Weighting and Pruning of Decision Rules by Attributes and Attribute Rankings

Urszula Stańczyk(✉)

Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
urszula.stanczyk@polsl.pl

Abstract. Pruning is a popular post-processing mechanism used in search for optimal solutions when there is insufficient domain knowledge to either limit learning data or govern induction in order to infer only the most interesting or important decision rules. Filtering of generated rules can be driven by various parameters, for example explicit rule characteristics. The paper presents research on pruning rule sets by two approaches involving attribute rankings, the first relaying on selection of rules referring to the highest ranking attributes, which is compared to weighting of rules by calculated quality measures dependent on weights coming from attribute rankings that results in rule ranking.

Keywords: Decision rules · Pruning · Weighting · Attribute · Ranking

1 Introduction

Rule classifiers express patterns discovered in data in learning processes through conditions on attributes included in the premises and pointing to specific classes [5]. A variety of available approaches to induction enable construction of classifiers with minimal numbers of constituent rules, with all rules that can be inferred from the training samples, or with subsets of interesting elements [3].

To limit the number of considered rules [9] either pre-processing can be employed, with reducing rather data than rules, by selection of features or instances, or in-processing relaying on induction of only those rules that satisfy given requirements, or post-processing, which implements pruning mechanisms and rejection of some unsatisfactory rules. The paper focuses on this latter approach.

One of the most straightforward ways to prune rules and rule sets involves exploiting direct parameters of rules, such as their support, length [11], strength [1]. Also specific condition attributes can be taken into account and indicate rules to be selected by appearing in their premises [12]. Such process can lead to improved performance or structure and in the presented research it is compared to weighting of rules by calculated quality measures, also based on attributes [13], both procedures actively using rankings of considered characteristic features [7].

The paper is organised as follows. Section 2 briefly describes some elements of background, that is feature weighting and ranking, and aims of pruning of

T. Czachórski et al. (Eds.): ISCIS 2016, CCIS 659, pp. 106–114, 2016.
DOI: 10.1007/978-3-319-47217-1_12

rules and rule sets. Section 3 explains the proposed research framework, details experimental setup, and gives test results. Section 4 concludes the paper.

2 Background

The research described in this paper incorporates characteristic feature weights and rankings into the problem of pruning of decision rules and rule sets.

2.1 Feature Ranking

Roles of specific features exploited in any classification task can vary in significance and relevance in a high degree. The importance of individual attributes can be discovered by some approach leading to their ranking, that is assigning values of a score function which causes putting them in a specific order [7].

Rankings of characteristic features can be obtained through application of statistical measures, machine learning approaches, or systematic procedures [12]. The former assign calculated weights to all variables, while the latter can return only the positions in a ranking, reflecting discovered order of relevance.

Information Gain coefficient (*InfoGain, IG*) is defined by employing the concept of entropy from information theory for attributes and classes:

$$InfoGain(Cl, a_f) = H(Cl) - H(Cl|a_f), \tag{1}$$

where $H(Cl)$ denotes the entropy for the decision attribute Cl and $H(Cl|a_f)$ condition entropy, that is class entropy while observing values of attribute a.

An attribute relevance measure can be based on rule length [11], with special attention given to the shortest rules that often possess good generalisation properties:

$$MREVM(a) = Nr(a, MinL) : Nr(a, MinL + 1), \tag{2}$$

where $Nr(a, L)$ denotes the number of rules with length L in which attribute a appears, and $MinL$ is the length of the shortest rule containing a. The attribute ranking constructed in this way is wrapped around the specific inducer, not its performance, since other parameters of rules are disregarded, but structure.

2.2 Pruning of Decision Rules

To limit the number of rules three approaches can be considered [8]:

- pre-processing — the input data is reduced before the learning stage starts by rejecting some examples or cutting down on characteristic features. With less data to infer from, it follows that fewer rules are induced.
- at the algorithm construction stage — by implementation of specific procedures only some rules meeting requirements are found instead of all possible.
- post-processing — the set of inferred rules is analysed and some of its elements discarded while others selected.

When lower numbers of rules are found the learning stage can be shorter, yet solutions are not necessarily the best. If higher numbers of rules are generated, more thorough and in-depth analysis is enabled, yet even for rule sets with small cardinalities some measures of quality or interestingness can be employed [6].

