

## 1. Experimental studies illustrating improper use of evaluation methodology

This section aims to show the chosen flaws of commonly used experimental protocols, i.e., how their improper use can dramatically change the results of classifier comparison. In the first experiment, we will focus on the dataset shift problem in cross-validation while in the second, we show the problems with mean-ranks in Friedman test.

### 1.1. Problems with dataset shift in cross-validation

In this section, we will describe the details of the conducted experimental study to illustrate how partitioning methods in cross-validation, the commonly used metric estimation method, introduce dataset shift (here, strictly the covariate shift) and show its impact on the classifier comparison results.

We want to show that:

*Depending on the choice of the dataset folds in cross-validation, one can change the result of comparing performances of the particular two classifiers.*

Both the source code used to conduct the described experiment as well as a report with all the results are available in public Git repository<sup>1</sup>.

Eighteen popular datasets from UCI *Machine Learning Repository* (see Table ??) [?] and three classifiers (as implemented in the SCIKIT-LEARN library and with default parameters)

- GNB — Gaussian Naive Bayes,
- KNN — k-Nearest Neighbors,
- DTC — CART decision tree.

were adopted for the described experimental study.

Table 1: Overview of datasets applied in tests for Section 3.1.

DATASET	SAMPLES	FEATURES	CLASSES
<i>breastcan</i>	683	9	2
<i>wisconsin</i>	699	9	2
<i>ionosphere</i>	351	34	2
<i>sonar</i>	208	60	2
<i>balance</i>	625	4	3
<i>monkone</i>	556	6	2
<i>heart</i>	270	13	2
<i>liver</i>	345	6	2
<i>hayes</i>	160	4	3
<i>german</i>	1000	24	2
<i>iris</i>	150	4	3
<i>monktwo</i>	601	6	2
<i>wine</i>	178	13	3
<i>yeast3</i>	1484	8	2
<i>monkthree</i>	554	6	2

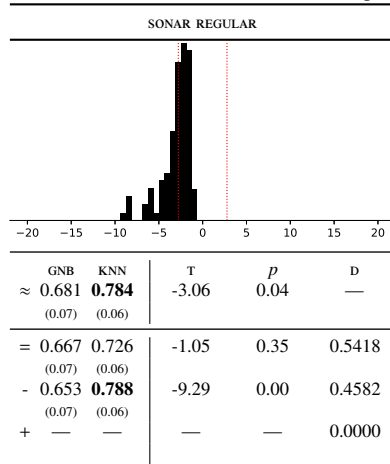
For each dataset, 100 000 partitionings into 5 folds with standard stratification were performed each one being the basis for the comparison of performance of 3 classifier pairs (constructed from the above mentioned). Thus, there were 45 different classifier comparisons (18 datasets with 3 pairs of classifiers). Classifier performance was measured by the described accuracy metric (i.e. the ratio of properly classified to all elements in a test set) estimated using standard stratified 5-fold cross-validation. To assess the statistical significance of differences between the achieved classifier performances, the *T-test* for cross validation was used (with parameter *corr* = 0.4) in T-statistic, taking the significance level  $\alpha = 0.05$ .

<sup>1</sup><https://github.com/w4k2/fair-evaluation>

Each of 54 classifier comparisons were summarized in the separate table associated with a plot presenting in a graphical form the empirical distribution (histogram) of the values of *T-statistic* in the population of 100 000 partitionings. The entries in such table are as follows (see for example Table ?? for sonar dataset depicted as the header). In the first two columns are the accuracies of the compared classifiers (average or exact - see the description of the meaning of rows below), followed by the standard deviations, the calculated value of *T-statistic* (see the description below), the corresponding *p-value* and the percentage of partitionings (in a pool of 10 000) in which the situation described in the line occurred (D). The meaning of the values in the following rows in each table are as follows (the presented in column with heading *p-value* corresponds to the listed value of T-statistic in a given row):

- $\approx$  : the average values of the classifier accuracies and T-statistic in 10 000 repetitions,
- $=$  : the values of the classifier accuracies with the T-statistic value nearest to zero, contained in the critical region, symbolizing the lack of significant differences between the compared classifiers,
- $-$  : the values of the classifier accuracies with the lowest value of T-statistic outside the critical region, symbolizing a statistically significant advantage of the classifier from the right column,
- $+$  : the values of the classifier accuracies with the highest value of T-statistic outside the critical region, symbolizing a statistically significant advantage of the classifier from the left column.

Table 2: Example summary of 100 000 repetitions of 5-fold cross-validation used to compare two classifiers – see the description in the text.



Therefore, the interpretation of the example table will be as follows. The average value of the observation indicates no predominance of any considered algorithm with the accuracies of approximately 78% for KNN and 68% for GNB. The average T-statistics value is -3.06, which gives *p value* at 0.04, which makes it a statistically significant difference.

