Dataset Weighting via Intrinsic Data Characteristics for Pairwise Statistical Comparisons in Classification

José A. Sáez[1](✉), Pablo Villacorta[2], and Emilio Corchado[1]

[1] Department of Computer Science and Automatics, University of Salamanca,
Plaza de los Caídos s/n, 37008 Salamanca, Spain
{joseasaezm,escorchado}@usal.es

[2] Department of Computer Science and Artificial Intelligence, CITIC-UGR,
University of Granada, 18071 Granada, Spain
pjvi@decsai.ugr.es

Abstract. In supervised learning, some data characteristics (e.g. presence of errors, overlapping degree, etc.) may negatively influence classifier performance. Many methods are designed to overcome the undesirable effects of the aforementioned issues. When comparing one of those techniques with existing ones, a proper selection of datasets must be made, based on how well each dataset reflects the characteristic being specifically addressed by the proposed algorithm. In this setting, statistical tests are necessary to check the significance of the differences found in the comparison of different methods. Wilcoxon's signed-ranks test is one of the most well-known statistical tests for pairwise comparisons between classifiers. However, it gives the same importance to every dataset, disregarding how representative each of them is in relation to the concrete issue addressed by the methods compared. This research proposes a hybrid approach which combines techniques of measurement for data characterization with statistical tests for decision making in data mining. Thus, each dataset is weighted according to its representativeness of the property of interest before using Wilcoxon's test. Our proposal has been successfully compared with the standard Wilcoxon's test in two scenarios related to the noisy data problem. As a result, this approach stands out properties of the algorithms easier, which may otherwise remain hidden if data characteristics are not considered in the comparison.

1 Introduction

Classification tasks aim to create a model, called a classifier, from labeled examples of the problem. The classifier is then used to predict the class label of new examples from the value of their attributes. Thus, the characteristics of the data used to build the classifier directly influence it, affecting its complexity and classification performance. Moreover, the presence of some undesirable properties

H. Pérez García et al. (Eds.): HAIS 2019, LNAI 11734, pp. 61–72, 2019.
https://doi.org/10.1007/978-3-030-29859-3_6

in the data, which are usually quantified using well-known metrics, may negatively affect classifier learning, causing problems broadly studied in the literature [6,9,12]. For instance, data that are characterized by an unequal distribution of the examples among the classes (which is usually measured using the *imbalance ratio*) results in the *imbalanced classification* problem [6]. Data that are characterized by a large quantity of unknown attribute values (which is usually measured as the percentage of these values in the data) results in the *missing values* problem in classification [9].

In order to build a classifier from data heavily characterized by these undesirable characteristics, many techniques have been proposed in the literature. In these works, new algorithms are typically compared against existing ones considering a set of datasets, which should be representative of the issues being addressed [14] (degree of imbalance between classes, presence of missing values, errors, etc.). In such a way, it is possible to analyze the effects of these properties on the classifiers built by all these methods. However, the selection of the datasets must be done carefully, since some of them may be much more representative of the problematic characteristic than others.

In the last decade, the analysis of results has been closely related to the usage of statistical tests [3,15], which are needed to confirm whether a new method provides a significant improvement with respect to existing ones. Among them, *Wilcoxon's signed-ranks* test [3] (hereafter called *Wilcoxon*'s test) is commonly used in the machine learning literature to perform pairwise comparisons. It is a non-parametric statistical test that, in the context of classification algorithms, compares the performance of two methods when they are applied to a set of problems to find differences between them. In this test, each problem (i.e. dataset) has the same relevance. However, the relevance of each dataset when determining the output of the test should be different, according to the degree to which a dataset reflects the characteristic of interest. Datasets in which the presence of the characteristic being addressed is stronger should have a stronger influence on the result of the test, as they are more representative of such characteristic.

