Separability of Split Value criterion with weighted separation gains

Krzysztof Grąbczewski

Department of Informatics, Nicolaus Copernicus University, Toruń, Poland
http://www.is.umk.pl/~kg
mailto:kg@is.umk.pl

Abstract. An analysis of the Separability of Split Value criterion in some particular applications has led to conclusions about possible improvements of the criterion. Here, the new formulation of the SSV criterion is presented and examined. The results obtained for 21 different benchmark datasets are presented and discussed in comparison with the most popular decision tree node splitting criteria like information gain and Gini index. Because the new SSV definition introduces a parameter, some empirical analysis of the new parameter is presented. The new criterion turned out to be very successful in decision tree induction processes.

Keywords: Decision trees, split criteria, separability.

1 Introduction

Since the advent of the first decision tree (DT) learning algorithms, several decades ago, the researchers have come up with a number of criteria (called *split criteria* or *split quality measures* or *selection measures*) for top-down DT construction [1, 10, 8]. Some comparisons of such criteria [9, 2] have been published. Although they still do not exhaustively explore the subject, many researchers claim that the criteria measuring split quality do not significantly differ from each other. It is an over-interpretation of the results, as it will be visible below (although it is not the main subject of this article). The fact is, that there is no approach outperforming all the others in all possible applications, but for many datasets, the results obtained with different methods are significantly different. Therefore, there is still room for improvement of existing criteria and defining new ones, if only they introduce some new quality. Provided many different algorithms one can analyze them and select the most adequate ones for modeling particular data.

The Separability of Split Value (SSV) criterion [4, 5] was defined as an alternative to the most popular criteria like the measure of *information gain* or *Gini index*. Here, a modified version of the SSV is presented and examined. To keep the comparison fair, it was implemented within Intemi [6, 7] – a system that has been designed and implemented recently as a perfect framework for such tasks.

The following section shortly presents SSV and other criteria used in the most popular DT algorithms. Then, section 3 defines the new version of the SSV. Thorough comparative analysis of the new criterion is contained within section 4.

2 Split criteria

Although, the trees built in the analysis presented below, are binary, the split quality measures used, are more general and can estimate multipart splits. In general a split s can be defined as a collection $s_1, \ldots, s_{n_s}$ that unambiguously determines a partition $\{D_{s_i} : i = 1, \ldots, n_s\}$ for each data set D in given domain $\mathcal{D}$. Binary univariate splits are defined differently for ordered and unordered features of $\mathcal{D}$. For an ordered feature F, it is determined by a split threshold t and splits data into two subsets of elements x satisfying $F(x) < t$ and $F(x) \geq t$ respectively. For unordered feature F, each binary split is determined by a set of possible values V of F and splits data into two subsets of elements x satisfying $F(x) \in V$ and $F(x) \notin V$ respectively.

The most popular approach to measure split quality is the use of the *purity gain* (or in other words: *impurity reduction*) criterion:

$$\Delta I(s, D) \stackrel{\text{def}}{=} I(D) - \sum_{i=1}^{n_s} p_i I(D_{s_i}). \tag{1}$$

It can be used with different impurity measures I, for example, the one based on classification accuracy:

$$I_A(D) \stackrel{\text{def}}{=} \frac{\max_{C \in \mathcal{C}} |D_C|}{|D|}, \tag{2}$$

where $\mathcal{C}$ is the set of classes of objects from $\mathcal{D}$, $D_C = D \cap C$ and $|\cdot|$ is the set cardinality operator.

The most popular impurity measures are the *Gini index* of CART [1]:

$$I_G(D) \stackrel{\text{def}}{=} 1 - \sum_{C \in \mathcal{C}} P(C|D)^2, \tag{3}$$

and the one based on entropy, used in ID3, its many descendants and also in CART:

$$I_E(D) \stackrel{\text{def}}{=} -\sum_{C \in \mathcal{C}} P(C|D) \log_2 P(C|D). \tag{4}$$

Here, $P(C|D)$ is shorthand for $P(x \in C | x \in D)$.

The purity gain criterion with entropy measure is called the *information gain* (IG) criterion. To overcome its bias towards multivalued features (when building multisplit trees), C4.5 [10] introduced the *information gain ratio* (IGR) which is the IG divided by the entropy of the split:

$$IGR(s, D) \stackrel{\text{def}}{=} \frac{\Delta I_E(s, D)}{\sum_i p_i \log_2 p_i}, \tag{5}$$

where $p_i = \frac{|D_{s_i}|}{|D|}$.