Rule quality can be weighted by conditional attributes [13]:

$$QM(r_i) = \prod_{j=1}^{K_{r_i}} w(a_j), \tag{3}$$

where K_{r_i} denotes the number of conditions included in rule r_i and $w(a_j)$ weight of a_j attribute taken from a ranking. It is assumed that $w(a_j) \in (0, 1]$.

3 Experimental Setup and Obtained Results

The research works presented were executed within the general framework:

- Initial preparation of learning and testing data sets
- Obtaining rankings of attributes
- Induction of decision algorithms
- Pruning of decision rules in two approaches:
  - Selecting rules referring to specific attributes in the ranking
  - Calculating measures for all rules while exploiting weights assigned to positions in the attribute rankings, which led to weighting of rules and their rankings, and from these rankings rules in turn were selected
- Comparison and analysis of obtained test results

Steps of these procedures are described in the following subsections.

3.1 Input Datasets

As a domain of application for the research stylometric analysis of texts was selected. Stylometry enables authorship attribution while basing on employed linguistic characteristic features. Typically they refer to lexical and syntactic markers, giving frequencies of occurrence for selected function words and punctuation marks that reflect individual habits of sentence and paragraph formation.

Learning and testing samples corresponded to parts of longer works by two pairs of writers, female and male, giving binary classification with balanced data.

As attribute values specified usage frequencies of textual descriptors, they were small fractions, which means that for data mining there was needed either some technique that can deal efficiently with continuous numbers, or some discretization strategy was required [2]. Since regardless of a selected method discretization always causes some loss of information, it was not attempted.

3.2 Rankings of Attributes

In the research presented two attribute rankings were tested. The first one relied on statistical properties detected in input datasets and was completely independent on the classifier used later for prediction, and the other was wrapped around characteristics of induced rules, observing how often each variable occurs in shortest rules, which usually are of higher quality as they are better at generalisation and description of detected patterns than those with many conditions. Orderings of variables for both rankings and both datasets are given in Table 1.

Table 1. Rankings of condition attributes

No	$w(a)$	Female writers		Male writers	
		InfoGain	*MREVM*	*InfoGain*	*MREVM*
1	1	not	not	and	and
2	1/2	:	:	that	by
3	1/3	;	but	by	from
4	1/4	,	and	but	of
5	1/5	-	.	from	in
6	1/6	on	,	what	:
7	1/7	?	by	for	!
8	1/8	(	for	-	on
9	1/9	as	to	?	,
10	1/10	but	this	if	as
11	1/11	by	as	at	(
12	1/12	that	what	with	with
13	1/13	for	!	not	
14	1/14	to	from	:	this
15	1/15	at	?	to	at
16	1/16	.	-	in	not
17	1/17	and	of	(	;
18	1/18	in	in	as	?
19	1/19	this	that	!	-
20	1/20	!	with	;	to
21	1/21	with	if	on	if
22	1/22	of	at	.	what
23	1/23	what	(	of	for
24	1/24	if	on	this	but
25	1/25	from	;	,	that

InfoGain returns a specific score for each feature while *MREVM* gives a ratio. To unify numbers considered as attribute weights they were assigned in an arbitrary manner, listed in column denoted $w(a)$, and equal $1/i$, where i is a position in the ranking. Thus the distances between weights decrease while going down the ranking. It is assumed that each variable has nonzero weight.

3.3 DRSA Rule Classifiers

The rules were induced with the help of 4eMka Software (developed at the Poznań University of Technology, Poland), which implements Dominance-Based Rough Set Approach (DRSA). By substituting the original indiscernibility relation [4] of classical rough sets with dominance DRSA observes ordinal properties in datasets and enables both nominal and ordinal classification [10].

As the reference points classification systems with all rules on examples were taken. For female writers the algorithm consisted of 62383 rules, which with constraints on minimal rule support to be equal at least 66 resulted in 17 decision rules giving the maximal classification accuracy of 86.67 %. For male writers the algorithm contained 46191 rules, limited to 80 by support equal at least 41, and it gave the correct recognition of 76.67 % of testing samples. In all cases ambiguous decisions were treated as incorrect, without any further processing.