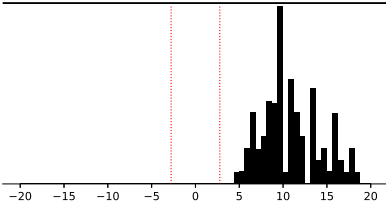
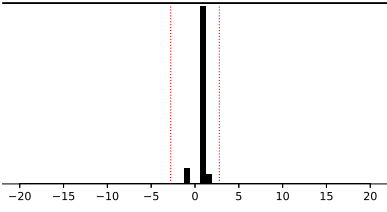
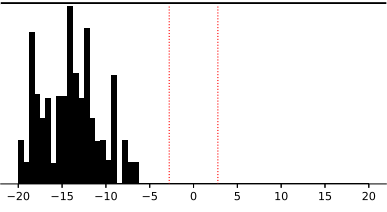
However, we may also observe situations in which a significant difference between algorithms does not occur (line  $=$ ), where T-statistic equal -1.05, which gives *p value* at 0.36. Similar situations (no significant difference) occurs in 54% of the considered cases. Divisions which show a significant advantage of the KNN algorithm occur in 46% of cases, and the extreme value of T-statistics is -9.26, which is an outlier in the context of the problem under consideration.

In connection with the above observations, we can properly validate two contradictory research hypotheses by pulling the dataset into folds adequately many times. Moreover, with the average value of T-statistics close to the significance level, approximately a quarter of the experiments will give us information about the KNN advantage, when all the other cases of the divisions will deny any statistical difference.

The simplest of situations encountered in performed experiments is that in which each of ten thousand experiments leads to the same conclusion, as is shown in the examples in the Table ???. The distribution of T-statistics in such cases is either narrow enough to fit within the critical region (SOYBEAN) or far enough from it, so that even outlier

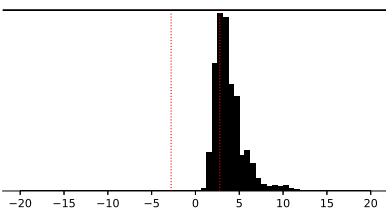
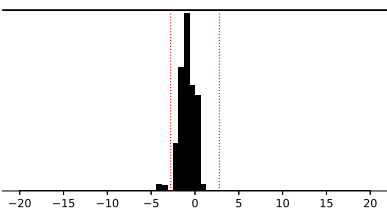
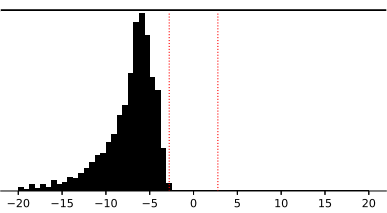
observations leading to other conclusions do not happen. It is worth noting, however, that even in such cases, outliers reveal a major deviation from the mean value (for example, an average of 11.74 in the WINE dataset and an outlier observation of 33.48). It is particularly important to note at this point that only twelve of the fifty-four comparisons are characterized by such unambiguous conclusions.

Table 3: Three example summaries of ten thousand repetitions of 5-fold cross-validation used to compare two classifiers using t test over accuracy score showing **results evident for one hypothesis**.

WINE REGULAR						SOYBEAN REGULAR						MONKONE REGULAR					
																	
GNB	KNN	T	p	D		GNB	CART	T	p	D		GNB	KNN	T	p	D	
≈ <b>0.974</b>	0.690	11.74	0.00	—		≈ 0.998	0.989	0.87	0.43	—		≈ 0.664	<b>0.945</b>	-18.24	0.00	—	
(0.03)	(0.06)					(0.01)	(0.04)					(0.04)	(0.02)				
-	—	—	—	0.0000		= 0.978	0.978	0.00	1.00	0.4388		-	—	—	—	0.0000	
(0.03)	(0.06)					(0.01)	(0.04)					(0.04)	(0.02)				
-	—	—	—	0.0000		-	—	—	—	0.0000		-	0.664	<b>0.957</b>	-47.89	0.00	1.0000
(0.03)	(0.06)					(0.03)	(0.03)					(0.04)	(0.02)				
+	<b>0.977</b>	0.702	33.48	0.00	1.0000	+	—	—	—	—	0.0000	+	—	—	—	—	0.0000
(0.03)	(0.06)																

The dominant majority, as many as 34 out of 54 comparisons correspond to the situation presented in the example in Table 1, where we can draw two contradictory conclusions from the appropriate (let's emphasize - random) combination of patterns division into folds. Examples here are presented in Table ???. In the case of the WINE dataset, we may observe a situation in which the GNB algorithm in averaging achieves an advantage over the CART, but in a quarter of the cases, random division of the dataset will lead to the conclusion that there are no significant differences. A much more interesting case is the IRIS dataset, where in 94% of divisions there is no significant difference between the compared classifiers, but 2% of the problem instances leads to the conclusion that the KNN has a significant advantage over GNB. An even stronger example of this type is another case of the WINE dataset, where, despite the predominance of the DTC over the KNN and the average difference in quality at 21%, we can still find one part-per-thousand of situations in which the statistical difference between them disappears.

