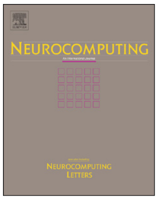
Neurocomputing 176 (2016) 26–35

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Evaluating the classifier behavior with noisy data considering performance and robustness: The Equalized Loss of Accuracy measure

José A. Sáez [a],*, Julián Luengo [b], Francisco Herrera [a]

[a] Department of Computer Science and Artificial Intelligence, University of Granada, CITIC-UGR, Granada 18071, Spain
[b] Department of Civil Engineering, LSI, University of Burgos, Burgos 09006, Spain

ARTICLE INFO

Article history:
Received 12 March 2014
Received in revised form
31 October 2014
Accepted 2 November 2014
Available online 8 May 2015

Keywords:
Noisy data
Behavior against noise
Robustness measures
Classification

ABSTRACT

Noise is common in any real-world data set and may adversely affect classifiers built under the effect of such type of disturbance. Some of these classifiers are widely recognized for their good performance when dealing with imperfect data. However, the noise robustness of the classifiers is an important issue in noisy environments and it must be carefully studied. Both performance and robustness are two independent concepts that are usually considered separately, but the conclusions reached with one of these metrics do not necessarily imply the same conclusions with the other. Therefore, involving both concepts seems to be crucial in order determine the expected behavior of the classifiers against noise. Existing measures fail to properly integrate these two concepts, and they are also not well suited to compare different techniques over the same data. This paper proposes a new measure to establish the expected behavior of a classifier with noisy data trying to minimize the problems of considering performance and robustness individually: the Equalized Loss of Accuracy (ELA). The advantages of ELA against other robustness metrics are studied and all of them are also compared. Both the analysis of the distinct measures and the empirical results show that ELA is able to overcome the aforementioned problems that the rest of the robustness metrics may produce, having a better behavior when comparing different classifiers over the same data set.

1. Introduction

It is widely known that classifier performance is influenced by the quality of the training data upon which this classifier is built [25]. Since real-world data sets are rarely clean of corruptions or *noise* [27,8], these can therefore affect the decisions taken by the classifiers built from these data. However, the maximum achievable performance depends not only on the quality of the data, but also on the appropriateness of the chosen classification algorithm for the data.

Knowing what kind of classification algorithms are most suitable when working with noisy data is a challenging proposition [14,18,19]. Ideally, since the systems must be adapted to the data they treat, if the data that we train are characterized by their inaccuracy, then systems that create classifiers capable of handling some degree of imprecission are needed [25]. One may wonder how to know which systems are more suitable or are better adapted to deal with these noisy data. Even though some classifiers have been related to this capability of working with imperfect data, this fact is usually based on only checking the accuracy of those and other classifiers over a concrete collection of data sets, with independence of the type and noise level present in the data. This analysis procedure has important disadvantages in noisy environments. First of all, the study of the performance alone does not provide enough information on the classifier behavior affected by the noise [12,11,20]. Moreover, a study with a controlled (probably artificial) noise level for each data set is also necessary to reach meaningful conclusions when evaluating the classifier behavior against noise [27]. Finally, it is also desirable to fairly compare different classifiers over the same data, taking into account not only the decrement in performance when noise increases, but the performance when no noise is present as well.

This paper proposes a new single score to perform an analysis of the classifier behavior with noisy data trying to solve the aforementioned problems. This will be done from a double point of view focusing on the classic performance assessment of the methods but also on their robustness [12,11,20], an important issue in noisy environments that must be carefully studied. We understand as performance the accuracy of a classifier predicting the class of a new example, whereas the noise robustness has been defined as the classifier accuracy loss rate [12,11], which is produced by presence of noise in the data, with respect to the case without noise. Since performance and robustness are

* Corresponding author. Tel.: +34 958 240598; fax: +34 958 243317.
E-mail addresses: smja@decsai.ugr.es (J.A. Sáez), jluengo@ubu.es (J. Luengo), herrera@decsai.ugr.es (F. Herrera).

http://dx.doi.org/10.1016/j.neucom.2014.11.086
0925-2312/© 2015 Elsevier B.V. All rights reserved.

different concepts, the conclusions that they provide may also be different, yet this comparative analysis remains disregarded in the literature.

Even though the robustness of the methods is important in dealing with noisy data, there are lack of proposals of robustness-based measures in the literature and the few existing ones also present several drawbacks. This paper will analyze the existing robustness measures in the classification framework focusing on their advantages and disadvantages. We will motivate the necessity of combining the robustness and performance concepts to obtain a unified conclusion on the expected behavior of the methods with noisy data. We will propose a new behavior-against-noise measure to characterize the behavior of a method with noisy data, the Equalized Loss of Accuracy (ELA) measure, which tries to minimize the problems of considering performance and robustness measures individually and can be used to compare different classifiers with ease.

In order to complete our analysis, we will perform an experimental evaluation of the behavior and representativeness of the different measures, considering several classifiers with a known behavior against noise (concretely, the C4.5 decision tree generator [17] and a Support Vector Machine [7]). The behavior of such classifiers described by using ELA will be tested using 32 data sets from the KEEL-dataset repository [2], over which we will introduce a 10% of noise level into the class labels in a controlled way [27]. All these data sets and other complementary material associated with this paper, such as the performance and robustness-based metrics results, are available at the web-page http://sci2s.ugr.es/ela_noise.

The rest of this paper is organized as follows. Section 2 presents an introduction to noisy data and robustness in classification. Next, Section 3 describes the new proposed measure ELA. Section 4 shows the details of the experimental framework including the noise introduction process, the parameter setup for the algorithms and the comparison methodology. Section 5 includes the analysis of the experimental results obtained with the different robustness-based metrics. Finally, in Section 6 we point out some concluding remarks.

2. Classification with noisy data

This section presents an introduction to noisy data in the field of classification, found in Section 2.1. Then, the concept of robustness in classification is explained in Section 2.2.

2.1. Introduction to noisy data

The quality of any data set is determined by a large number of components as described in [23]. Two of these are the source of the data and the input of the data, which are inherently subject to error. Thus, real-world data is rarely perfect; it is often affected by corruptions that hinder the models built as well as the interpretations and decisions made from them. In the particular framework of classification, the most notable effect of noise is that it negatively affects the system performance in terms of classification accuracy, time in building, size and interpretability of the model obtained [26,27]. In the literature there are two types of noise distinguished [24]:

1. *Class noise* [4,1]: Also known as labeling errors, they occur when an instance belongs to the incorrect class due to, for example, data entry errors or inadequacy of the information used to label each instance.
2. *Attribute noise* [27,22]: This is used to refer to corruptions in the attribute values of instances in a data set. Examples of attribute noise include: erroneous attribute values, missing or unknown attribute values, and incomplete attributes or “do not care” values.

