The effects of classifiers diversity on the accuracy of stacking

Mariele Lanes
Centro de Ciências Computacionais
Universidade Federal do Rio Grande
Rio Grande, Brazil
Email: mariele.lanes@furg.br

Eduardo N. Borges
Centro de Ciências Computacionais
Universidade Federal do Rio Grande
Rio Grande, Brazil
Email: eduardoborges@furg.br

Renata Galante
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Porto Alegre, Brazil
Email:galante@inf.ufrgs.br

***Abstract*—In recent years several data classification techniques have been proposed. However, it is not a trivial task to choose the most appropriate classifier for deal with a particular problem and set it up properly. In addition, there is no optimal algorithm to solve all prediction problems. In order to improve the result of the classification process, the stacking strategy combines the knowledge acquired by individual learning algorithms aiming to discover new patterns not yet identified. Stacking combines the outputs of base classifiers, induced by several learning algorithms using the same dataset, by means of a meta-classifier. The main goal of this paper is to evaluate the effects of classifier diversity on the accuracy of stacking. We have performed a lot of experiments which results show the impact of multiple diversity measures on the gain of stacking, considering many real datasets extracted from UCI machine learning repository and three synthetic two-dimensional datasets. The results revealed connections between some measures and the gain of stacking, but they imply a weak or moderate relationship that suggest predicting the improvement on the best base classifier accuracy using diversity measures is inappropriate.**

I. Introduction

Classification is the most usual task among data mining tasks. It is based on the discovery of prediction rules that aid in making decisions. Generally, this task is used when there are large amounts of records in a database, which have several fields, and it is necessary to extract from this base some relevant knowledge with predictive capacity.

Classification algorithms can be organized into different types according to the technical features they use in learning. Each type is best suited for a particular dataset. Although some classifiers individually provide solutions which are considered effective, the experimental evaluation performed by [1] shows a drop in the quality when there are large sets of patterns and/or a significant number of incomplete data samples or irrelevant features. That is, such classifiers may not effectively and/or efficiently recognize patterns in complex problems.

Classifiers that implement different algorithms potentially provide additional information on the patterns to be classified. The combination of the outputs of a set of different classifiers aims to get a more precise classification, i.e. to reach a greater accuracy. In this context, stacking [2] is a widely used method for combining multiple classifiers generated from different learning algorithms applied on the same dataset. It is also known in the literature as stacked generalization [3], [4].

DOI reference number: 10.18293/SEKE2017-016

In the stacking method the choice of base algorithms is very important. It is desirable that there be different solutions to the problem to be solved, i.e., it is important to obtain a diversity among the results found by these algorithms. According to [5], the performance of the stacking method strongly depends on the accuracy and diversity of classifiers results. To verify this diversity, there are several measures based on the agreement and/or disagreement of the classifiers [6].

Therefore, the use of base algorithms with different particulars is ideal, since the patterns learned tend not to be the same. Thus, even low accuracy classifiers combined can generate a strong classifier, providing gain for stacking. Otherwise, when several classifiers agree on the vast majority of responses (no diversity), the combination will possibly have the same result, with no improvement in the stacking quality.

The purpose of this paper is to evaluate the effects of classifier diversity on the accuracy of stacking. The experiments we have performed show the relationship between multiple diversity measures and the gain of stacking, considering a lot of real datasets extracted from UCI machine learning repository and three other synthetic two-dimensional datasets used for visual inspection.

II. Related work

Stacking method combines multiple base classifiers trained by using different learning algorithms L on a single dataset *S*, by means of a meta-classifier [7], [8]. Each training sample $s_j = (X_j, y_j)$ is a pair composed by an array of features X_j and the class label y_j.