This study proposes an hybrid system which weights each one of the datasets used in the comparison performed by *Wilcoxon*'s test according to its representativeness with respect to the characteristic under study. To the best of our knowledge, this approach has not been previously used in the context of pairwise statistical comparisons between classifiers. Weighting schemes do have been applied to statistical comparisons between multiple classifiers (not pairwise). *Quade*'s test [11] considers that some problems are more difficult than others, and proposes scaling each problem depending on the differences observed in the algorithms' performances. However, there are two main differences between *Quade*'s test and our proposal. First, our weighting scheme is for pairwise comparisons, whereas *Quade*'s test is aimed at detecting differences between more than two classifiers and requires the usage of post-hoc procedures to characterize these differences. Second, *Quade*'s test only allows a weighting scheme based on the performance of the classifiers, whereas our proposal can be used with the performance of the methods, but also with any other metric computable from the data, as explained in Sect. 3.

The novel approach proposed implies a modification of *Wilcoxon*'s test in order to include information about the relevance of each dataset in the comparison, giving more importance to those problems that are interesting from the point of view of the metric under study. When comparing supervised classification algorithms, experiments are usually conducted to show that one of the algorithms is better than some others. However, it is well-known that no algorithm can be better than any other for any condition (*no free lunch theorem* [17]). Thus, instead of this kind of comparisons trying to find the best overall algorithm, experiments should be conducted to show which characteristics of the problem (such as the class imbalance, missing values, noisy data, etc.) lead to a better or worse performance of a classification algorithm. The weighting scheme enables us to easily focus on the exploration of such conditions, which is recommended when proposing new methods [14]. For this reason, the modification of *Wilcoxon*'s test presented in this study is an interesting analysis tool that allows to include additional information about the characteristics of each classification problem in order to deal with the comparison of two classification algorithms and better identify the conditions that are most favorable for each of them.

To assess the feasibility of our proposal, *Wilcoxon*'s test and its weighted version will be experimentally compared in different scenarios in the framework of the *noisy data* in classification [12]–note that the proposal is not exclusively applicable to this task, but can be used in any comparison of classification methods. In them, different metrics that can be computed from the data will be studied and the performance of several classifiers will be compared using the unweighted and weighted *Wilcoxon*'s test. The output of both tests will be compared with findings already published in the literature about noisy data.

The rest of this research is organized as follows. Section 2 provides an overview of metrics to characterize datasets and the description of *Wilcoxon*'s test. Section 3 introduces the hybrid version of *Wilcoxon*'s test for classifier comparison, including data weighting. Then, Sect. 4 includes the comparison of the unweighted and weighted versions of *Wilcoxon*'s test. Finally, Sect. 5 points out some concluding remarks.

2 Background

2.1 Measuring Data Characteristics

Each dataset has particular characteristics that define it, such as its size, the generality of the data and the inter-relationships among the variables. These properties are generally quantified defining numerical metrics directly computable from the data, with the aim of increasing the knowledge about the problem and determining the best way to deal with the data.

Classification datasets are composed by examples that are described by several attributes (numerical or nominal) and a class label (always nominal). This formation enables one to compute any numerical measure that summarizes the full data or a concrete part of these, such as relative or absolute frequencies of concrete values, means and medians, correlations among variables and so on.

Among the simplest metrics that can be computed from a dataset are the number of examples, attributes and classes. These are traditionally used to quantify, in a simple way, data characteristics such as the size and complexity of a dataset. Thus, higher values of these metrics usually represent a higher size and complexity of the corresponding dataset, resulting in some of the most studied problems within the framework of classification tasks - a clear parallelism among the characteristics of the data, the metrics employed to quantify them and the type of problems studied can be established. For example, some works focus on datasets with a high number of attributes, a fact that results in the *curse of dimensionality* problem [1]. The presence of unknown attribute values (which can be quantified by the ratio of missing values) for some examples results in *missing values* problems [9]. *Big data* problems [2] are characterized, among other factors, by the presence of a high number of attributes and examples simultaneously. Datasets with a highly unequal number of examples in each class (quantified using the imbalance ratio) lead to *imbalanced data* problems [6].