In this paper we consider the most common type of class noise, which is also the most disruptive; this is known as *misclassifications* and refers to those examples incorrectly labeled with a wrong class label [27].

It is important to note that the concept of noise used in this paper is different from that of outlier [10]. Thus, an outlier e_o is an example of a concrete class L_A which appears to be inconsistent with respect to other examples of the same class L_A, since it is situated within a different class L_B. Even though this definition may seem similar to that of noise, the outlier example e_o has not errors in its class label or its attributes (therefore, it is different from a noisy example) and its correct classification, although surprising, is the class L_A.

Since errors in real-world data sets are common, actions must be taken to mitigate their consequences [24]. Several methods have been studied in the literature to deal with noisy data [27]. They follow two main postulates: (i) the adaptation of the algorithms to properly handle the noise [17,6] and; (ii) the preprocessing of the data sets aiming to remove or correct the noisy examples [3].

The methods that follow the first postulate are also known as *robust learners* and they are characterized by being less influenced by noisy data. An example of a robust learner is the C4.5 algorithm [17] considered in the experimental case of study of this paper, which uses pruning strategies to reduce the chances that the trees are overfitting to noise in the training data [16]. However, if the noise level is relatively high, even a robust learner may have a poor performance.

Noise preprocessing techniques, which follow the second postulate, have shown a good behavior dealing with noisy data in standard classification (particularly with class noise [3,13]). However, these methods may present problems with imbalanced data [5], since they may recognize the examples belonging to the minority class as noisy examples and therefore, these examples may be removed from the data. In spite of that fact, since we relate noise to the idea of errors, minority class examples are not necessarily noisy if they are error-free.

2.2. Robustness measures

Noise hinders the knowledge extraction from the data and spoils the models obtained using these noisy data when they are compared to the models learned from clean data from the same problem [27]. In this sense, robustness [9] is the capability of an algorithm to build models that are insensitive to data corruptions and suffer less from the impact of noise; that is, the more robust an algorithm is, the more similar the models built from clean and noisy data are. Thus, a classification algorithm is said to be more robust than another if the former builds classifiers which are less influenced by noise than the latter. Robustness is considered very important when dealing with noisy data, because it allows one to expect *a priori* the amount of variation of the learning method's performance against noise with respect to the noiseless performance in those cases where the characteristics of noise are unknown. It is important to note that a higher robustness of a classifier does not imply a good behavior of that classifier with noisy data, since a good behavior implies a high robustness but also a high performance without noise.

In the literature, the measures that are used to analyze the degree of robustness of the classifiers in the presence of noise compare the performance of the classifiers learned with the original (without controlled noise) data set with the performance of the classifiers learned using a noisy version of that data set. Therefore, those classifiers learned from noisy data sets that are more similar (in terms of results) to the noise-free classifiers will be the most robust ones. To the best of our knowledge, the robustness-based measures found in the literature are the two following:

1. The robustness measure proposed in [12] considers the performance of Bayesian Decision rule as a reference, which is considered as the classifier providing the minimal risk when the training data are not corrupted. Concretely, the next expression is used

$$BRM_{x\%} = \frac{E_{x\%} - E}{E}, \tag{1}$$

where $E_{x\%}$ is the risk (that we will understand as the classification error rate in our case) of the classifier at a noise level $x\%$ and E is the risk of the Bayesian Decision rule without noise. This classifier is a theoretical decision rule, that is, it is not learned from the data, which depends on the data generating process. Its error rate is by definition the minimum expected error that can be achieved by any decision rule.
2. The *Relative Loss of Accuracy* (RLA) is the robustness measure employed in [20] and was defined as

$$RLA_{x\%} = \frac{A_{0\%} - A_{x\%}}{A_{0\%}}, \tag{2}$$

where $A_{0\%}$ is the accuracy of the classifier with a noise level 0%, and $A_{x\%}$ is the accuracy of the classifier with a noise level $x\%$. RLA evaluates the robustness as the loss of accuracy with respect to the case without noise $A_{0\%}$, weighted by this value $A_{0\%}$. This measure has two clear advantages: (i) it is simple and interpretable and (ii) to the same values of loss $A_{0\%} - A_{x\%}$, the methods having a higher value of accuracy without noise $A_{0\%}$ will have a lower RLA value.

In the next section, some shortcomings of these measures are explained.

3. The Equalized Loss of Accuracy measure

This section discusses the problems of the existing robustness-based measures as indicators of the behavior of a classifier with noisy data (Section 3.1) and the necessity of combining the robustness and the performance of the classifier (Section 3.2). Finally, we present the ELA measure as our proposal (Section 3.3).

3.1. Problems of the existing robustness measures

Even though the robustness-based measures presented in the above section let us to evaluate the higher or lower robustness of the classifiers, they have a series of important disadvantages which do not make their usage recommendable. In [11] two main points that a robust algorithm must satisfied are established:

1. It must have a good performance without noise.
2. When the noise level is increased, it must suffer a low loss if the noise level is low and the performance must not be drastically deteriorated when the noise level is high.

The measure proposed in [12] (Eq. (1)) considers these two points, even though it has a clear disadvantage: it is based on a theoretical model. Thus, the quality of the performance without noise is determined with respect to that of the optimal Bayesian classifier for the data. This optimal performance can be rarely computed since it is typically characterized by probability distributions on the input/output space, which are intrinsic to the data set and are rarely known. The estimation of these distributions from the data generally requires the choice of a known probability distribution that could not properly represent the characteristics of the data. For this reason this measure is not feasible for practical cases, where mixed data and computation time bound the obtention of the Bayesian classifier.

On the other hand, the RLA measure (Eq. (2)) has several points considered as drawbacks:

- From the two points implied in the definition of a robust algorithm given above, point 1, that is, the necessity to have a good initial performance $A_0\%$, has a very low influence in the RLA equation; being point 2 the main aspect computed by it.
- Classifiers obtaining a poor generalization from the training data without noise, that is, those in which $A_{0\%}$ is low, are usually affected in a lower degree by the presence of noise (their RLA value is therefore lower) than classifiers with a high $A_{0\%}$. In these cases, the lower loss of accuracy is not due to the better capability of the algorithm to get adapted to noise but for their inability to successfully model the data and for creating too general models that are little affected by noise.
- The RLA values do not represent the behavior against noise. For example, consider a random classifier in a balanced data set with $A_{0\%} = 50\%$. This classifier may maintain an accuracy $A_{x\%} = 50\%$ for different noise levels $x\%$ implying a robustness $RLA_{x\%} = 0$. On the other hand, a classifier with a higher starting accuracy suffering from a very low loss of accuracy when noise level increases has higher RLA values always $RLA_{x\%} > 0$, and then it is less robust, even though its behavior with noisy data is better.
- The RLA measure presents problems if $A_{x\%}$ is higher than the base accuracy $A_{0\%}$, obtaining negative numbers of RLA. This fact is more frequent with classifiers with a low base accuracy $A_{0\%}$, whereas it is more rare with a good classifier with a high base accuracy $A_{0\%}$. These low negative values are interpreted as an excellent robustness, but they denote a very bad working of the classifier without noise.