The process can be described in two distinct levels. The first level-0 defines a set of N base classifiers, where $C_i = L_i(S)|1 \leq i \leq N$. Level-0 classifiers are trained and tested using the cross-validation or leave-one-out procedure. The output dataset D used for training the meta-classier is composed by examples $((y_j^1, \ldots, y_j^i), y_j)$, i.e. a vector of predictions for each base classifier $y_j^i = C_i(X_j)$ and the same original class label y_j [3]. In the second level-1, the meta-classifier combines base classifiers outputs from D into a final prediction y_j^f. The stacking pseudocode can be seen in Algorithm 1.

Algorithm 1: Combining classifiers with stacking

Input: training samples $s_j \in S$
Output: final predictions y_j^f

```
begin
    Select N learning algorithms (L_1, L_2, ..., L_N);
    for i = 1, 2, ..., N do
        Train C_i = L_i(S) using cross-validation;
        y_j^i = C_i(X_j);
    end
    Make up a new dataset D combining all predictions y_j^i;
    Train M = L(D) using cross-validation;
    y_j^f = M(D);
end
return y_j^f
```

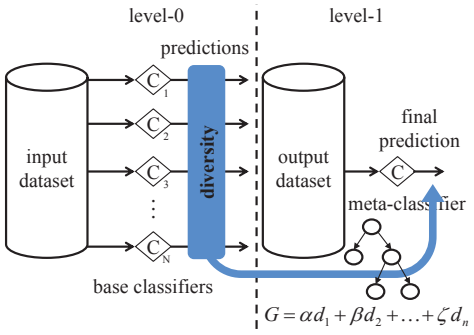

Fig. 1. The proposed method for analyzing the effects of classifiers diversity on the accuracy of stacking.

Several approaches have proposed the use of stacking to increase the classification quality in recent years [9], [10], [11], [12], [13]. Considering the covered related work, the classification techniques most used at level-0 are those based on decision trees, artificial neural networks and Bayes' theorem, which are usually combined by a function-based classifier. The majority of approaches stacks the predicted labels or the confidence of these predictions to compose the feature vector of level-1.

The diversity of predictions y_j^i is a key issue in the combination of classifiers. Reference [6] defines several measures of diversity and relate them to the quality of classification system. Experiment results revealed the used measures were strongly correlated between themselves and the possible inadequacy of the diversity measures for predicting the improvement on the best individual accuracy. Although there are proven connections between diversity and accuracy in some special cases, the results raised some doubts about the usefulness of diversity measures in constructing classifier sets in real-life pattern recognition problems. References [14], [15] also conclude that diversity measures can hardly be used as selection criteria for building ensembles. However, other authors report success using diversity to detect noise [16] and to generate more precise classification systems [17], [18], [19], [20].

III. Proposed Method

The proposed method for analyzing the effects of classifiers diversity on the accuracy of stacking are graphically represented in Figure 1. For each analyzed dataset, different learning algorithms are used to train multiple base classifiers. The predictions returned by these classifiers are evaluated and used to perform several measures of diversity [6]: double-fault df, Disagreement Dis, Q statistic, Correlation coefficient ρ, Kohavi-Wolpert variance KW, Interrater agreement k and Entropy E. These measures check whether and how the classifiers agree or disagree on the predicted class label. At level-1, classifiers predictions for each original instance are used to compose a new dataset that is submitted to another algorithm for training the meta-classifier. Final prediction is determined from the combination of knowledge learned by the base classifiers.

The feature vectors for each training set are made up of all data fields and the class label. The following algorithms are used at level-0 of the stacking method: Multilayer Perceptron (MLP) [21], a variation of Support Vector Machine (SVM) [22] called SMO [23], Nave Bayes (NB) [24], RIPPER [25], C4.5 [26] and Random Forest (RF) [27]. This choice was motivated mainly because the algorithms are quite heterogeneous, since they are based on distinct particulars. The meta-classifier is trained using any of these classification algorithms combining the knowledge learned by the base classifiers, and it is finally used to get a final prediction. The gain of stacking G is computed by the ratio between accuracy values achieved by the meta-classifier and by the best base classifier.