Another recent trend proposes more sophisticated *data complexity measures* [5] to quantify more subtle characteristics of the data which are considered difficult in classification tasks, such as the overlapping among classes, their separability or the linearity of the decision boundaries. Among the metrics estimating the overlapping existing in a dataset, one can find the `F1` metric, which computes the maximum *Fisher's discriminant ratio*; the `F2` metric, which estimates the volume of the overlapping region; and the `F3` metric, representing the maximum feature efficiency, which is the maximum fraction of examples within the overlapping region distinguishable with only one attribute. Class separability measures include the `N2` metric, which is used to estimate whether the examples of the same class lie close in the feature space, and the `N3` metric, which denotes how close the examples of different classes are. Other metrics, such as `L3` or `N4`, are used to estimate the linearity of the decision boundaries.

Finally, other artificial mechanisms have been proposed in the literature to build synthetic datasets from the modification of real-world ones, controlling the presence of a particular characteristic of the data [9,12]. These schemes enable one to extract conclusions based on the properties of the data which are modified. Examples of these mechanisms are, for example, the introduction of errors into the data (resulting in *noisy data* problems [12]) or the introduction of missing values [9]. In all these cases, the metric to study usually represents the amount (commonly, a percentage) of the property modified in the data. In this research, we consider both the computation of data complexity metrics and the use of artificial mechanisms to modify the datasets in two different experiments related to the problem of noisy data in classification.

2.2 Wilcoxon's Signed-Ranks Test

Wilcoxon's signed-ranks test [3] is a simple, yet safe and robust, nonparametric procedure that aims to detect whether two related samples come from two different populations. When applied to classification algorithms comparison, it performs an statistical comparison between the performance of two techniques

X and Y when they are applied to a common set of n problems. Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be the performance results of X and Y in the n datasets considered in the comparison. *Wilcoxon*'s test proceeds as follows:

1. Compute the difference between the performance results of X and Y for each one of the n datasets, that is, $d_i = x_i - y_i, i = 1, \ldots, n$.
2. Rank the differences from the lowest value of $|d_i|$ (with $rank(d_i) = 1$) up to the highest one (with $rank(d_i) = n$), by increments of 1. If there are t tied differences $|d_i|, \ldots, |d_{i+t}|$, they are sorted in any order, but the final rank of all of them is computed as the average rank $(rank(d_i) + \ldots + rank(d_{i+t}))/t$.
3. Let R^+ be the sum of ranks for the datasets in which X outperforms Y $(d_i > 0)$, and R^- the sum of ranks for the opposite $(d_i < 0)$. Those $rank(d_i) \mid d_i = 0$ are evenly splitted between R^+ and R^- and, if there is an odd number of them, one is ignored –note that $R^+ + R^- = n \cdot (n+1)/2$:

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2} \sum_{d_i=0} rank(d_i) \qquad R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2} \sum_{d_i=0} rank(d_i)$$

4. Let $W = min(R^+, R^-)$ be the *Wilcoxon* statistic. If this value is less than or equal to the theoretical value of the specific distribution this statistic is known to follow (Table B.12 in [18]), the null hypothesis of mean equality is rejected. This fact implies that a given method outperforms the other one, with the corresponding p-value associated.

Note that *Wilcoxon*'s test is analogous to the paired t-test, but it does not require any parametric assumption. It is more sensitive; thus, higher differences $|d_i|$ have a greater influence than in the t-test when determining the final result of the comparison, which is probably desired, although the absolute magnitudes of these differences are ignored. This means that two differences d_i, d_j can be equal, but we lose the information about the magnitudes of the performance values of the classifiers x_i, y_i, x_j, y_j they come from. This fact is one of the reasons that motivates the proposal of a weighting scheme for the data considered.