Apart from the aforementioned problems, the main drawback of the RLA measure is that it assumes that both methods have the same robustness ($RLA_{0\%} = 0$) in the case without controlled noise $x = 0\%$. However, the information of the robustness without controlled noise must be also taken into account. If we are interested in analyzing only a single classifier, the RLA measure may suffice, but it fails when comparing two different methods when their performance without noise are different as important information is being ignored. RLA analyzes the robustness in the classic sense of variation with respect to the case without noise and thus the problem of knowing which methods will behave better with noisy data is considered partially. Therefore, it seems necessary to somehow combine the robustness in the sense of performance variation (as RLA makes) with the behavior without noise as it is performed in [12], but determining the quality of that initial accuracy without depending of the results of any external nor theoretical classifier.

3.2. Combining performance and robustness

As we have commented above, focusing only in the robustness, such as RLA makes it, in order to determine the behavior of several methods against noise is a partial way to address the problem. Thus, it is also important to properly consider the performance of these methods without noise. For example, consider a classifier C_1 that is only slightly affected by the noise and another classifier C_2 that is affected by the noise in a higher degree. If we ignore the initial accuracy $A_{0\%}$ of both methods without noise, the following two cases can be produced:

- If both methods C_1 and C_2 obtain two high and similar performances without noise, we would probably choose C_1 as the more robust method as it probably outperforms the method C_2 so far when we deal with new noisy data sets.
- If the performance of C_1 is significantly lower than that of C_2 without noise, then the method C_2 could be expected to be more accurate than the method C_1 when noise appears thanks to C_2's initial good behavior in spite of having a higher degradation of performance.

If we consider the usage of RLA, the second case would be incorrectly described. Furthermore, with the RLA measure, bad classifiers will have less probability to deteriorate their results to the same scale that a good classifier (they could indeed improve their results in an extreme case) when noise is introduced. Finally, in order to better understand the importance of the initial performance (A_0), consider how the RLA measure is defined for the limits of the initial performance A_0. For a classifier C_1 with initial performance $A_0 = 1$, the only possible variation when introducing noise is that the classifier to be hindered as $A_x \le A_0$. However, for another classifier C_2 with an very low initial accuracy $A_0 \approx 0$ the opposite may occur, being probably $A_x \ge A_0$. Therefore, we will obtain that $RLA(C_1) \le RLA(C_2)$. That would mean that the worst imaginable classifier behaves better with noise than the almost perfect classifier.

3.3. The ELA measure

In order to overcome the problems mentioned in the above sections, we propose a correction of the RLA measure inspired in the measure proposed in [12]. The new measure is

$$ELA_{x\%} = \frac{100 - A_{x\%}}{A_{0\%}}. \tag{3}$$

Using a pessimistic approach comparing to the perfect classifier instead of the optimal theoretical Bayesian classifier, it is possible to derive the expression mentioned as

$$\begin{aligned} ELA_{x\%} &= \frac{100 - A_x}{A_0} = \frac{100 - A_x + A_0 - A_0}{A_0} = \frac{A_0 - A_x}{A_0} + \frac{100 - A_0}{A_0} \\ &= RLA_{x\%} + f(A_0) \end{aligned} \tag{4}$$

Therefore, ELA combines the robustness computed by RLA and a factor depending on the initial accuracy A_0 ($f(A_0)$ in Eq. (4). Please, note that this factor $f(A_0)$ is precisely $ELA_{0\%}$, that is, $f(A_0) = ELA_{0\%} = (100 - A_0)/A_0$. Therefore, if we define $ELA_{x\%}$ as a measure of *behavior with noise* at a given noise level $x\%$, then $ELA_{x\%}$ is based on:

- The robustness of the method, that is, the loss of performance at a controlled noise level $x\%$ ($RLA_{x\%}$).
- The *behavior with noise* for the clean data, that is, without controlled noise ($ELA_{0\%}$).

Fig. 1 shows a graphical representation of the RLA and ELA measures. Both of them are functions of two variables: the performance without noise ($A_{0\%}$) and the performance at a noise level $x\%$ ($A_{x\%}$). Therefore, they require a 3-dimensional representation. For each pair of values of $A_0\%$ and $A_{x\%}$, a value of either RLA or ELA is obtained (in the vertical axis z). In these figures, several similarities and differences among the two metrics (RLA and ELA) can be appreciated. First of all, both metrics have similar values when $A_{0\%}$ is high. However, they diverge along with the decrement of $A_{0\%}$ (even though both RLA and ELA have higher values when $A_{x\%}$ is lower and lower values when $A_{x\%}$ is higher). This divergence is produced thanks to the correction obtained by considering the initial accuracy without noise in the ELA measure and it is essential to overcome the limitations of RLA.

The ELA measure changes the initial reference $A_{0\%}$ of the RLA measure by a constant value. As expressed by the BRM measure, the constant value should be the best attainable accuracy value, and as proposes BRM the optimal Bayesian Decision Rule should be used to obtain a theoretical best accuracy value based on the joint underlying distributions of the data set. However, and as we have previously stated, this optimal value is rarely known. For this reason we choose an upper bound to this unknown in practice optimal Bayesian classifier's accuracy instead. The safer and most pessimistic value used for A_0 is fixed to 100% considering it as the accuracy of a perfect classifier. In this way, the loss of accuracy respect to the perfect classifier is weighted by the base accuracy $A_{0\%}$. As a result when taking into account the same loss of accuracy $100 - A_{x\%}$, the classifier with better value of base accuracy $A_{0\%}$ is considered to have a better behavior against noise.

This measure used to evaluate the behavior of a classifier with noisy data overcomes some problems of the RLA measure:

1. It takes into account the noiseless performance $A_{0\%}$ when considering which classifier is more appropriate with noisy data. This fact makes ELA more suitable to compare the behavior against noise between 2 different classifiers. We must take into account that a benchmark data set might contain implicit and not controlled noise with a noise level $x = 0\%$.
2. A classifier with a low base accuracy $A_{0\%}$ that is not deteriorated at higher noise levels $A_{x\%}$ is not better than another better classifier suffering from a low loss of accuracy when the noise level is increased.