The effects of diversity on stacking can be analyzed by observing the relationship between diversity measures values and the gain of stacking for multiple datasets. It is expected that the most diverse sets of classifiers will contribute to the quality of stacking. This relationship is quantified using a linear regression function and a regression model tree induced by the algorithm M5 [28] with the gain of stacking G as the target field. The vector of features is composed by the diversity measures previously computed $(d_1, d_2, \ldots, d_n)$. The regression models show how much each measure pitches in with the gain of stacking.

IV. Experimental Evaluation

We have used 54 real classification datasets extracted from UCI[1]. The full list is available in the research group website[2]. The chosen datasets cover several areas of knowledge: business, computer, financial, game, life, physical and social. Many of them were widely cited in the scientific literature and they have sundry objectives. The field data types can be integer, real or categorical. The amount of instances ranges from 187 to 12,960. The number of fields and class labels varies from 4 to 216 and from 2 to 48 respectively. These datasets were deposited in the UCI repository from the year 1987 to 2015.

Furthermore, have used three two-dimensional synthetic datasets with very different spatial distributions that allow us to interpret results and algorithms behavior by visual inspection

[1] archive.ics.uci.edu/ml
[2] ginfo.c3.furg.br

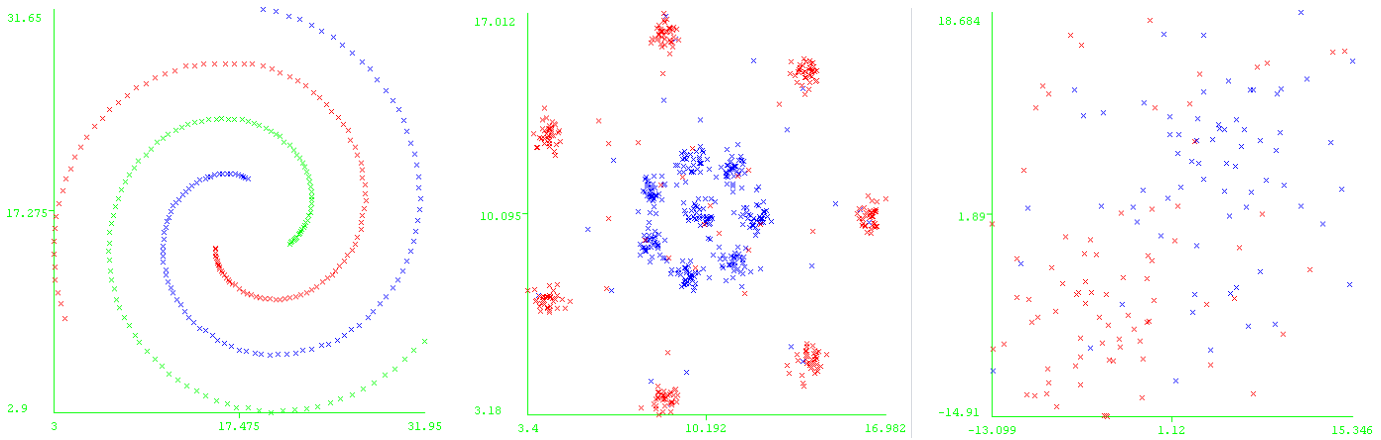

Fig. 2. Datasets Spiral, R15 and D13.

(see Figure 2). Spiral [29] consists of three concentric spirals. R15 [30] is composed by 15 similar Gaussian distributions. We reduced it to 2 classes and included little noise by changing the label of some instances manually selected. D13 [31] presents two Gaussians randomly parameterized and with a lot of noise (half of the instances randomly selected). The number of instances varies from 150 to 600.

The experiments were performed using the data mining tool Weka [32]. The algorithms were parameterized to improve the accuracy or using default values.