Table 4: Three example summaries of ten thousand repetitions of 5-fold cross-validation used to compare two classifiers using t test over accuracy score showing results pointing **two different hypotheses**.

WINE REGULAR						IRIS REGULAR						WINE REGULAR					
																	
GNB	CART	T	p	D		GNB	KNN	T	p	D		KNN	CART	T	p	D	
≈ <b>0.974</b>	0.905	3.91	0.02	—		≈ 0.954	0.965	-1.00	0.37	—		≈ 0.690	<b>0.905</b>	-7.48	0.00	—	
(0.03)	(0.05)					(0.03)	(0.03)					(0.06)	(0.05)				
= 0.967	0.961	0.32	0.76	0.2751		= 0.960	0.960	0.00	1.00	0.9389		= 0.724	0.847	-2.30	0.08	0.0011	
(0.03)	(0.05)					(0.03)	(0.03)					(0.06)	(0.05)				
-	—	—	—	0.0000		- 0.947	<b>0.973</b>	-4.00	0.02	0.0204		- 0.702	<b>0.921</b>	-57.30	0.00	0.9989	
(0.03)	(0.03)					(0.03)	(0.03)					(0.06)	(0.05)				
+	<b>0.972</b>	0.944	69.20	0.00	0.7249	+	—	—	—	0.0000		+	—	—	—	0.0000	
(0.03)	(0.05)																

The most interesting, however, is the third, last group of observations, which consists of 8 out of 54 examples (Table ??). There is a set of cases in which, depending on which of the random dataset divisions we select for the experiment, we may get the validation of each possible conclusion. The liver dataset is particularly interesting here, where only 3 out of 100 000 cases show the CART algorithm superiority over KNN, with approximately 94% of their equal quality and 5% of KNN advantage cases.

This may lead to the hypothesis that if we have the sufficiently large computational power to repeat random dataset divisions, we will lead to a situation in which, using the standard approach to experiments, we will be able to reasonably support any hypothesis about the statistical relationship between the compared classifiers. Moreover, based on all separate groups of cases, we can undoubtedly conclude that the existence of such combinations of data sets and classification algorithms is common, against which we can support contradictory hypotheses, depending on the applied random division of the data set into folds.

Table 5: Three example summaries of ten thousand repetitions of 5-fold cross-validation used to compare two classifiers using t test over accuracy score showing results pointing **all the possible hypotheses**.

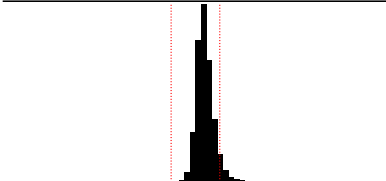
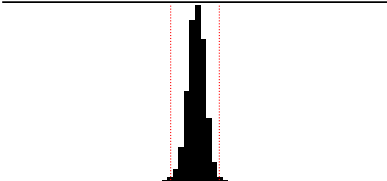
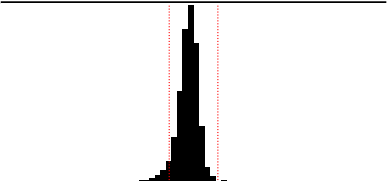
LIVER REGULAR						GERMAN REGULAR						DIABETES REGULAR					
																	
KNN ≈ <b>0.663</b> (0.05)	CART 0.634 (0.05)	T 2.6*10 <sup>10</sup>	p 0.00	D —		KNN ≈ 0.691 (0.02)	CART 0.688 (0.03)	T 0.15	p 0.89	D —		KNN ≈ 0.691 (0.03)	CART 0.701 (0.04)	T -0.67	p 0.54	D —	
= 0.643 (0.05)	0.643 (0.05)	0.00	1.00	0.9413		= 0.688 (0.02)	0.688 (0.03)	0.00	1.00	0.9807		= 0.682 (0.03)	0.682 (0.04)	0.00	1.00	0.9286	
- 0.638 (0.05)	<b>0.655</b> (0.05)	-6.00	0.00	0.0003		- 0.693 (0.02)	<b>0.709</b> (0.03)	-16.00	0.00	0.0096		- 0.681 (0.03)	<b>0.716</b> (0.04)	-21.90	0.00	0.0610	
+ <b>0.699</b> (0.05)	0.641 (0.05)	2.6*10 <sup>15</sup>	0.00	0.0584		+ <b>0.712</b> (0.02)	0.682 (0.03)	18.97	0.00	0.0097		+ <b>0.714</b> (0.03)	0.677 (0.04)	11.06	0.00	0.0104	

Table 6: ABCD

TEST USED	NUMBER OF SUPPORTED HYPOTHESES		
	I	II	III
<i>without correction</i>	12	34	8
<i>parametric (corr=.1)</i>	11	35	8
<i>parametric (corr=.2)</i>	11	35	8
<i>parametric (corr=.3)</i>	9	39	6
<i>parametric (corr=.4)</i>	9	39	6
<i>parametric (corr=.5)</i>	9	39	6
<i>parametric (corr=.6)</i>	13	36	5
<i>non-parametric</i>	23	31	0