The SSV criterion is not based on the purity gain rule, but on a simple idea that splitting pairs of vectors belonging to different classes is advantageous, while splitting pairs of vectors of the same class should be avoided if possible. It has got two forms:

$$\text{SSV}(s, D) \stackrel{\text{def}}{=} 2 \cdot \text{SSV}_1(s, D) - \text{SSV}_2(s, D), \tag{6}$$

$$\mathrm{SSV}_{lex}(s, D) \stackrel{\text{def}}{=} \Big(\mathrm{SSV}_1(s, D), -\mathrm{SSV}_3(s, D)\Big), \tag{7}$$

where:

$$\mathrm{SSV}_1(s, D) \stackrel{\text{def}}{=} \sum_{i=1}^{n_s} \sum_{j=i+1}^{n_s} \sum_{C \in \mathcal{C}} |D_{s_i,C}| \cdot |D_{s_j} \setminus D_{s_j,C}|, \tag{8}$$

$$\mathrm{SSV}_2(s, D) \stackrel{\text{def}}{=} \sum_{C \in \mathcal{C}} (D_C - \max_{i=1,\ldots,n_s} |D_{s_i,C}|), \tag{9}$$

$$\mathrm{SSV}_3(s, D) \stackrel{\text{def}}{=} \sum_{i=1}^{n_s} \sum_{j=i+1}^{n_s} \sum_{C \in \mathcal{C}} |D_{s_i,C}| \cdot |D_{s_j,C}|. \tag{10}$$

The SSV_{lex} version provides pairs of values, which are compared in lexicographic order, so the second value is important only in the case of equal first elements.

Many other criteria have also been proposed. A review can be found in [8].

3 Weighting separability gains

A toy example data set presented in figure 1 reveals the weakness of the original SSV definitions that inspired the modification described here. The example can be solved

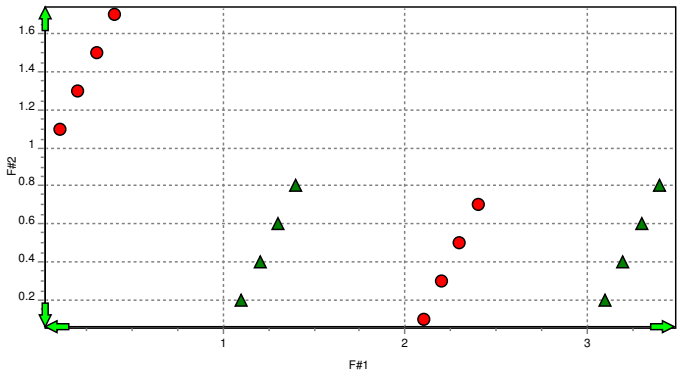

Fig. 1. Example 2D data

by quite simple DTs: one of them splits the scope of the feature `F#1` in three points, another one splits `F#2` in one point and `F#1` in two points. The solutions are trivial to find with a quick look at the scatter plot, so DT learning algorithms should not miss them either. Recursive partitioning with IG or Gini index finds the solution, but with SSV of the form of (6) or (7), it does not. The topmost split in the resulting DT is defined by the condition `F#1 < 0.65`, because it generates the split (3,6) vs (5,2),

i.e. keeps 3 circles and 6 triangles below the threshold and 5 circles and 2 triangles above it. This gives $\mathrm{SSV}_1 = 36$, $\mathrm{SSV}_2 = 5$, $\mathrm{SSV}_3 = 27$, while in the case of a split (4,0) vs (4,8) we have $\mathrm{SSV}_1 = 32$, $\mathrm{SSV}_2 = 6$, $\mathrm{SSV}_3 = 16$. Therefore, both definitions of SSV favor the former split (more pairs are separated).

Since manipulating the penalty term of (6) to repair such cases may easily spoil the functionality in other cases, the idea followed here is to weight the pairs of separated objects when counting the separability index. The heuristic is based on the idea that separating pairs of objects is more advantageous, when the objects belong to the majority classes within their sides of the split, and less valuable if the objects are still misclassified after the split. Therefore, we introduce a parameter weight α as a factor to diminish the contribution of the minority objects in separated pairs, and obtain the following definition:

$$SSV_\alpha(s, D) \stackrel{\text{def}}{=} \sum_{i=1}^{n_s} \sum_{j=i+1}^{n_s} \sum_{\substack{A \in C \\ B \in C \\ A \neq B}} W_\alpha(D_{s_i}, A) \cdot |D_{s_i,A}| \cdot W_\alpha(D_{s_j}, B) \cdot |D_{s_j,B}|, \quad (11)$$

where

$$W_\alpha(D, C) = \begin{cases} 1 & \text{if } C \text{ is the majority class within } D, \\ \alpha & \text{otherwise.} \end{cases} \quad (12)$$

Such definition introduces three levels of contribution of the separated pairs (1, α and α^2), dependent on whether the objects represent the majorities or not. If more than one class is represented in a sample with maximum count, one of them is arbitrarily selected as the majority class (in practice, the one with the smallest index).