3 Hybridizing Wilcoxon's Signed-Rank Test with Dataset Weighting Based on Data Characteristics

When applying *Wilcoxon*'s test to compare two methods over a set of problems, only the performance results of each algorithm in each dataset are considered. The intrinsic properties of these datasets are not usually taken into account in the statistical comparison, and each one of the problems receives the same importance for determining the result of the test. As mentioned previously, sometimes we are interested in comparing methods which explicitly address a specific problematic characteristic of the data, such as noise, class imbalance, missing values, etc. However, the datasets considered may present this characteristic to different degrees, which can be quantified by an appropriate metric. Consequently, each dataset should have a different influence on the test result.

Our novel hybrid proposal in this scenario is to weight each one of the datasets of the comparison depending on its relevance with respect to the characteristic being studied when applying *Wilcoxon*'s test: the higher the value of the metric in the dataset, the higher the weight, that is, the importance of the dataset when determining the test result. Therefore, the test conclusions are based on the representativeness of each dataset with respect to the metric used.

The weighting proposed for *Wilcoxon*'s test is based on the following steps:

1. **Computation of the metric to study on each dataset.** A metric of interest is computed over each one of the n datasets considered in the comparison of the methods X and Y, resulting in the values m_i, $i = 1, \ldots, n$.
2. **Computation of the weight associated to each dataset.** The values m_i are normalized to obtain a weight w_i for each dataset as follows:

$$w_i = m_i / \sum_{j=1}^{n} m_j \tag{1}$$

   Note that the aforementioned equation is used to maximize the weight of those datasets with higher values m_i. In case we want to maximize the weight of the datasets having lower values, the equation of the weights is $w_i = (M - m_i)/\sum_{j=1}^{n}(M - m_j)$, being $M = \max_i\{m_i\}$.
   Due to the large variety of metrics that can be computed from the data, this paper proposes the calculation of weights linearly distributed across the domain of the metric (as shown by Eq. 1). However, other schemes are also possible, such as logarithmic or exponential. The choice of a concrete scheme should be made by the data analyst based on the problems involved in the comparison and the distribution of values m_i. In any case, it is important to remark that we do not manipulate the weight of each dataset arbitrarily, since the weights are determined by the values m_i of the metric.
3. **Modification of the performance values of the methods to compare.** The performance results x_i and y_i of the two algorithms X and Y involved in the comparison are replaced by their weighted versions x_i^w and y_i^w as follows:

$$x_i^w = x_i \cdot w_i \qquad\qquad y_i^w = y_i \cdot w_i$$

4. **Application of *Wilcoxon*'s test over the weighted results.** Finally, *Wilcoxon*'s test is applied to the weighted data x_i^w and y_i^w, $i = 1, \ldots, n$.

Note this weighting scheme is not a new statistical test; it requires a modification of the original data samples based on the relevance of each dataset with respect to the characteristic of interest. It neither intends to be a substitute of *Wilcoxon*'s test, but a complementary tool to be used when analyzing the results obtained in experiments. Even though a classifier should show differences against another one when using unweighted *Wilcoxon*'s test, the use of its weighted version enhances this issue. Unlike the unweighted *Wilcoxon*'s test, the weighting can capture the relevance of those data more representative with respect to the property of interest. Our novel proposal modifies (considering the property of the data) the differences in performance on the two methods, which determine the rankings assigned to each method and, thus, the final test result.

Although in some works the only goal of new proposals is to improve the classification performance of existing methods (without focusing on particular data characteristics), the use of the weighting scheme for statistical comparisons could be also interesting. Several metrics could be computed from the data and applied to the statistical comparisons to obtain another batch of results. With them, we could check whether giving more importance to some property of the data makes a particular algorithm stand out. Thus, the weighting approach can be useful to uncover properties of the algorithms, that is, their better behavior in datasets presenting a particular characteristic to a high degree, even when they were not specifically designed for dealing with this type of data.