A. Diversity and stacking results

Experimental results are summarized in Table I and II that show for each dataset the following information: the computed diversity measures double fault df, disagreement Dis, statistic Q, correlation coefficient ρ, interrater agreement k, Kohavi-Wolpert variance KW and entropy E; the algorithm used to learn the best base classifier (L_0) and its accuracy in percentage (A_{L_0}); the algorithm used to learn the best meta-classifier (L_1) and its accuracy in percentage (A_{L_1}); and the gain of stacking (G), used to sort the results, also in percentage. Values of df, Dis, Q and ρ are averages of the computed values for each pair of base classifiers.

Table I presents experiments with real datasets that reached the worst and best values of G, i.e. we have omitted the results for any dataset where the gain of stacking ranges between -1 and 1%. Table II covers the synthetic datasets.

Observing Table I, we notice that stacking worked well only for 8 out of 54 datasets, where the gain in accuracy ranged from 1.2 to 5.1% (lines 1-8). The best gain of stacking was reached by Balance Scale dataset, in an already accurate result (90.7%) which is difficult to improve. The most frequent algorithm that reaches the best accuracy for level-0 was MLP ranging $26.6 \leq A_{L_0} \leq 90.7$, followed by RF with $84.8 \leq A_{L_0} \leq 92.9$. At level-1, the best meta-classifiers were trained with SMO $(26.9 \leq A_{L_1} \leq 94.9)$ and RF $(83.7 \leq A_{L_1} \leq 95.4)$.

However, stacking decreased the classification quality for some datasets (lines 9-17) reaching in the worst case $G = -9.4\%$. The most frequent algorithms with best accuracy were RF (L_0) and SMO (L_1). For some datasets, more than one classifier used at level-1 returned the same result. For instance, SMO and MLP reaches equal values $(A_{L_1} = 78.8\%)$ for Leaf dataset (line 9).

We have considered good values of diversity those that were sufficiently larger or smaller than the average for all 54 datasets, taking into account whether the measure is directly or inversely proportional to the diversity among the classifiers. These good values are in bold. A general analysis of them indicates that there is more classification diversity for the experiments in which there was gain of stacking (lines 1-8) than for those in which there was loss of quality (lines 9-17).

Abalone dataset (line 8) had the best value of double fault $(df = 0.09)$ due to the low accuracy $(A_{L_0} = 26.6\%)$ presented by the base classifiers. Many of them fail together because these is a multi-classification problem involving 28 distinct class labels. For datasets Connect. Bench (S,M vs. R), Statlog (Vehicle Silh.) and Diabetic Retinopat. Debrec. (lines 2-4) df values were less significant, compared to Abalone, but good in relation to the average. These df values are related to the better accuracy presented by the base classifiers and to the lower number of classes. We observe that for these datasets, all the measures of diversity return good values, collaborating with the hypothesis that the greater the diversity, the greater the stacking accuracy.

However, the experiment involving Low Resolution Spectrometer dataset (line 16) revealed the opposed behavior where the gain of stacking was negative $(G = -5.6\%)$, i.e. the quality of classification decreased considerably, even with high values for all measures of diversity. These high values are returned because there are 531 instances distributed in 48 classes, making even hard the agreement of many classifiers. Balance Scale dataset (line 1) is another counterexample in which there was no diversity among classifiers, but the stacking has reached the best G among all the performed experiments. Furthermore, diversity was considered non-existent or low in 33 out of 37 datasets where the gain of stacking ranged from -1 to 1%.

TABLE I
DIVERSITY MEASURES AND STACKING RESULTS FOR REAL DATASETS.