Table 1 shows four simple examples that extend the aforementioned ideas about RLA and ELA. These examples describe the possible problems caused by RLA and how they are solved by the ELA measure. In these examples, we consider a classifier C_1 with a good initial performance without noise and a classifier C_2 with a worse performance without noise. Each one of the four toy examples is composed of three different scenarios describing it (one scenario per row), where the performances of the two classifiers C_1 and C_2 without noise ($A_{0\%}$) and with a noise level $x\%$ ($A_{x\%}$) are used to compute the RLA and ELA metrics. For a graphical representation of the performance of these two classifiers of each one of the toy examples shadowed in this table see Fig. 2. In the following we describe the studied scenarios along with the four examples:

Example 1. C_1 and C_2 are equally hindered when the noise is introduced.

In this case, C_1 is always more robust than C_2 with both measures (RLA and ELA). To the same amount of loss of accuracy, both ELA and RLA give more importance to the method with a higher performance (C_1). However, ELA makes more remarkable the difference between C_1 and C_2 since it takes into account the performance without noise ($A_{0\%}$) and the loss of accuracy ($A_{x\%}$). This information shows that is much probably that C_1 behaves well with noisy data considering ELA, whereas with RLA the differences would be much lower.

Furthermore, it is important to note that, when the performance of both classifiers without and with noise do not vary, RLA gives the same importance to both of them, but ELA clearly establishes C_1 as that classifier with the best behavior with noisy data (and this fact fits more to the desirable answer than that provided by RLA).

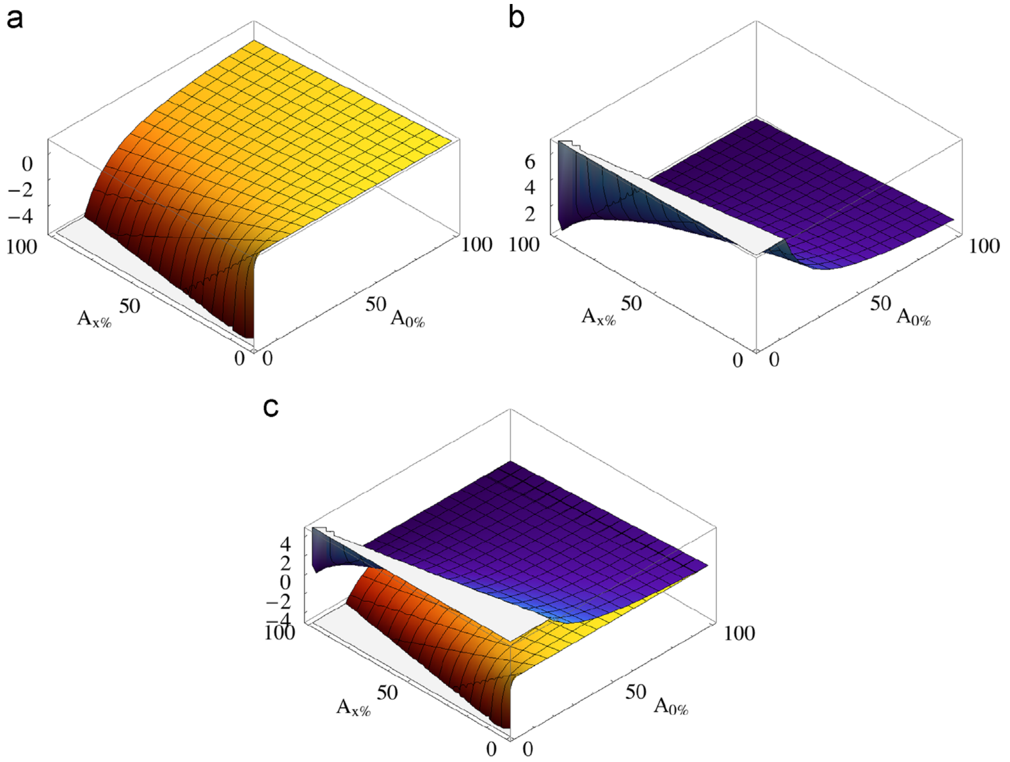


Fig. 1. Representations of the RLA and ELA metrics. The vertical axis represents the value of each metric. (a) $\text{RLA}_{x\%}$. (b) $\text{ELA}_{x\%}$. (c) $\text{RLA}_{x\%}$ vs $\text{ELA}_{x\%}$.

Table 1
Four different illustrative toy examples comparing RLA versus ELA.

	Classifier C_1				Classifier C_2			
Example	$A_{0\%}$	$A_{x\%}$	RLA	ELA	$A_{0\%}$	$A_{x\%}$	RLA	ELA
Example 1	100	100	0	0	50	50	0	1
	100	96	0.04	0.04	50	46	0.08	1.08
Fig. 2(a)	100	92	0.08	0.08	50	42	0.16	1.16
Example 2	100	100	0	0	50	50	0	1
	100	96	0.04	0.04	50	48	0.04	1.04
Fig. 2(b)	100	92	0.08	0.08	50	46	0.08	1.08
Example 3	80	80	0	0.25	30	30	0	2.33
	80	84	−0.05	0.2	30	34	−0.13	2.2
Fig. 2(c)	80	88	−0.1	0.15	30	38	−0.27	2.07
Example 4	100	100	0	0	50	50	0	1
	100	99.996	0.00004	0.00004	50	50.004	−0.00008	0.99992
Fig. 2(d)	100	99.992	0.00008	0.00008	50	50.008	−0.00016	0.99984

Example 2. C_1 and C_2 are hindered when the noise is introduced but both have the same RLA values.

In this case, even though the classifier C_2 is equally robust than C_1, C_2 obviously is not better than C_1 and ELA is able to reflect this issue. This situation (to a same RLA value equal to 0) also occurs if the classifiers do not alter their performance when the noise is introduced: from $A_{0\%}=100$ to $A_{x\%}=100$, then $\text{ELA}(C_1)=0$, whereas from $A_{0\%}=50$ to $A_{x\%}=50$, $\text{ELA}(C_2)=1$, so C_1 will behave better with noisy data than C_2.

Example 3. C_1 and C_2 are benefited when the noise is introduced.

In this rare case, both methods gain the same amount of performance when the noise is introduced. RLA shows that method C_2, which has a very poor performance, is more robust than the method C_1. ELA solves this problem and shows that the classifier C_1 has a better behavior with noisy data than the classifier C_2.

Example 4. C_1 is slightly affected by the noise whereas C_2 is slightly benefited by the noise.

This case clearly shows that RLA does not take into account the initial accuracy. Again, the classifier C_2 is more robust with RLA, whereas it clearly behaves worse with the ELA measure.

4. Experimental framework

In this section, we present the details of the experimentation developed in this paper. We first show how to build the noisy data sets in Section 4.1. Then, Section 4.2 indicates the classification methods used and their parameters. Finally, Section 4.3 establishes the analysis methodology carried out.