Finally, even if we are not studying any specific characteristic of the data, a relevance metric based on the accuracy itself can still be defined. The weighting procedure can be applied to assign more importance to those datasets that are intrinsically more difficult for constructing a model (and therefore every classifier performs poorly on them in terms of accuracy), since those are more interesting from a classification point of view. This would partially remedy the fact that *Wilcoxon's signed-ranks* test does not consider absolute magnitudes of the samples to reach the final result, which in our scenarios may often represent an omission of valuable information. In these cases, datasets could be weighted, for example, depending on the intrinsic difficulty of the problem to be learnt by any classifier, which can be quantified, for instance, as the minimum theoretical error achievable in the dataset by the *Bayes* classifier. Other option could be the minimum error obtained experimentally by any known classifier in that dataset.

4 Combining Data Weighting and Wilcoxon's Test in the Framework of Noisy Data in Classification

In this section we assess the differences of applying *Wilcoxon*'s test with or without the weighting scheme proposed in Sect. 3. Two scenarios are studied focusing on a common problem in classification, namely, the existence of errors or *noise* in the data [12]. These scenarios are not aimed at showing that the weighted version allows to reach statistically significant results, neither if it is better or worse than the unweighted version, but showing that giving more importance to more representative datasets for the comparison of algorithms leads to different conclusions that are closer to those claimed in the literature.

4.1 Robustness of Different Classifiers to Noise

This scenario considers the results of two classification algorithms, `C4.5` [8] and `SVM` [16], over a collection of 30 datasets taken from the UCI repository [4], which are affected by different amounts of noise (added artificially).

These algorithms are known to behave differently dealing with noisy data [10]. On the one hand, C4.5 is a well-known robust learner, which is less affected by the presence of noise thanks to the pruning mechanisms which reduce the chances that the trees overfit the data [8]. On the other hand, SVM usually obtains better results when data have little or no noise at all, but performs poorly in presence of severe noise [10].

Ideally, *Wilcoxon*'s test without weights should be able to find statistically significant differences in this setting, as already noted in existing literature. Furthermore, since we are focusing on noisy data, the amount of noise present in each dataset should be considered when drawing a conclusion.

The performance of the methods is calculated as the average of the accuracy in test sets of five independent runs of 5-fold stratified cross validation (SCV).

To control the presence of noise, different noise levels $l\%$ are introduced into each training set. To this end, the *uniform class noise scheme* [12] is used, in which $l\%$ of the examples are corrupted randomly changing their class labels by other randomly chosen one. For each one of the 30 datasets used we introduce a different random noise level $l \in [5, 50]\%$, in steps of 5%. The procedure to introduce a noise level $l\%$ in the original dataset is the following:

1. A noise level $l\%$ is introduced into a copy of the full original dataset.
2. The original dataset and the noisy copy are partitioned into 5 folds, maintaining the same examples in each one.
3. The training sets are built from the noisy copy, whereas the test sets are built from the original dataset.

The algorithms have been executed with the following configurations:

- C4.5: *confidence* $= 0.25$, *instances per leaf* $= 2$, *prune* after the tree building.
- SVM: $C = 100$, *tolerance* $= 0.001$, $\epsilon = 10^{-12}$, *kernel* = PUK ($\sigma = 1$, $\omega = 1$).

Table 1 shows the datasets considered in the experiments, their noise levels and the performance of C4.5 and SVM. The usage of unweighted *Wilcoxon*'s test over the results shown above, comparing C4.5 versus SVM in the 30 datasets, provides a total sum of ranks $R^+ = 289$ in favor of C4.5 and a sum of ranks $R^- = 176$ for SVM. These results show that C4.5 presents, as expected, a better behavior than SVM dealing with this type of data (since it obtains a sum of ranks higher than that of SVM). However, this difference is not statistically significant to a significance level of 0.05 since the p-value obtained is 0.241.