	Dataset	$df \downarrow$	$Dis \uparrow$	$Q \downarrow$	$\rho \downarrow$	$k \downarrow$	$KW \uparrow$	$E \uparrow$	L_0	A_{L_0}	L_1	A_{L_1}	G
1	Balance Scale	0.78	0.13	0.87	0.50	0.48	0.06	0.17	MLP	90.7	RF	95.4	5.1
2	Connect. Bench (S,M vs. R)	**0.62**	**0.27**	**0.58**	**0.28**	**0.26**	**0.11**	**0.35**	MLP	82.2	Jrip	85.6	4.1
3	Statlog (Vehicle Silh.)	**0.55**	**0.30**	**0.64**	**0.32**	**0.28**	**0.13**	**0.37**	MLP	81.7	RF	83.7	2.5
4	Diabetic Retinopat. Debrec.	**0.49**	**0.32**	**0.56**	**0.29**	**0.29**	**0.13**	**0.41**	MLP	72.0	SMO	73.8	2.4
5	Ionosphere	0.83	0.12	0.83	0.41	0.38	0.05	0.13	RF	92.9	SMO	94.9	2.2
6	Contrac. Method Choice	**0.38**	**0.29**	**0.71**	0.42	0.42	**0.12**	**0.36**	MLP	54.2	SMO	55.3	2.0
7	Vertebral Column	0.72	0.19	0.74	0.39	0.38	0.08	0.23	RF	84.8	RF	86.1	1.5
8	Abalone	**0.09**	**0.28**	**0.47**	**0.21**	**0.21**	**0.12**	**0.36**	MLP	26.6	SMO	26.9	1.2
9	Leaf	**0.52**	**0.30**	**0.70**	0.37	**0.33**	**0.12**	**0.37**	MLP	79.7	SMO$^{\#}$	78.8	-1.1
10	Glass Identification	**0.49**	**0.32**	**0.61**	**0.33**	**0.30**	**0.13**	**0.40**	RF	79.9	RF	79.0	-1.2
11	Credit Approval	0.77	0.14	0.85	0.50	0.48	0.06	0.17	RF	86.7	NB	85.7	-1.2
12	Ecoli	0.79	0.11	0.92	0.57	0.57	0.05	0.14	RF	87.2	RF	86.0	-1.4
13	Solar Flare	**0.62**	0.15	0.91	0.64	0.64	0.06	0.18	J48	72.1	SMO	70.9	-1.7
14	Dresses Attribute Sales	**0.46**	**0.26**	0.77	0.47	0.47	**0.11**	**0.32**	JRip	63.0	NB	60.2	-4.4
15	Audiology (Std.)	0.72	0.13	0.94	0.64	0.63	0.05	0.16	MLP	83.2	SMO	79.2	-4.8
16	Low Resolution Spectrom.	**0.18**	**0.36**	**0.47**	**0.25**	**0.23**	**0.15**	**0.46**	RF	54.0	SMO	51.0	-5.6
17	Primary Tumor	**0.33**	0.19	0.89	0.61	0.60	0.08	**0.25**	NB	50.1	RF	45.4	-9.4
	Average (all 54 datasets)	0.74	0.15	0.76	0.41	0.39	0.06	0.19					

$^{\#}$ MLP reaches equal results

Table II shows diversity and stacking results for synthetic datasets. Using Weka's default values to parametrize the classification algorithms (lines 1-3), R15 and Spiral present gain of stacking, but only classifiers applied to Spiral dataset were high diverse (values in bold). Manually configuring the parameters to reach best accuracy (lines 4-6), stacking decreased the classification quality for all 3 datasets, even with high diversity among classifiers.

To check whether the gain of stacking are in fact statistically significant, we performed a paired Student's T-test [33] comparing the accuracy values of the best meta-classifier (A_{L_1}) with those of the best base classifier (A_{L_0}). We used the T-test because it evaluates whether the means of two normal distributions of values are statistically different, showing good results even when the distributions are not perfectly normal.

Table III shows the gain of stacking and the values of 1-tailed p for each dataset where $p < 0.05$, i.e. when the stacking was statistically superior or inferior than the best base classifier. The number of observations (Obs.) was set to the amount of instances. The statistical difference occurred in 18 of the 54 experiments considering real datasets and only in two involving the synthetic ones. The gain of stacking for the other datasets was considered a technical tie. Datasets in which there was high diversity of the classifiers are highlighted in bold.