4 The analysis

To examine the advantages of the new definition of SSV, we compare the results obtained with different versions of the criterion and four other split criteria described in section 2. To provide an equal chance comparison, all the other components of the decision tree induction algorithm are the same in the case of all criteria and the algorithms are run for the same training and test data. 10-fold cross-validation (CV) tests were repeated 10 times with different randomization, but each algorithm received the same sets in all 100 training and testing runs. Moreover, because pruning was made with the *cost complexity algorithm* [1] based on inner (i.e. performed within the training data) 10-fold cross-validation, the inner data splits were also exactly the same for all the algorithms being compared. The tests were performed for 21 different datasets from the UCI repository [3], summarized in table 1. The selection of datasets was done before the tests (no test results were discarded so as to obtain satisfactory but unfair conclusions). Some datasets were not selected because they would need some preprocessing (for example to delete classes with very few examples) and that would spoil the clarity of the tests. The mushroom data was rejected, because of a priori knowledge, that all the DT algorithms would be 100% accurate with zero variance.

Symbol	Dataset	classes	instances	features	ordered f.
APP	appendicitis	2	106	7	7
AUS	Australian credit	2	690	14	6
BRE	breast cancer (Wisconsin)	2	699	9	9
FLA	flag	8	194	28	10
GLA	glass	6	214	9	9
HEA	heart	2	303	13	13
IMA	image	7	2310	19	19
ION	ionosphere (trn+tst)	2	351	34	34
IRI	iris	3	150	4	4
KVK	kr-vs-kp	2	3196	36	0
LBR	Ljubjlana breast cancer	2	286	9	1
LET	letter recognition	26	20000	16	16
PIM	Pima indians diabetes	2	768	8	8
SON	sonar	2	208	60	60
SOY	soybean large	19	307	35	0
SPL	splice	3	3190	60	0
THY	thyroid (trn+tst)	3	7200	21	6
VOT	vote	2	435	16	0
VOW	vowel	6	871	3	3
WAV	waveform	3	5000	21	21
WIN	wine	3	178	13	13

Table 1. Datasets used for the tests.

4.1 Comparison of the split criteria

The mean accuracy and the standard deviation within the population of 100 test results (10 times 10-fold CV) for each dataset are presented in table 2. The results with the highest mean for given dataset are underlined. Bold face and italics mark the results that are not statistically significantly different than the one with the highest mean: bold face represents t test significance decision and italics—the Wilcoxon test judgment. The confidence level of 0.01 was applied in both kinds of tests.

Table 3 reports the counts of obtaining the best mean result for a dataset and the counts of obtaining insignificant difference of the mean values with the best result. It can be seen from both tables that the new definition of SSV is very successful. The highest mean is obtained for 5 datasets, but to be fair, we should count 8 wins, because if we had not included the two older versions of SSV in the comparison, their wins (three cases) would move to the account of SSV_α. 8 wins is the maximum obtained also by IG criterion. More sensible information (than the number of the highest means obtained) is contained in the counts of obtaining insignificant differences from the best results (the last two rows of table 3). These numbers also put IG and SSV_α at the top.

Another interesting point (a bit aside the main topic of the article, but worth to be mentioned) is that for 7 of the 21 datasets there is a single definite winner i.e. all other methods obtained significantly worse results (with 0.01 significance level). IG outperformed all the others in 5 cases (IMA, KVK, LET, SPL and WIN). The new