3.4 Pruning of Rule Sets by Attributes

Selection of decision rules while following attribute rankings was executed as follows: at i-th step only the rules with conditions on the i highest ranking features were taken into account. The rules could refer to all or some proper subsets of variables considered, and these with at least one condition on any of lower ranking attributes were discarded. Thus at the first step only rules with single conditions on the highest ranking variable were filtered, while at the last 25-th step all features and all rules were included. For example at 5-th step for female writer dataset for *InfoGain* ranking only rules referring to any combination of attributes: not, colon, semicolon, comma, hyphen, were selected. The detailed results for both datasets and both rankings are listed in Table 2.

It can be observed that with each variable added to the studied set the numbers of recalled rules rose significantly, but the classification accuracy equal to or even higher than the reference points was detected quite soon in processing, for *InfoGain* for female dataset after selection of just four highest ranking attributes, for male writers and *MREVM* for just three most important features.

3.5 Pruning of Rule Sets Through Rule Rankings

Calculation of *QM* measure for rules can be understood as translating feature rankings into rule rankings. Depending on cardinalities of subsets of rules selected at each step, the total number of executed steps can significantly vary. The minimum is obviously one, while the maximum can even equal the total number of rules in the analysed set, if with each step only a single rule is added.

Table 2. Characteristics of decision algorithms with pruning of rules referring to specific conditional attributes: N indicates the number of considered attributes, (a) number of recalled rules, (b) maximal classification accuracy [%], (c) minimal support required of rules, (d) number of rules satisfying condition on support

N	Female								Male							
	InfoGain				*MREVM*				*InfoGain*				*MREVM*			
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
1	10	61.11	55	4	10	61.11	55	10	6	13.33	14	4	6	13.33	14	4
2	27	81.11	55	13	27	81.11	55	13	15	21.11	9	8	27	55.56	9	23
3	36	81.11	55	13	56	82.22	32	27	45	61.11	6	35	80	**80.00**	25	39
4	79	**86.67**	55	27	97	81.11	55	14	73	61.11	10	41	127	**80.00**	25	50
5	91	86.67	55	27	203	82.22	44	26	153	75.56	21	62	219	**81.11**	20	115
6	128	86.67	55	27	324	**86.67**	55	28	198	75.56	21	65	290	**80.00**	**41**	25
7	167	86.67	55	27	578	86.67	55	30	239	75.56	26	46	562	75.56	41	28
8	202	86.67	55	27	877	86.67	**66**	13	307	75.56	21	72	778	75.56	41	29
9	356	86.67	**66**	11	1317	86.67	66	13	422	75.56	21	79	1073	75.56	41	30
10	570	86.67	66	11	1923	86.67	66	15	531	75.56	21	89	1355	75.56	41	31
11	1011	86.67	66	12	2755	86.67	66	16	689	75.56	32	44	1591	**78.89**	41	41
12	1415	86.67	66	14	3793	86.67	66	16	866	75.56	32	48	1975	**76.67**	41	45
13	2201	86.67	66	14	4995	86.67	66	**17**	1395	75.56	32	65	3169	76.67	41	45
14	3137	86.67	66	14	6671	86.67	66	17	1763	75.56	32	67	4456	**78.89**	41	53
15	4215	86.67	66	14	8099	86.67	66	17	2469	75.56	32	67	5774	**78.89**	41	53
16	5473	86.67	66	14	9485	86.67	66	17	3744	75.56	**41**	42	8476	76.67	41	63
17	7901	86.67	66	14	13255	86.67	66	17	4336	75.56	41	56	11055	76.67	41	66
18	10732	86.67	66	14	17589	86.67	66	17	5352	75.56	41	57	13428	76.67	41	69
19	14187	86.67	66	16	21238	86.67	66	17	7214	75.56	41	60	16188	76.67	41	69
20	18087	86.67	66	**17**	26821	86.67	66	17	9819	75.56	41	63	22035	76.67	41	75
21	23408	86.67	66	17	33834	86.67	66	17	14282	75.56	41	64	26674	76.67	41	78
22	31050	86.67	66	17	43225	86.67	66	17	18590	75.56	41	64	30846	76.67	41	78
23	39235	86.67	66	17	52587	86.67	66	17	26474	75.56	41	70	36630	76.67	41	78
24	48583	86.67	66	17	58097	86.67	66	17	35014	**76.67**	41	79	40024	76.67	41	78
25	62383	86.67	66	17					46191	76.67	41	**80**				

On the other hand, once the core sets of rules, corresponding to the decision algorithms limited by constraints on minimal support of rules and giving the best results for the complete algorithms, are retrieved, there is little point in continuing, thus the results presented in Table 3 stop when only fractions of the whole rule sets are recalled, for female writers just few hundreds, and for male writers close to ten thousand (still less than a quarter of the original algorithm).