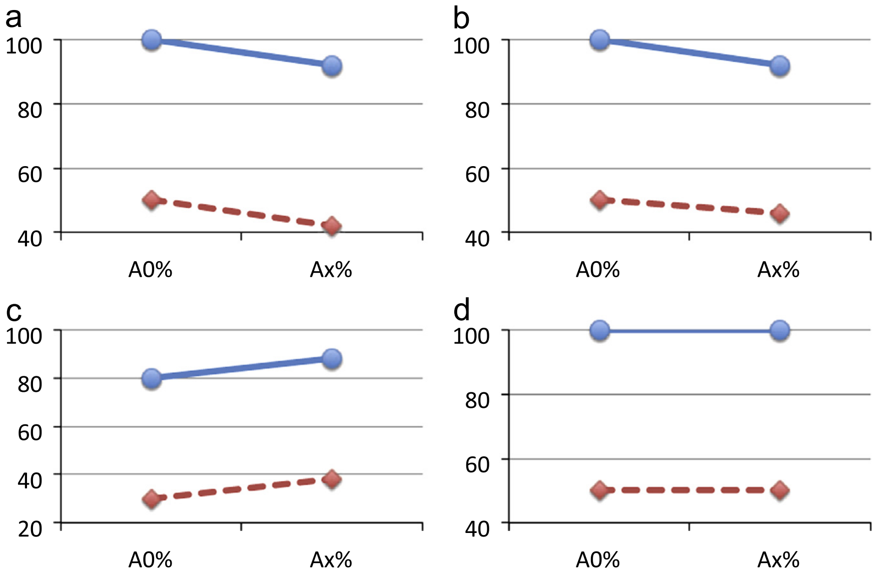


Fig. 2. Graphical representation of the performance of the two classifiers C_1 and C_2 of four different illustrative toy examples shadowed in Table 1. Each graphic shows the performance of C_1 and C_2 without noise ($A_{0\%}$) and with a noise level $x\%$ ($A_{x\%}$). Classifier C_1 is represented with a continuous line (—) and Classifier C_2 is represented with a dashed line (– – –). (a) Example 1. (b) Example 2. (c) Example 3. (d) Example 4.

4.1. *Data sets*

The experimentation has been based on 32 data sets taken from the KEEL-dataset repository[1] [2]. Table 2 summarizes the properties of the originally selected data sets. For each data set, the number of examples (#EX), the number of attributes, differentiating between numeric and nominal attributes (#AT (Numeric/Nominal)) and the number of classes (#CL) are presented. The percentage of examples of the class with the lowest number of examples (%Min) and the class with the highest number of examples (%Maj) are also shown.

In order to control the noise level in the existing data, we use a manual mechanism to add noise into each training data set. Thus, we have considered the introduction of class noise following the scheme proposed in [27,21]. This scheme, also known as *random class noise scheme*, introduces a noise level $x\%$ into a data set by randomly changing the class labels of exactly $x\%$ of the examples by other one out of the other classes.

The accuracy estimation of each classifier is obtained by means of 5 runs of a stratified 5-fold cross-validation. The data set is divided into 5 partition sets with equal numbers of instances and maintaining the proportion between classes in each fold. Each partition set is used as a test for the classifier learned from the four remaining partitions. This procedure is repeated 5 times.

4.2. *Parameters*

Two learning algorithms have been chosen to be used in this paper: C4.5 [17] and SVM [7]. This choice is based on their good behavior in a large number of real-world problems; moreover, they were selected because these methods have a highly differentiated and well known noise-robustness. In the following, their noise-tolerance is described along with the parameter configuration used for the experimentation:

- *C4.5 decision tree generator* [17]: C4.5 is considered a robust learner, which uses pruning strategies to reduce the chances of classifiers being affected by noisy examples [16]. The parameter setup for C4.5 used in this paper is the following: confidence level (0.25), minimal instances per leaf (2) and prune after the tree building.
- *Support Vector Machine* [7]: Since SVM relies on the support vectors (that are training examples lying near the separating hyperplane) to derive the decision model, this can be easily altered including or excluding a single noisy example [15]. Thus, SVM should *a priori* be more noise-sensitive than C4.5. The parameter setup for SVM used in this paper is the following: type of Kernel (*Puk* with $\sigma=1$, $\omega=1$), cost ($C=100$), tolerance (0.001) and parameter for the round-off error ($\epsilon=10^{-12}$).

4.3. *Methodology of analysis*

The experimental analysis of the capabilities of the ELA measure will be based on a complete case of study which involves the two aforementioned classification algorithms with a different noise tolerance: the noise-robust algorithm C4.5 and the noise-sensitive method SVM. These methods will be tested over the 32 base data sets without noise, that is $x=0\%$, and another 32 noisy data sets with the noise level $x=10\%$, which will be created with the *random class noise scheme*. All the data sets created can be found on the web-page associated with this paper.

The classification accuracy of C4.5 and SVM will be computed on the 64 data sets (without and with noise), along with their corresponding ELA and RLA results for the noise level 10%. Please note that it is not our intention to establish the most robust method between C4.5 and SVM, but to provide an ample and varied test bed where the two methods' behavior will help us to show the benefits of ELA measure against RLA. Because of this, our analysis will be based on studying the similarities and differences between the evaluations of ELA and RLA on the behavior with noise of each classification algorithm with each data set.

5. Benefits of ELA against other robustness metrics: a case of study

In this section we focus on the analysis of the behavior of the classifiers to study (C4.5 and SVM) when training with noisy data

[1] http://www.keel.es/datasets.php

Table 2
Base data sets used in the experimentation.

Data set	#EX	#AT	#CL	%Min	%Maj	Data set	#EX	#AT	#CL	%Min	%Maj
automobile	159	25(15/10)	6	1.89	30.19	magic	19 020	10(10/0)	2	35.16	64.84
balance	625	4(4/0)	3	7.84	46.08	monk	432	6(6/0)	2	47.22	52.78
banana	5300	2(2/0)	2	44.83	55.17	new-thyroid	215	5(5/0)	3	13.95	69.77
car	1728	6(0/6)	4	3.76	70.02	phoneme	5404	5(5/0)	2	29.35	70.65
cleveland	297	13(13/0)	5	4.38	53.87	pima	768	8(8/0)	2	34.90	65.10
contraceptive	1473	9(9/0)	3	22.61	42.70	ring	7400	20(20/0)	2	49.51	50.49
dermatology	358	33(1/32)	6	5.59	31.01	segment	2310	19(19/0)	7	14.29	14.29
ecoli	336	7(7/0)	8	0.60	42.56	sonar	208	60(60/0)	2	46.63	53.37
flare	1066	11(0/11)	6	4.03	31.05	spambase	4597	57(57/0)	2	39.42	60.58
german	1000	20(13/7)	2	30.00	70.00	twonorm	7400	20(20/0)	2	49.96	50.04
glass	214	9(9/0)	7	4.21	35.51	vehicle	846	18(18/0)	4	23.52	25.77
hayes-roth	160	4(4/0)	3	22.73	38.64	vowel	990	13(13/0)	11	9.09	9.09
heart	270	13(13/0)	2	44.44	55.56	wdbc	569	30(30/0)	2	37.26	62.74
ionosphere	351	33(33/0)	2	35.90	64.10	wine	178	13(13/0)	3	26.97	39.89
iris	150	4(4/0)	3	33.33	33.33	yeast	1484	8(8/0)	10	0.34	31.20
lymphography	148	18(3/15)	4	1.35	54.73	zoo	101	16(0/16)	7	3.96	40.59