Data	Accuracy	IG	IGR	Gini	SSV	SSV_{lex}	$SSV_{\alpha=0.5}$
APP	81,69 ±9,23	83,13 ±9,46	82,05 ±10,52	82,12 ±9,41	***86,79*** ±***9,10***	***86,70*** ±***9,55***	***86,65*** ±***9,92***
AUS	***85,29*** ±***4,29***	***85,06*** ±***4,42***	***85,22*** ±***4,25***	***84,75*** ±***4,11***	***85,07*** ±***4,39***	***85,22*** ±***4,43***	***84,90*** ±***4,21***
BRE	93,81 ±2,31	94,03 ±2,61	***94,76*** ±***2,63***	94,10 ±2,85	***95,26*** ±***2,52***	***95,25*** ±***2,58***	***94,96*** ±***2,34***
FLA	***62,53*** ±***9,73***	***63,28*** ±***9,21***	***63,27*** ±***9,81***	61,84 ±8,29	***64,45*** ±***9,34***	***64,55*** ±***9,18***	***64,28*** ±***9,30***
GLA	***70,19*** ±***8,61***	69,08 ±9,77	***72,27*** ±***9,16***	***70,51*** ±***8,33***	68,92 ±8,35	68,68 ±8,26	***71,27*** ±***8,53***
HEA	73,14 ±7,86	***79,71*** ±***7,87***	***78,73*** ±***7,00***	***79,69*** ±***7,69***	***78,86*** ±***6,87***	***78,76*** ±***6,68***	77,51 ±6,44
IMA	95,42 ±1,38	***96,86*** ±***1,21***	96,29 ±1,17	96,30 ±1,28	95,90 ±1,07	96,00 ±1,14	95,94 ±1,17
ION	***89,34*** ±***5,17***	***89,45*** ±***4,83***	***88,63*** ±***4,60***	***88,77*** ±***4,54***	87,43 ±4,94	87,46 ±5,17	***88,57*** ±***5,00***
IRI	92,40 ±5,61	***93,47*** ±***5,36***	***93,40*** ±***5,40***	***93,47*** ±***5,44***	***93,80*** ±***5,30***	***93,80*** ±***5,30***	***94,00*** ±***5,15***
KVK	98,73 ±0,79	***99,61*** ±***0,34***	98,83 ±0,58	99,52 ±0,36	98,82 ±0,66	98,83 ±0,65	98,94 ±0,58
LBR	69,67 ±4,54	70,96 ±5,42	69,63 ±3,91	71,17 ±5,24	71,53 ±5,16	71,29 ±5,26	***74,12*** ±***6,06***
LET	84,58 ±0,84	***88,34*** ±***0,76***	87,47 ±0,78	87,61 ±0,85	86,00 ±0,68	85,98 ±0,75	86,58 ±0,71
PIM	***73,49*** ±***4,61***	***74,18*** ±***4,60***	***73,89*** ±***4,58***	***74,12*** ±***4,54***	***73,93*** ±***4,35***	***73,88*** ±***4,51***	***73,74*** ±***4,36***
SON	71,18 ±7,86	***73,88*** ±***8,70***	72,96 ±8,96	71,24 ±8,12	***75,53*** ±***9,34***	***75,49*** ±***9,08***	***75,73*** ±***9,52***
SOY	***79,11*** ±***5,67***	58,75 ±6,98	58,54 ±7,94	62,43 ±7,26	76,32 ±6,58	76,19 ±6,52	***79,92*** ±***6,82***
SPL	90,60 ±1,47	***94,71*** ±***1,26***	93,82 ±1,36	94,48 ±1,31	93,91 ±1,43	93,84 ±1,48	94,44 ±1,32
THY	99,53 ±0,23	***99,58*** ±***0,22***	99,37 ±0,28	***99,61*** ±***0,23***	99,47 ±0,27	99,45 ±0,28	99,54 ±0,24
VOT	91,02 ±8,61	93,95 ±7,13	91,25 ±9,30	92,48 ±8,89	94,99 ±4,48	94,91 ±4,55	***96,20*** ±***2,60***
VOW	***86,88*** ±***3,28***	***86,42*** ±***3,05***	84,78 ±3,35	***86,91*** ±***2,95***	85,88 ±3,20	86,03 ±3,43	**86,35** ±**3,18**
WAV	76,71 ±1,89	***77,82*** ±***2,10***	***77,95*** ±***1,72***	77,24 ±1,90	***77,80*** ±***2,00***	***77,77*** ±***1,98***	***77,90*** ±***1,83***
WIN	89,44 ±6,58	***94,25*** ±***5,63***	92,56 ±5,92	89,09 ±6,28	90,89 ±6,74	90,89 ±6,74	91,01 ±6,88

Table 2. Means and standard deviations of 10 repetitions of 10-fold CV.

	Accuracy	IG	IGR	Gini	SSV	SSV_{lex}	$SSV_{\alpha=0.5}$
Best	1	8	2	2	2	1	5
t test within 0.01	7	15	9	8	9	9	14
Wilcoxon within 0.01	7	15	8	8	9	9	13

Table 3. The best and insignificantly different (with 0,01 confidence level) result counts.

SSV_α won significantly over all the others in 2 cases (LBR and VOT), though the case of the APP dataset may be counted as the third such case, because also here, SSV_α significantly defeats all the non-SSV methods. Moreover, if we do not count SSV and SSV_{lex}, then there are another 4 datasets (BRE, SON, SOY, THY) with two winners that significantly outperform all the others. Despite the fact, that there are two datasets (AUS and PIM), for which no significant differences could be observed between any two tested methods, it is definitely justified to claim that for many datasets, different indices result in significantly different average accuracies. This conclusion confirms the need for accurate meta-learning algorithms, capable of finding the most advantageous DT induction method for given data.