3.6 Summary of the Best Results

Out of the two tested and compared approaches to rule filtering, selection governed by attributes included when following their rankings enabled to reject more rules from the reference algorithms, even over 35 % and 48 %, respectively for female and male datasets, with prediction at the reference level. For male writers recognition could be increased (at maximum by over 4 %) either with keeping or lowering constraints on minimal support required of rules.

Table 3. Characteristics of decision algorithms with pruning of rules while weighting them by measures based on rankings of conditional attributes: N indicates the weighting step, (a) number of recalled rules, (b) maximal classification accuracy [%], (c) minimal support required of rules, (d) number of rules satisfying condition on support

N	Female								Male							
	InfoGain-RDD				*MREVM*-RDD				*InfoGain*-RDD				*MREVM*-RDD			
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
1	10	61.11	55	4	10	61.11	55	4	36	55.56	9	26	27	55.56	9	23
2	12	61.11	55	4	12	61.11	55	4	113	61.11	13	58	48	61.11	13	39
3	29	81.11	55	13	39	83.33	32	23	128	61.11	13	62	60	61.11	13	45
4	46	**87.78**	52	25	55	84.44	14	37	154	61.11	13	70	71	61.11	13	53
5	48	87.78	52	25	70	84.44	14	45	185	66.67	10	99	112	**80.00**	25	52
6	67	87.78	52	25	104	**87.78**	52	28	215	66.67	10	120	127	73.33	26	56
7	71	87.78	52	25	129	87.78	52	31	231	66.67	10	130	149	73.33	26	63
8	80	**90.00**	46	29	161	87.78	52	36	265	73.33	26	86	189	73.33	26	66
9	94	90.00	46	33	182	87.78	52	39	301	73.33	26	90	251	73.33	26	79
10	106	90.00	46	33	212	**88.89**	52	45	329	73.33	26	99	288	73.33	26	87
11	131	90.00	46	38	226	88.89	52	48	384	73.33	26	110	331	73.33	**41**	33
12	166	**86.67**	**66**	12	265	**86.67**	**66**	16	396	73.33	26	116	368	73.33	41	41
13	181	86.67	66	14	279	86.67	66	**17**	511	73.33	26	124	382	73.33	41	44
14	202	86.67	66	14	327	86.67	66	17	667	75.56	25	143	451	73.33	41	48
15	206	86.67	66	14	339	86.67	66	17	794	75.56	32	91	483	75.56	27	130
16	221	86.67	66	14	362	86.67	66	17	912	73.33	32	94	514	**76.67**	27	135
17	237	86.67	66	14	388	86.67	66	17	949	73.33	26	148	624	75.56	37	74
18	268	86.67	66	14	441	86.67	66	17	1011	73.33	**41**	54	848	75.56	37	77
19	285	86.67	66	16	452	86.67	66	17	1117	75.56	27	153	937	**78.89**	35	87
20	305	86.67	66	**17**	498	86.67	66	17	1189	75.56	27	155	1236	76.67	35	91
21									1228	75.56	27	157	1965	76.67	41	65
22									1900	75.56	**41**	61	2160	76.67	41	67
23									1993	75.56	41	63	2264	76.67	41	68
24									2667	**76.67**	41	67	3291	76.67	41	71
25									3610	76.67	41	68	4036	76.67	41	72
26									4577	76.67	41	70	4519	76.67	41	74
27									4825	76.67	41	71	5637	76.67	41	76
28									5725	76.67	41	74	6269	76.67	41	77
29									7901	76.67	41	76	9820	76.67	41	79
30									9250	76.67	41	78	9830	76.67	41	**80**
31									9394	76.67	41	79	9841	76.67	41	80
32									9404	76.67	41	**80**	9844	76.67	41	80

When rules were wighted, ranked, and then selected the quality of prediction was enhanced at maximum by over 3 % for both datasets, and for female and male writers datasets respectively over 29 % and 18 % of rules could be pruned.