considering the usage of the ELA and RLA measures. As we cannot know the probability distribution of the benchmark data sets, we cannot use BRM as a comparison measure. Table 3 shows the performance results of C4.5 and SVM for all the data sets considered in this paper (at the noise levels 0% and 10% as indicated in Section 4.3), and their ELA and RLA results. From this case of study, several observations can be appreciated, which can be grouped into two main parts: global results (including the *average* and *best* rows in Table 3) and the individual results for each data set. These remarks on the results presented in this table will be focused on the similarities and differences between RLA and ELA, attending to the problems of RLA and how ELA can solve them.

Analysis of the global results: This part of the analysis compares the average results for both C4.5 and SVM across all the data sets, by using the performance and those average results of the ELA and RLA metrics, and the number of data sets where each classifier is the best.

Regarding the performance results, SVM has a better performance without noise than C4.5, obtaining an average performance of 82.24 versus a 81.28 of C4.5. Furthermore, SVM is also better in more data sets than C4.5, concretely in 20 of the 32 data sets. The situation reverses when noise is considered, and C4.5 obtains a better average performance (80.37 versus 78.50 of SVM) and the same number of data sets in which each classifier is the best (16 data sets in total). Note that these results without and with noise are consistent with the expected behavior of both classifiers.

Since the results of SVM for many of the data sets drop in an higher degree than those of C4.5, their average RLA value is therefore higher than that of C4.5 (0.0457 versus 0.0115). Thus, SVM is clearly less robust considering this metric. The average of the ELA measure also offers the same final result, showing to C4.5 as the method that globally behaves better with noisy data. These average results of both ELA and RLA are again in concordance with the expected behavior of the two implied classifiers, since we knew that C4.5 could probably have a better behavior with noise than SVM due to the punning mechanism.

However, as the accuracy of SVM is notably better than that of C4.5 without noise, the number of data sets in which each method is the best considering ELA is not so clear in favor of C4.5 like that of RLA. This fact is due to ELA which considers the initial performance without noise, but it is not taken into account by RLA, that only considers the percentage variation of the performances with and without noise.

These global results highlights that ELA does not only use the loss of performance to evaluate the behavior with noisy data as RLA makes, but also the performance without noise, that must be considered to obtain a good evaluation metric of the behavior against noise as we have previously commented in Section 3.

Analysis of the individual results for each data set: Even though the aforementioned average results give an idea of how ELA and RLA work, it is interesting to observe their results in each single data set to better understand the differences and coincidences between both metrics depending on the behavior of the classifiers. In order to properly analyze these results, we categorize them into two different groups – a group devoted to those data sets in which the evaluation of ELA and RLA agrees and another for those data sets in which this evaluation is different:

1. *ELA and RLA predict the same classifier with the best behavior*: In this case, it is important to note that, even though both metrics provide the same final result, the difference in these values for each classifier could be very different depending on the drop in performance with noise but also on the performance without noise (when ELA is considered). We can differentiate three scenarios within this case:
   - *One of the classifiers is better than the other one with and without noise and both of them are deteriorated from the effect of noise.* There are some data sets (such as *autos* and *wine*) in which both methods have a remarkable and similar loss of performance, whereas in other data (such as *banana*, *dermatology*, *heart*, *phoneme* and *twonorm*) this loss of performance is very low. In all these data sets, SVM has always a better performance than C4.5. However, there are also other data sets, in which C4.5 has the best performance and its loss of accuracy is also lower than that of SVM, see for example *car*, *iris*, *pima*, *zoo* and *monk-2*. In these data sets the method with the best performance is chosen by ELA and RLA as that behaving best with noise.
   - *One of the classifiers improves and the other deteriorates its accuracy without noise when noise is considered.* All these data sets are characterized by C4.5 being the method that behaves best with noisy data. For example, in the *cleveland* data set, both classifiers have a low accuracy without noise (being SVM worse without noise and deteriorating its accuracy, whereas C4.5 improves in presence of noise). The same situation occurs with *ecoli* and *hayes-roth*, even though the initial performance is higher than with *cleveland*.
   - *One of the classifiers has a better performance without noise, but it suffers a very high drop in performance with noise and finally it has a worse performance in the noisy data set.* SVM is usually this classifier with a better performance without noise but it is more affected in the noisy version of the data set (see, for

Table 3
Performance results for C4.5 and SVM at 0% and 10% of class noise level, and their ELA and RLA results at 10% for all the data sets considered. Best results are remarked in bold. Those data sets where ELA and RLA predict a different classifier with the best behavior against noise (in which the result of RLA is not appropriate) are shadowed in gray.