Table 4 presents the summary of win-draw-loss counts between each two algorithms, according to the two statistical tests. The last rows (printed in bold face) show the relative performance of the proposed SSV modification. They prove the value of

	Accuracy	IG	IGR	Gini	SSV	SSV_{lex}	$SSV_{\alpha=0.5}$
Accuracy		1-10-10	3-11-7	1-12-8	3-7-11	4-6-11	0-9-12
IG	10-10-1		8-11-2	7-13-1	7-11-3	7-11-3	6-10-5
IGR	7-11-3	2-11-8		2-13-6	4-10-7	3-12-6	2-11-8
Gini	8-12-1	1-13-7	6-13-2		6-8-7	6-8-7	4-8-9
SSV	11-7-3	3-11-7	7-10-4	7-8-6		0-20-1	1-12-8
SSV_{lex}	11-6-4	3-11-7	6-12-3	7-8-6	1-20-0		0-13-8
$SSV_{\alpha=0.5}$	**12-9-0**	**5-10-6**	**8-11-2**	**9-8-4**	**8-12-1**	**8-13-0**	

	Accuracy	IG	IGR	Gini	SSV	SSV_{lex}	$SSV_{\alpha=0.5}$
Accuracy		1-10-10	3-10-8	1-11-9	4-6-11	4-7-10	1-8-12
IG	10-10-1		9-10-2	7-13-1	7-11-3	7-11-3	6-10-5
IGR	8-10-3	2-10-9		3-12-6	3-10-8	2-12-7	2-11-8
Gini	9-11-1	1-13-7	6-12-3		7-7-7	6-8-7	5-7-9
SSV	11-6-4	3-11-7	8-10-3	7-7-7		0-20-1	1-12-8
SSV_{lex}	10-7-4	3-11-7	7-12-2	7-8-6	1-20-0		1-12-8
$SSV_{\alpha=0.5}$	**12-8-1**	**5-10-6**	**8-11-2**	**9-7-5**	**8-12-1**	**8-12-1**	

Table 4. Pairwise win-draw-loss counts by t test (top) and Wilcoxon test (bottom).

the new method, as it has the best record of results relative to the most naive criterion based on accuracy, and shows more wins than losses in relation to IGR and Gini indices than the original SSV definitions. The only defeat is registered in relation to the IG index, but the score is 5-10-6, so it is probable, that another selection of datasets could easily invert the result. The table confirms that the proposed modification significantly improves the SSV criterion, as for many datasets the test results are significantly better and only in the case of one dataset (HEA)—significantly worse.

An interesting observation is that the IGR index does not perform well in the test. It does not mean, however, that the IGR index is useless. A guess is, that the poor results of the IGR is a consequence of using a binary tree construction algorithm, so the correction which reduces the bias towards multisplits, is not usable in this case (hampers more than helps). Probably, a similar test exploiting another tree search technique, would be much more advantageous for the IGR.

4.2 Analysis of the α parameter

The analysis described above was done for the new algorithm with $\alpha = 0.5$ chosen as the middle point between no attention payed to separated objects when they do not belong to the majority class in their data part, and full attention on them (treating them as equally important as the pairs separated and properly classified thanks to the split). To check, whether the intuition of $\alpha = 0.5$ is accurate, let's have a look at the results of a similar analysis as the one performed above, but comparing the results obtained with α values of 0, 0.1, 0.2, ..., 0.9, 1.

Table 5 presents full results table for the 21 datasets and for the 11 values of α. It is not surprising that many of the results differences are not statistically significant. The summary of obtaining the results insignificantly different from the best one, is presented