In order to include information about the noise level of each dataset, the weighting scheme of Eq. 1 is used. In this case, we want to give more importance to those datasets with higher amount of noise when determining the outcome of the test. The weight w_i in this problem is computed by dividing each noise level l_i in Table 1 by the sum of noise levels, $S = 10 + 25 + \ldots + 15 + 15 = 635$. The weighted version of *Wilcoxon*'s test provides $R^+ = 333$ for C4.5 and $R^- = 132$ for SVM, with an associated p-value $= 0.038420$.

As these results show (Table 2), C4.5 obtains a higher sum of ranks than SVM when using the weighting scheme (like in the first comparison without using weights), but in this case the p-value obtained is much lower and now determines statistical significance, which is consistent with the results published in the literature [10]. Thus, by weighting each dataset based on the noise level, we were able to uncover a particular property of C4.5 versus SVM, that is, a better behavior when dealing with datasets affected by severe noise.

Table 1. Datasets, noise levels and accuracy of C4.5 and SVM.

Dataset	Noise	C4.5	SVM	Dataset	Noise	C4.5	SVM	Dataset	Noise	C4.5	SVM
autos	10	73.56	64.5	heart	40	70.81	69.33	satimage	5	85.4	89.25
banana	25	88.03	90.04	ionosp.	40	80	75.05	segment	20	93.09	87.77
cleveland	20	49.96	36.82	iris	25	90.27	86.8	sonar	20	68.76	83.46
contracep.	10	50.52	46.8	led7digit	35	68.72	66.88	twonorm	5	84.61	96.96
dermat.	5	93.07	96.53	lymph.	10	77.31	80.96	vehicle	45	52.84	53.12
ecoli	15	77.73	63.28	magic	45	83.36	83.73	vowel	10	74.91	87.11
flare	30	73.81	70.71	pageblocks	30	95.59	95.77	wdbc	25	88.61	85.77
german	30	69.68	64.46	penbased	20	93.28	84.57	wine	5	89.55	94.93
glass	5	67.28	67.75	phoneme	10	84.76	86.66	yeast	15	51.97	54.3
hayes	15	81.24	73.15	pima	50	70.94	60.21	zoo	15	92.26	72.85

Table 2. *Wilcoxon*'s test with and without weights using the results of Table 1.

Wilcoxon	C4.5 (R^+)	SVM (R^-)	p-value
Unweighted	289	176	0.241038
Weighted	333	132	0.038420

4.2 Noise Filtering Efficacy

This scenario focuses on using *noise filters* in classification problems [12]. These are preprocessing methods to identify and remove noisy data before building a classifier. The removal of noisy examples has shown to be beneficial in many cases, improving the performance of the classifiers used later [12]. However, examples containing valuable information may also be removed, which implies that filters do not always provide an improvement in performance [13].

The work of Sáez et al. [13] show that the efficacy of noise filters, i.e., whether their usage causes an improvement in classifier performance, is somehow related to the characteristics of the data. The authors show that the overlapping among the classes, measured with the F2 metric, is important to determine whether filters will improve classifier performance. Thus, when the amount of overlapping is high enough, filters usually improve classifier performance. The F2 metric computes the volume of the overlapping region among the examples of two different classes C_1 and C_2, by means of the following equation:

$$F2 = \prod_{i=1}^{d} \frac{minmax_i - maxmin_i}{max(f_i, C_1 \cup C_2) - min(f_i, C_1 \cup C_2)}, \tag{2}$$

being d the number of attributes, $max(f_i, C_j)$ and $min(f_i, C_j)$ the maximum and minimum values of the feature f_i in the set of examples of class C_j, $minmax_i$ the minimum of $max(f_i, C_j)$ and $maxmin_i$ the maximum of $min(f_i, C_j)$.