For female dataset for both approaches to rule pruning better results were obtained while exploiting *InfoGain* attribute ranking, and for male dataset the same can be stated for *MREVM* ranking.

4 Conclusions

The paper presents research on selection of decision rules while following rankings of considered conditional attributes and exploiting weights assigned to them, which constitute alternatives to the popular approaches to rule filtering. Two ways to prune rules were compared, the first relying on selection of the rules with conditions only on the highest ranking attributes, while those referring to lower ranking features were rejected. Within the second methodology, the weights of attributes from their rankings formed a base from which for all rules the defined quality measures were calculated, and their values led to rule rankings. Next, the highest ranking rules were filtered out. For both described approaches two attribute rankings were tested, and the test results show several possibilities of constructing optimised rule classifiers, either with increased recognition, decreased lengths of decision algorithms, or both.

Acknowledgments. The research presented was performed at the Silesian University of Technology, Gliwice, Poland, within the project BK/RAu2/2016.

References

1. Amin, T., Chikalov, I., Moshkov, M., Zielosko, B.: Relationships between length and coverage of decision rules. Fundamenta Informaticae **129**, 1–13 (2014)
2. Baron, G.: On approaches to discretization of datasets used for evaluation of decision systems. In: Czarnowski, I., Caballero, A., Howlett, R., Jain, L. (eds.) Intelligent Decision Technologies 2016. Smart Innovation, Systems and Technologies, vol. 56, pp. 149–159. Springer, Switzerland (2016)
3. Bayardo Jr., R., Agrawal, R.: Mining the most interesting rules. In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 145–154 (1999)
4. Cyran, K.A., Stanczyk, U.: Indiscernibility relation for continuous attributes: application in image recognition. In: Kryszkiewicz, M., Peters, J.F., Rybiński, H., Skowron, A. (eds.) RSEISP 2007. LNCS (LNAI), vol. 4585, pp. 726–735. Springer, Heidelberg (2007)
5. Fürnkranz, J., Gamberger, D., Lavrač, N.: Foundations of Rule Learning. Springer, Heidelberg (2012)

6. Gruca, A., Sikora, M.: Rule based functional description of genes – estimation of the multicriteria rule interestingness measure by the UTA method. Biocybernetics Biomed. Eng. **33**, 222–234 (2013)
7. Mansoori, E.: Using statistical measures for feature ranking. Int. J. Pattern Recog. Artitf. Intell. **27**(1), 1350003–1350014 (2013)
8. Sikora, M.: Induction and pruning of classification rules for prediction of microseismic hazards in coal mines. Expert Syst. Appl. **38**(2), 6748–6758 (2013)
9. Sikora, M., Wróbel, Ł.: Data-driven adaptive selection of rules quality measures for improving the rules induction algorithm. In: Kuznetsov, S.O., Ślęzak, D., Hepting, D.H., Mirkin, B.G. (eds.) RSFDGrC 2011. LNCS (LNAI), vol. 6743, pp. 278–285. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21881-1_44
10. Słowiński, R., Greco, S., Matarazzo, B.: Dominance-based rough set approach to reasoning about ordinal data. In: Kryszkiewicz, M., Peters, J.F., Rybinski, H., Skowron, A. (eds.) RSEISP 2007. LNCS (LNAI), vol. 4585, pp. 5–11. Springer, Heidelberg (2007). doi:10.1007/978-3-540-73451-2_2
11. Stańczyk, U.: Decision rule length as a basis for evaluation of attribute relevance. J. Intell. Fuzzy Syst. **24**(3), 429–445 (2013)
12. Stańczyk, U.: Selection of decision rules based on attribute ranking. J. Intell. Fuzzy Syst. **29**(2), 899–915 (2015)
13. Stańczyk, U.: Measuring quality of decision rules through ranking of conditional attributes. In: Czarnowski, I., Caballero, A., Howlett, R., Jain, L. (eds.) Intelligent Decision Technologies 2016. Smart Innovation, Systems and Technologies, vol. 56, pp. 269–279. Springer, Switzerland (2016)