	0	0.1	0.2	0.3	0.4	0.5
APP	**86,64** ±10,00	**86,64** ±10,12	**86,64** ±10,12	**86,55** ±9,94	**86,45** ±9,85	**86,65** ±9,92
AUS	**85,14** ±4,33	**85,14** ±4,26	**84,83** ±4,05	**84,90** ±4,03	**84,87** ±3,90	**84,90** ±4,21
BRE	**94,92** ±2,38	**94,92** ±2,32	**94,94** ±2,17	**94,96** ±2,22	**94,96** ±2,41	**94,96** ±2,34
FLA	60,42 ±9,93	61,97 ±9,79	62,74 ±9,84	**63,52** ±9,86	**63,84** ±9,26	**64,28** ±9,30
GLA	**70,67** ±8,74	**70,93** ±8,90	**71,77** ±9,74	**71,97** ±8,30	**72,15** ±8,18	**71,27** ±8,53
HEA	76,85 ±6,61	77,21 ±6,76	76,72 ±6,83	77,21 ±6,84	77,51 ±6,66	77,51 ±6,44
IMA	95,80 ±1,33	95,77 ±1,12	95,99 ±1,13	**96,32** ±1,17	**96,24** ±1,00	95,94 ±1,17
ION	**88,40** ±4,55	**88,17** ±5,32	**88,17** ±5,21	**88,42** ±4,94	**88,51** ±4,80	**88,57** ±5,00
IRI	**93,87** ±5,42	**93,93** ±5,11	**93,93** ±5,11	**93,93** ±5,11	**93,93** ±5,11	**94,00** ±5,15
KVK	**98,75** ±0,72	98,71 ±0,60	98,82 ±0,59	**98,92** ±0,59	**98,90** ±0,62	**98,94** ±0,58
LBR	**74,16** ±6,01	**74,27** ±5,99	**73,95** ±5,93	**74,12** ±6,06	**74,12** ±6,06	**74,12** ±6,06
LET	85,66 ±0,78	**86,42** ±0,81	**86,41** ±0,89	86,21 ±0,87	86,20 ±0,80	**86,58** ±0,71
PIM	**74,58** ±4,53	74,10 ±4,53	**74,01** ±4,73	73,83 ±4,52	73,78 ±4,46	73,74 ±4,36
SON	**75,99** ±8,99	**75,65** ±9,16	**75,74** ±9,11	**75,79** ±9,10	**75,85** ±9,26	**75,73** ±9,52
SOY	79,14 ±5,72	**80,05** ±6,68	**80,97** ±6,59	**81,10** ±6,40	**80,87** ±6,67	**79,92** ±6,82
SPL	93,86 ±1,49	**94,36** ±1,34	**94,34** ±1,32	**94,37** ±1,37	**94,42** ±1,40	**94,44** ±1,32
THY	**99,59** ±0,24	**99,58** ±0,24	**99,58** ±0,24	**99,57** ±0,25	99,55 ±0,25	99,54 ±0,24
VOT	92,55 ±7,30	**95,81** ±2,72	**95,81** ±2,72	**96,02** ±2,65	**96,09** ±2,71	**96,20** ±2,60
VOW	**86,45** ±3,18	**86,79** ±3,25	**86,64** ±3,26	**86,56** ±3,27	**86,39** ±3,37	**86,35** ±3,18
WAV	76,96 ±2,06	77,15 ±1,91	77,19 ±1,95	**77,56** ±2,02	**77,78** ±1,91	**77,90** ±1,83
WIN	**91,45** ±6,44	**91,11** ±6,49	**91,69** ±6,95	**91,01** ±7,11	**90,90** ±6,87	**91,01** ±6,88

	0.6	0.7	0.8	0.9	1
APP	**86,93** ±9,65	**86,93** ±9,65	**87,03** ±9,53	**86,93** ±9,55	**86,52** ±9,47
AUS	**84,87** ±4,22	**84,99** ±4,15	**85,07** ±4,35	**85,09** ±4,32	**85,06** ±4,40
BRE	**94,96** ±2,33	**94,78** ±2,45	**94,82** ±2,57	**95,15** ±2,42	**95,24** ±2,54
FLA	**64,64** ±9,20	**64,65** ±9,26	**64,39** ±10,05	**64,35** ±9,49	**64,03** ±9,00
GLA	**71,45** ±8,15	**71,59** ±7,79	**70,79** ±7,97	69,24 ±8,23	69,24 ±8,23
HEA	**78,30** ±6,27	**78,82** ±6,49	**78,86** ±6,64	**78,79** ±6,41	**78,40** ±6,88
IMA	95,66 ±1,26	95,80 ±1,26	95,81 ±1,25	95,90 ±1,18	**96,01** ±1,12
ION	**88,37** ±4,96	**88,45** ±4,82	**88,60** ±4,98	**87,97** ±5,18	87,38 ±5,14
IRI	**94,00** ±5,15	**93,93** ±5,11	**93,93** ±5,11	**93,80** ±5,30	**93,80** ±5,30
KVK	**98,96** ±0,58	**98,95** ±0,62	**98,91** ±0,64	**98,89** ±0,64	98,85 ±0,64
LBR	**74,16** ±5,99	**74,37** ±5,98	**74,58** ±6,14	**74,58** ±6,23	71,40 ±5,33
LET	**86,54** ±0,74	**86,47** ±0,74	**86,49** ±0,74	**86,55** ±0,80	85,88 ±0,73
PIM	73,39 ±4,90	73,41 ±4,39	73,42 ±4,56	**74,02** ±4,69	**73,92** ±4,51
SON	**76,07** ±9,34	**76,60** ±9,66	**75,63** ±9,79	**76,52** ±9,56	**75,35** ±8,95
SOY	78,02 ±6,15	76,92 ±5,97	76,37 ±5,81	75,87 ±6,09	76,19 ±6,72
SPL	**94,46** ±1,34	94,31 ±1,37	**94,30** ±1,40	**94,24** ±1,42	93,84 ±1,50
THY	99,54 ±0,23	**99,54** ±0,23	99,49 ±0,24	99,48 ±0,25	99,45 ±0,29
VOT	**96,25** ±2,53	**96,20** ±2,54	**95,91** ±3,46	**95,59** ±3,74	94,78 ±4,66
VOW	**86,46** ±3,21	85,98 ±3,40	86,10 ±3,68	85,82 ±3,45	85,35 ±3,42
WAV	**77,77** ±1,75	**77,81** ±1,84	**77,75** ±1,87	**77,74** ±1,96	**77,83** ±1,96
WIN	**90,73** ±6,51	**90,56** ±6,83	**90,44** ±6,62	**90,84** ±6,76	**90,89** ±6,74

Table 5. Results for different values of α.

in table 6. It shows that the values of α close to 0 or 1 are worse than the values closer to the middle of the scope. The best result, according to the t test, seems to be 0.3, but the differences in the area between 0.3 and 0.8 are so small that no winner can be announced with large probability.