In this case, we compare a well-known noise filter, *Edited Nearest Neighbor* (ENN) [7], versus not considering any preprocessing of the data. As in [13], the classifier used in both cases is the *Nearest Neighbor* (NN) rule [8], known to be sensitive to noise. ENN is run considering $k = 3$, and both ENN and NN consider the HVDM distance, which is valid for nominal and numerical attributes.

These two methods are compared over 20 datasets taken from the UCI repository [4]. The performance is measured using AUC, which is an evaluation metric less sensitive to class imbalance, an issue posed by some of the data used. The AUC shown in Table 3 is the average of 5 independent runs of a 5-fold SCV, when no preprocessing is done before NN (None) and when ENN is used prior to NN.

Table 3. Datasets, F2 metric and performance of ENN and None.

Dataset	F2	None	ENN	Dataset	F2	None	ENN
appendicitis	4.50E-02	0.7551	0.7511	monk-2	6.67E-01	0.7531	0.7601
australian	3.00E-03	0.8236	0.8120	phoneme	2.71E-01	0.8683	0.8440
banana	6.26E-01	0.8710	0.8917	pima	2.52E-01	0.6487	0.6601
breast	1.88E-01	0.5570	0.5978	sonar	1.00E-06	0.8614	0.7964
bupa	7.30E-02	0.6218	0.6116	spambase	2.53E-33	0.8965	0.8749
crx	3.00E-03	0.8233	0.8172	spectfheart	3.60E-19	0.6299	0.6203
haberman	7.18E-01	0.5519	0.5658	tic-tac-toe	1.00E+00	0.9088	0.8970
heart	1.96E-01	0.7663	0.8043	twonorm	4.12E-03	0.9424	0.9518
housevotes	1.00E+00	0.9484	0.9550	wdbc	5.90E-11	0.9507	0.9469
mammograp.	7.44E-01	0.7494	0.7968	wisconsin	2.17E-01	0.9547	0.9691

The results of the tests are summarized in Table 4. Regarding the unweighted version, the test slightly favors ENN as it has a larger $R^- = 117$ than None, $R^+ = 93$, but no statistically significant differences are found (p-value = 0.640744). To sum up, no interesting conclusions can be drawn about the efficacy of ENN when the only information is the AUC.

When the data are weighted by their F2 metric, the output of the test is the same, but it is closer to statistical significance. Thanks to such additional information, ENN is now clearly favored, with $R^- = 141$ versus a much smaller value $R^+ = 69$ obtained when no preprocessing is done. The p-value corresponding to this comparison is 0.189340, closer to the significance threshold.

This fact shows that, giving more importance to datasets with higher degrees of overlapping, the preprocessing can work better than not considering it, which is in concordance with the results claimed in [13]. This conclusion can be drawn from the great decrease in the p-value when the weighted version of *Wilcoxon*'s

Table 4. *Wilcoxon*'s test with and without weights using the results of Table 3.

Wilcoxon	None (R^+)	ENN (R^-)	p-value
Unweighted	93	117	0.640744
Weighted	69	141	0.189340

test is employed instead of the conventional one, even though the new p-value is still larger than the significance threshold. It is important to note that the conclusions claimed in [13] on the data properties that determine in which cases the filtering is statistically beneficial are based on a combination of several data complexity metrics considered simultaneously, among which F2 is included. This fact can explain why the weighting scheme does not show significant differences in the comparison, since we only consider the isolated metric F2 in our study.

5 Concluding Remarks

This research proposes a hybrid approach to weight data before using *Wilcoxon*'s test and give more or less importance to the different data in a comparison. The weights of the datasets are computed using characteristics of the datasets used in the comparison. The conclusions reached by the statistical test consider the property that determines the weighting, which constitutes additional information not exploited by the unweighted version of *Wilcoxon*'s test.

