (2023)
12. Wu, H., Zhang, Z., Wu, Q.: Exploring syntactic and semantic features for authorship attribution. Applied Soft Computing **111**, 107815 (2021)