The pairwise win-draw-loss counts in table 7 do not show any leader. All the values within the interval [0.3, 0.8] look attractive, and obtain similar scores (with similar counts of wins and losses). The competition between the values of 0.3 and 0.4 shows no significant difference, and both values get attractive scores in relation to the others, so one could see them as the most promising ones.

A summary of win-draw-loss counts of the competitions between the selected values of α and other indices measuring split quality are presented in table 8. Again, the whole range from 0.3 to 0.8 demonstrates interesting results. The most promising result is the one obtained with $\alpha = 0.4$—it has reached a draw with the IG index (another draw can be noticed for $\alpha = 0.7$). It is very important to keep in mind, that the selection of $\alpha = 0.4$, made afterwards (after the analysis of the whole set of results) would not be fair with respect to other criteria. Also, the conclusions about superiority of one value of α over another, on the basis of the results presented in tables 7 and 8, would not be reasonable: another selection of datasets could result in another winner values, so no particular value of α can be pointed as definitely the best.

What we can claim reliably, is that the intuitions about the value of α are confirmed by the experiments: it should be neither close to 0 nor close to 1 and the value of 0.5 is a good choice.

5 Conclusions

This article shows a simple but very successful modification to the SSV criterion introducing weighted separation gains. The comparative analysis, presented here, confirms that:

- the new definition of SSV criterion is a significant improvement to the original version,
- the intuitions of $\alpha = 0.5$ are quite accurate,
- together with IG index, the SSV_α is one of the most successful in the competition between different indices,
- there exist significant differences between performances of different DT split criteria, so to provide as accurate DTs as possible we should always check several alternative indices.

There is still a need for further analysis of DT algorithm components, including split quality measures. More indices should be tested to gain meta-knowledge about when to use different indices. Such tests will be performed in the closest future. Similar analysis of other kinds of components like validation methods, data transformations etc. will certainly bring very successful meta-learning algorithms, gathering useful meta-knowledge which will eventually lead to efficient complex algorithms constructing as successful DTs as possible.

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Best	2	2	1	2	1	2	4	2	3	1	1
t test within 0.01	13	15	16	18	17	17	17	16	16	16	11
Wilcoxon within 0.01	13	14	14	17	17	16	17	16	16	13	11

Table 6. Best and insignificantly worse than the best (within 0,01 confidence level).

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0		1-17-3	0-17-4	1-13-7	2-13-6	2-14-5	2-13-6	2-13-6	3-12-6	2-13-6	5-12-4
0.1	3-17-1		0-20-1	0-18-3	0-17-4	0-18-3	1-17-3	2-15-4	3-14-4	3-14-4	6-13-2
0.2	4-17-0	1-20-0		0-19-2	0-20-1	0-19-2	1-16-4	2-16-3	2-17-2	3-16-2	6-13-2
0.3	7-13-1	3-18-0	2-19-0		0-21-0	1-19-1	2-18-1	3-16-2	3-16-2	5-14-2	8-13-0
0.4	6-13-2	4-17-0	1-20-0	0-21-0		1-19-1	2-18-1	2-17-2	2-17-2	4-16-1	9-12-0
0.5	5-14-2	3-18-0	2-19-0	1-19-1	1-19-1		2-19-0	1-19-1	1-19-1	1-20-0	8-13-0
0.6	6-13-2	3-17-1	4-16-1	1-18-2	1-18-2	0-19-2		3-18-0	1-20-0	2-19-0	8-12-1
0.7	6-13-2	4-15-2	3-16-2	2-16-3	2-17-2	1-19-1	0-18-3		1-20-0	2-19-0	9-12-0
0.8	6-12-3	4-14-3	2-17-2	2-16-3	2-17-2	1-19-1	0-20-1	0-20-1		1-19-1	6-15-0
0.9	6-13-2	4-14-3	2-16-3	2-14-5	1-16-4	0-20-1	0-19-2	0-19-2	1-19-1		4-17-0
1	4-12-5	2-13-6	2-13-6	0-13-8	0-12-9	0-13-8	1-12-8	0-12-9	0-15-6	0-17-4	