We have evaluated our proposal in two scenarios related to the problem of noisy data. In the first scenario, we have compared the C4.5 robust learner and the noise-sensitive SVM classifier when they are trained over data with different noise levels. The results revealed that the weighting scheme based on the noise ratio of each dataset leads to statistically significant differences that the unweighted *Wilcoxon*'s test could not find. Such differences support the claims done in the existing literature about the superiority of C4.5 over SVM on noisy data, particularly when the amount of noise is high enough [10].

In the second scenario, we have compared the efficacy of the ENN filter versus not-preprocessing. In the literature, it is claimed that noise filters are usually useful when the overlapping among the classes is noticeable. Neither of the tests were able to detect such differences, but the weighted version showed a clear advantage towards the use of ENN, supported by a large decrease of the p-value. This can be considered additional information which the unweighted version could not uncover, in accordance with the existing literature [13].

As a final note, it is clear that the information returned by the weighted *Wilcoxon*'s test is a revenue for the weights we have computed at the input of the test, but this constitutes a desirable approach: we are orienting our analysis and conclusions to be based on the properties of the data we are interested in.

Acknowledgment. José A. Sáez holds a *Juan de la Cierva-formación* fellowship (*Ref. FJCI-2015-25547*) from the Spanish Ministry of Economy, Industry and Competitiveness.

References

1. Bach, F.: Breaking the curse of dimensionality with convex neural networks. J. Mach. Learn. Res. **18**, 1–53 (2017)
2. Bello-Orgaz, G., Jung, J., Camacho, D.: Social big data: recent achievements and new challenges. Inf. Fusion **28**, 45–59 (2016)
3. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
4. Dua, D., Karra Taniskidou, E.: UCI machine learning repository (2017). http://archive.ics.uci.edu/ml
5. Jain, S., Shukla, S., Wadhvani, R.: Dynamic selection of normalization techniques using data complexity measures. Expert Syst. Appl. **106**, 252–262 (2018)
6. Khalilpour Darzi, M., Niaki, S., Khedmati, M.: Binary classification of imbalanced datasets: the case of coil challenge 2000. Expert Syst. Appl. **128**, 169–186 (2019)
7. Kuncheva, L., Galar, M.: Theoretical and empirical criteria for the edited nearest neighbour classifier, vol. January, pp. 817–822 (2016)
8. Larose, D.T., Larose, C.D.: Data Mining and Predictive Analytics, 2nd edn. Wiley Publishing, Hoboken (2015)
9. Luengo, J., García, S., Herrera, F.: A study on the use of imputation methods for experimentation with radial basis function network classifiers handling missing attribute values: the good synergy between RBFs and eventcovering method. Neural Networks **23**(3), 406–418 (2010)
10. Nettleton, D., Orriols-Puig, A., Fornells, A.: A study of the effect of different types of noise on the precision of supervised learning techniques. Artif. Intell. Rev. **33**, 275–306 (2010)
11. Quade, D.: Using weighted rankings in the analysis of complete blocks with additive block effects. J. Am. Stat. Assoc. **74**, 680–683 (1979)
12. Sáez, J.A., Galar, M., Luengo, J., Herrera, F.: INFFC: an iterative class noise filter based on the fusion of classifiers with noise sensitivity control. Inf. Fusion **27**, 19–32 (2016)
13. Sáez, J.A., Luengo, J., Herrera, F.: Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification. Pattern Recogn. **46**(1), 355–364 (2013)
14. Santafe, G., Inza, I., Lozano, J.: Dealing with the evaluation of supervised classification algorithms. Artif. Intell. Rev. **44**(4), 467–508 (2015)
15. Singh, P., Sarkar, R., Nasipuri, M.: Significance of non-parametric statistical tests for comparison of classifiers over multiple datasets. Int. J. Comput. Sci. Math. **7**(5), 410–442 (2016)
16. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
17. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE Trans. Evol. Comput. **1**(1), 67–82 (1997)
18. Zar, J.: Biostatistical Analysis. Prentice Hall, Upper Saddle River (2009)