Measure	Performance				ELA		RLA	
Noise level	0%		10%		10%		10%	
Data set	C4.5	SVM	C4.5	SVM	C4.5	SVM	C4.5	SVM
autos	**77.10**	69.29	**73.56**	64.50	**0.3429**	0.5123	**0.0459**	0.0691
balance	77.73	**89.09**	78.27	**81.50**	0.2796	**0.2077**	**−0.0069**	0.0852
banana	88.98	**90.28**	88.96	**90.27**	0.1241	**0.1078**	0.0002	**0.0001**
car	**91.33**	64.83	**90.20**	60.68	**0.1073**	0.6065	**0.0124**	0.0640
cleveland	**51.58**	45.85	**52.26**	41.47	**0.9256**	1.2766	**−0.0132**	0.0955
contraceptive	**52.14**	47.56	**50.52**	46.80	**0.9490**	1.1186	0.0311	**0.0160**
dermatology	93.91	**96.92**	93.02	**96.59**	0.0743	**0.0352**	0.0095	**0.0034**
ecoli	**79.05**	78.11	**79.29**	67.81	**0.2620**	0.4121	**−0.0030**	0.1319
flare	**73.86**	70.34	**74.15**	71.82	**0.3500**	0.4006	−0.0039	**−0.0210**
german	**71.54**	66.44	**71.02**	66.42	**0.4051**	0.5054	0.0073	**0.0003**
glass	66.07	**71.40**	64.46	**65.22**	0.5379	**0.4871**	**0.0244**	0.0866
hayes-roth	**81.67**	77.87	**82.87**	74.55	**0.2097**	0.3268	**−0.0147**	0.0426
heart	77.11	**78.52**	76.22	**77.93**	0.3084	**0.2811**	0.0115	**0.0075**
ionosphere	89.34	**91.91**	**87.52**	81.72	**0.1397**	0.1989	**0.0204**	0.1109
iris	**95.07**	94.53	**94.13**	87.20	**0.0617**	0.1354	**0.0099**	0.0775
lymphography	76.88	**80.82**	77.31	**80.96**	0.2951	**0.2356**	**−0.0056**	−0.0017
magic	85.10	**87.18**	84.69	**86.57**	0.1799	**0.1540**	**0.0048**	0.0070
monk-2	**100.00**	96.25	**100.00**	91.62	**0.0000**	0.0871	**0.0000**	0.0481
newthyroid	92.84	**95.81**	91.53	**92.65**	0.0912	**0.0767**	**0.0141**	0.0330
phoneme	85.88	**87.18**	84.76	**86.66**	0.1775	**0.1530**	0.0130	**0.0060**
pima	**73.99**	69.74	**73.15**	67.37	**0.3629**	0.4679	**0.0114**	0.0340
ring	90.06	**97.09**	88.72	**91.54**	0.1252	**0.0871**	**0.0149**	0.0572
segment	96.35	**97.28**	**95.19**	90.23	**0.0499**	0.1004	**0.0120**	0.0725
sonar	72.50	**86.63**	73.36	**85.94**	0.3674	**0.1623**	**−0.0119**	0.0080
spambase	92.57	**93.57**	**91.32**	91.20	**0.0938**	0.0940	**0.0135**	0.0253
twonorm	84.82	**97.35**	83.86	**96.31**	0.1903	**0.0379**	0.0113	**0.0107**
vehicle	71.04	**80.69**	68.91	**74.75**	0.4376	**0.3129**	**0.0300**	0.0736
vowel	78.59	**99.33**	74.91	**87.11**	0.3193	**0.1298**	**0.0468**	0.1230
wdbc	93.64	**94.41**	**92.79**	91.07	**0.0770**	0.0946	**0.0091**	0.0354
wine	92.23	**97.30**	89.75	**95.95**	0.1111	**0.0416**	0.0269	**0.0139**
yeast	55.54	**57.44**	53.34	**54.39**	0.8401	**0.7940**	**0.0396**	0.0531
zoo	**92.29**	80.64	**91.70**	73.08	**0.0899**	0.3338	**0.0064**	0.0938
average	81.28	**82.24**	**80.37**	78.50	**0.2777**	0.3117	**0.0115**	0.0457
best	12	**20**	**16**	**16**	**16**	**16**	**23**	9

example, the *ionosphere*, *segment*, *spambase* and *wdbc* data sets).

2. *ELA and RLA predict a different classifier with the best behavior*. This case is perhaps more interesting than that of above since each one of the metrics give more importance to a different classifier with noisy data. Thus, we can clearly check the differences between ELA and RLA. We differentiate two scenarios within this case:
   - *One or both classifiers improve in presence of noise*. Note that all the data sets under these circumstances are characterized by ELA giving a higher importance to SVM (the method with the best performance without noise), whereas RLA highlights C4.5 (the method that usually has a lower performance without noise but having a higher improvement when noise is considered). For example, with *balance* and *sonar*, even though C4.5 slightly improves with noisy data whereas SVM has a higher drop in performance, the latter is notably better than C4.5 without and with noise in terms of performance. With other data sets, such as *flare* and *lymphography*, both classifiers improve their results when noise is considered. In these cases, even though RLA gives more importance to the method with an higher improvement in performance, the other method has remarkable better results without noise, and this fact is also considered by ELA.
   - *Both classifiers deteriorate their performance when considering noise*. Some data sets (such as *vehicle*, *ring*, *newtiroid*, *vowel. magic* or *yeast*) are characterized by SVM being more deteriorated than C4.5 by the noise, but also having a higher performance without noise. Thus, RLA highlights C4.5 in this cases, whereas ELA also considered the importance of the higher performance of SVM estimating that it behaves better with noisy data. The opposite fact occurs with the *german* data set, where C4.5 and SVM interchange their behaviors. With the *glass* data set, although SVM is more affected than C4.5 by the presence of noise, it obtains a notable higher accuracy without noise. Thus, ELA establishes SVM as the method with the best behavior with noisy data since it valorizes more than RLA the initial accuracy (although the ELA value of SVM is very similar to that of C4.5), whereas RLA establishes C4.5 as the best method based on the higher drop in performance of SVM.

     Another case is that of the *contraceptive* data set, where the two classifiers obtain a very low accuracy without noise (SVM is indeed worse than C4.5, having a performance lower than 50%). This particular case shows that RLA favors those algorithms with lower classification performances without noise when the loss of performance of the classifiers are comparable (although not equal), whereas ELA does not harm so much the methods with higher performances.

Individual results for each data set again emphasize that RLA is only based on the percentage drop in performance, giving no chance, for example, to those methods that experiment an higher drop but having a very high performance without noise (being competitive enough when the noise is considered). In contrast, ELA takes into account both factors, making its usage overcoming some of the problems of the RLA measure.

6. Concluding remarks

Performance and robustness are two independent concepts that imply different conclusions. Considering both concepts together seems to be crucial in order determine the expected behavior of the classifiers against noise. Existing measures partially consider these aspects, and their values for different classifiers cannot be compared, making a comparative analysis of different classifiers with noise almost impossible. Therefore, a new measure is proposed to know the expected behavior of a classifier with noisy data, the ELA measure, which tries to overcome these problems of the existing robustness-based metrics.

In order to check the suitability of our proposal, we have analyzed the existing robustness measures pointing out their main drawbacks and how ELA can solve them. We have provided a variety of practical examples supporting our analysis. In order to complete this analysis, we have experimentally compared the ELA and RLA measures over real data, showing that the evaluation of the ELA and RLA metrics agree in some cases, but in other cases the behavior of the RLA is not appropriate since it is only based on the percentage variation of the performance without and with noise. Thus, ELA has shown being able to overcome the shortcomings that RLA produces and it is particularly useful when comparing different classifiers over the same data set, a scenario in which RLA usually does not behave as expected.

Acknowledgments

Supported by the Projects TIN2011-28488, P10-TIC-06858 and P11-TIC-7765. J.A. Sáez holds an FPU grant from the Spanish Ministry of Education and Science.