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0		1-16-4	1-15-5	1-13-7	2-13-6	2-14-5	2-13-6	2-13-6	3-12-6	3-12-6	4-12-5
0.1	4-16-1		0-20-1	0-18-3	0-16-5	0-17-4	2-15-4	3-13-5	3-14-4	3-14-4	6-13-2
0.2	5-15-1	1-20-0		0-17-4	1-17-3	1-16-4	3-13-5	3-14-4	2-17-2	4-15-2	7-12-2
0.3	7-13-1	3-18-0	4-17-0		0-21-0	2-18-1	2-17-2	3-16-2	3-16-2	5-14-2	8-13-0
0.4	6-13-2	5-16-0	3-17-1	0-21-0		2-18-1	2-18-1	2-17-2	2-17-2	5-15-1	8-13-0
0.5	5-14-2	4-17-0	4-16-1	1-18-2	1-18-2		2-19-0	1-19-1	1-19-1	2-18-1	10-11-0
0.6	6-13-2	4-15-2	5-13-3	2-17-2	1-18-2	0-19-2		3-18-0	1-20-0	4-17-0	9-11-1
0.7	6-13-2	5-13-3	4-14-3	2-16-3	2-17-2	1-19-1	0-18-3		1-20-0	2-18-1	8-13-0
0.8	6-12-3	4-14-3	2-17-2	2-16-3	2-17-2	1-19-1	0-20-1	0-20-1		2-18-1	7-13-1
0.9	6-12-3	4-14-3	2-15-4	2-14-5	1-15-5	1-18-2	0-17-4	1-18-2	1-18-2		4-17-0
1	5-12-4	2-13-6	2-12-7	0-13-8	0-13-8	0-11-10	1-11-9	0-13-8	1-13-7	0-17-4	

Table 7. Pairwise win-draw-loss counts by t test and Wilcoxon test for different values of α.

	Accuracy	IG	IGR	Gini	SSV	SSV_{lex}
0	12-8-1	4-8-9	6-10-5	6-10-5	3-13-5	3-13-5
0.1	9-11-1	5-8-8	8-9-4	7-10-4	6-12-3	6-12-3
0.2	12-8-1	5-8-8	8-10-3	7-11-3	7-12-2	6-13-2
0.3	12-8-1	6-8-7	8-12-1	7-11-3	7-13-1	5-15-1
0.4	11-10-0	6-9-6	8-12-1	9-8-4	7-14-0	6-15-0
0.5	12-9-0	5-10-6	8-11-2	9-8-4	8-12-1	8-13-0
0.6	11-10-0	5-9-7	8-10-3	8-9-4	8-13-0	8-12-1
0.7	11-8-2	5-11-5	8-10-3	7-9-5	7-14-0	7-14-0
0.8	10-9-2	5-10-6	8-10-3	7-9-5	7-14-0	6-15-0
0.9	11-7-3	4-10-7	8-9-4	7-8-6	3-18-0	3-18-0
1	11-6-4	3-10-8	5-13-3	6-9-6	1-18-2	0-19-2

Table 8. Win-draw-loss counts: α vs other split criteria.

Acknowledgements:
The author is grateful to Włodzisław Duch for the illustrative example data of figure 1, and (also to other colleagues from the Department) for fruitful discussions.
The research is supported by the Polish Ministry of Science with a grant for years 2010–2012.

References

1. Breiman, L., Friedman, J.H., Olshen, A., Stone, C.J.: Classification and regression trees. Wadsworth, Belmont, CA (1984)
2. Buntine, W., Niblett, T.: A further comparison of splitting rules for decision-tree induction. Machine Learning 8, 75–85 (1992), http://dx.doi.org/10.1007/BF00994006, 10.1007/BF00994006
3. Frank, A., Asuncion, A.: UCI machine learning repository (2010), http://archive.ics.uci.edu/ml
4. Grąbczewski, K., Duch, W.: A general purpose separability criterion for classification systems. In: Proceedings of the 4th Conference on Neural Networks and Their Applications. pp. 203–208. Zakopane, Poland (Jun 1999)
5. Grąbczewski, K., Duch, W.: The Separability of Split Value criterion. In: Proceedings of the 5th Conference on Neural Networks and Their Applications. pp. 201–208. Zakopane, Poland (Jun 2000)
6. Grąbczewski, K., Jankowski, N.: Versatile and efficient meta-learning architecture: Knowledge representation and management in computational intelligence. In: IEEE Symposium Series on Computational Intelligence (SSCI 2007). pp. 51–58. IEEE (2007)
7. Grąbczewski, K., Jankowski, N.: Efficient and friendly environment for computational intelligence. Knowledge-Based Systems p. 41 (2011), (in print)
8. L. Rokach, O.M.: Data Mining with Decision Trees: Theory and Applications. World Scientific (2008)
9. Mingers, J.: An empirical comparison of selection measures for decision-tree induction. Machine Learning 3, 319–342 (1989)
10. Quinlan, J.R.: Programs for machine learning (1993)