References

[1] J. Abellán, A. Masegosa, Bagging schemes on the presence of class noise in classification, Expert Syst. Appl. 39 (8) (2012) 6827–6837.
[2] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera, KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework, J. Mult. Valued Logic Soft Comput. 17 (2–3) (2011) 255–287.
[3] C.E. Brodley, M.A. Friedl, Identifying mislabeled training data, J. Artif. Intell. Res. 11 (1999) 131–167.
[4] J. Cao, S. Kwong, R. Wang, A noise-detection based ada boost algorithm for mislabeled data, Pattern Recognit. 45 (12) (2012) 4451–4465.
[5] N.V. Chawla, N. Japkowicz, A. Kotcz, Editorial: special issue on learning from imbalanced data sets, SIGKDD Explor. 6 (1) (2004) 1–6.
[6] W.W. Cohen, Fast effective rule induction, in: Proceedings of the Twelfth International Conference on Machine Learning, Morgan Kaufmann Publishers, Lake Tahoe, California, 1995, pp. 115–123.
[7] C. Cortes, V. Vapnik, Support vector networks, Mach. Learn. 20 (1995) 273–297.
[8] B. Frenay, M. Verleysen, Classification in the presence of label noise: a survey, IEEE Trans. Neural Netw. Learn. Syst. 25 (5) (2014) 845–869.
[9] P.J. Huber, Robust Statistics, John Wiley and Sons, New York, 1981.
[10] G.H. John, Robust decision trees: removing outliers from databases, in: Proceedings of the First International Conference on Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, California, 1995, pp. 174–179.
[11] Y. Kharin, Robustness in statistical pattern recognition, in: Mathematics and Its Applications, 1996, vol. 380.
[12] Y. Kharin, E. Zhuk, Robustness in statistical pattern recognition under contaminations of training samples, in: Proceedings of the 12th IAPR International Conference on Pattern Recognition, Conference B: Computer Vision and Image Processing, vol. 2, 1994, pp. 504–506.
[13] T.M. Khoshgoftaar, P. Rebours, Improving software quality prediction by noise filtering techniques, J. Comput. Sci. Technol. 22 (2007) 387–396.
[14] H.-X. Li, J.-L. Yang, G. Zhang, B. Fan, Probabilistic support vector machines for classification of noise affected data, Inf. Sci. 221 (2013) 60–71.
[15] D. Nettleton, A. Orriols-Puig, A. Fornells, A study of the effect of different types of noise on the precision of supervised learning techniques, Artif. Intell. Rev. 33 (2010) 275–306.
[16] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[17] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Francisco, CA, USA, 1993.
[18] J.A. Sáez, M. Galar, J. Luengo, F. Herrera, Tackling the problem of classification with noisy data using multiple classifier systems: analysis of the performance and robustness, Inf. Sci. 247 (2013) 1–20.
[19] J.A. Sáez, M. Galar, J. Luengo, F. Herrera, Analyzing the presence of noise in multi-class problems: alleviating its influence with the one-vs-one decomposition, Knowl. Inf. Syst. 38 (1) (2014) 179–206.
[20] J.A. Sáez, J. Luengo, F. Herrera, Fuzzy rule based classification systems versus crisp robust learners trained in presence of class noise's effects: a case of study, in: 11th International Conference on Intelligent Systems Design and Applications (ISDA 2011), 2011, pp. 1229–1234.
[21] C.-M. Teng, Correcting noisy data, in: Proceedings of the Sixteenth International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, CA, USA, 1999, pp. 239–248.
[22] C.M. Teng, Polishing blemishes: issues in data correction, IEEE Intell. Syst. 19 (2004) 34–39.
[23] R.Y. Wang, V.C. Storey, C.P. Firth, A framework for analysis of data quality research, IEEE Trans. Knowl. Data Eng. 7 (4) (1995) 623–640.
[24] X. Wu, Knowledge Acquisition from Databases, Ablex Publishing Corp., Norwood, NJ, USA, 1996.
[25] X. Wu, X. Zhu, Mining with noise knowledge: error-aware data mining, IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. 38 (4) (2008) 917–932.
[26] S. Zhong, T.M. Khoshgoftaar, N. Seliya, Analyzing software measurement data with clustering techniques, IEEE Intell. Syst. 19 (2) (2004) 20–27.
[27] X. Zhu, X. Wu, Class noise vs. attribute noise: a quantitative study, Artif. Intell. Rev. 22 (2004) 177–210.

José A. Sáez received his M.Sc. in computer science from the University of Granada (Granada, Spain), in 2009. He is currently a Ph.D. student in the Department of Computer Science and Artificial Intelligence in the University of Granada. His main research interests include noisy data in classification, discretization methods and imbalanced learning.

Julián Luengo received the M.S. degree in computer science and the Ph.D. degree from the University of Granada, Granada, Spain, in 2006 and 2011, respectively. His research interests include machine learning and data mining, data preparation in knowledge discovery and data mining, missing values, data complexity and fuzzy systems.

Francisco Herrera received his M.Sc. in mathematics in 1988 and Ph.D. in mathematics in 1991, both from the University of Granada, Spain.

He is currently a Professor in the Department of Computer Science and Artificial Intelligence at the University of Granada. He has been the supervisor of 30 Ph.D. students. He has published more than 260 papers in international journals. He is the coauthor of the book "Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases" (World Scientific, 2001). He currently acts as an Editor in Chief of the international journals "Information Fusion" (Elsevier) and "Progress in Artificial Intelligence (Springer). He acts as an area editor of the International Journal of Computational Intelligence Systems and associated editor of the journals: IEEE Transactions on Fuzzy Systems, Information Sciences, Knowledge and Information Systems, Advances in Fuzzy Systems, and International Journal of Applied Metaheuristics Computing; and he serves as a member of several journal editorial boards, among others: Fuzzy Sets and Systems, Applied Intelligence, Information Fusion, Knowledge-Based Systems, Evolutionary Intelligence, International Journal of Hybrid Intelligent Systems, Memetic Computation, and Swarm and Evolutionary Computation.

He received the following honors and awards: ECCAI Fellow 2009, IFSA Fellow 2013, 2010 Spanish National Award on Computer Science ARITMEL to the "Spanish Engineer on Computer Science", International Cajastur "Mamdani" Prize for Soft Computing (Fourth Edition, 2010), IEEE Transactions on Fuzzy System Outstanding 2008 Paper Award (bestowed in 2011), 2011 Lotfi A. Zadeh Prize Best paper Award of the International Fuzzy Systems Association, and 2013 AEPIA Award to a scientific career in Artificial Intelligence (September 2013).

His current research interests include computing with words and decision making, bibliometrics, data mining, data preparation, instance selection and generation, imperfect data, fuzzy rule based systems, genetic fuzzy systems, imbalanced classification, knowledge extraction based on evolutionary algorithms, memetic algorithms and genetic algorithms, biometrics, cloud computing and big data.