Experiments including datasets Connect. Bench (S,M vs. R), Abalone and Turkiye Student Eval. present good values for all diversity measures. Although the diversity measures applied to Spiral dataset have also returned great values, regardless of the parameter configuration, the gain of stacking was not considered significant because the best classifier achieves accuracy equal to 98.7%, which is very difficult to improve. For most experiments there seems to be no relationship between the classifiers diversity and the significant gain in stacking accuracy.

B. Diversity effects

We used regression models to quantify the effects of diversity on the gain of stacking applied to real datasets. We varied the learning algorithm (linear or M5) and the training set: 17 datasets with best and worst (b/w) G, where $-1 \geq G \geq 1$, present in Table I, 18 real datasets where the gain of stacking passed T-test present in Table III and considering all 54 real datasets. The models were evaluated using correlation coefficient and root relative squared error (RRSE).

For the training set composed by the 17 datasets with best and worst values of G, the minimum number of instances to allow at a leaf node in M5 algorithm ranged from 2 to 4, however the result was the same tree with only one node containing the model described by Equation 1. We notice that df had a positive effect on the gain but the influence of ρ was negative. Other diversity measures were irrelevant in estimating the gain. The correlation with the gain of stacking was 0.5243 and the RRSE was 79.67%.

$$G = 0.1278\ df - 0.2189\ \rho + 0.0168 \tag{1}$$

Considering only the 18 datasets that passed T-test, gain of stacking was affected only by ρ (Equation 2), which correlation was 0.3532 and RRSE equal to 99.89%.

$$G = -0.0847\ \rho + 0.0362 \tag{2}$$

Equation 3 shows the linear model trained with all 54 datasets. For this model, only df and KW had effect on the gain of stacking. ρ and other measures were not used. Correlation and RRSE was 0.4081 and 91.58% respectively.

$$G = 0.0971\ df + 0.3757\ KW - 0.0957 \tag{3}$$

TABLE II
DIVERSITY MEASURES AND STACKING RESULTS FOR SYNTHETIC DATASETS.

	Datasets	Parameters	$df \downarrow$	$Dis \uparrow$	$Q \downarrow$	$\rho \downarrow$	$k \downarrow$	$KW \uparrow$	$E \uparrow$	L_0	A_{L_0}	L_1	A_{L_1}	G
1	Spiral		**0.44**	**0.43**	**0.52**	**0.21**	**0.05**	**0.18**	**0.63**	RF	98.7	RF$^{\#}$	99.3	0.6
2	R15	default	0.76	0.19	**0.66**	0.48	**0.23**	0.08	0.21	J48*	93.8	SMO	94.5	0.7
3	D13		0.71	0.12	0.94	0.67	0.67	0.05	0.15	JRip	79.2	MLP	78.5	-0.8
4	Spiral'		0.75	**0.23**	**0.47**	**0.11**	**0.02**	**0.10**	**0.26**	RF	99.4	SMO$^{\#}$	99.0	-0.3
5	R15'	manually configured	0.93	0.02	1.00	0.87	0.87	0.01	0.02	J48	94.0	JRip^{*+}	93.7	-0.4
6	D13'		0.72	0.10	0.95	0.71	0.70	0.04	0.13	JRip^{+}	79.2	SMO	76.5	-3.4

$^{\#}$ MLP reaches equal results * NB reaches equal results $^{+}$ J48 reaches equal results

TABLE III
STATISTICAL TEST APPLIED TO THE GAIN OF STACKING.

	Set	Dataset	Obs.	p	G
1	real	**Connect. Bench (S,Mvs.R)**	208	0.008	4.1
2		Ionosphere	351	0.008	2.2
3		Vertebral Column	310	0.045	1.5
4		**Abalone**	4177	0.001	1.2
5		Mammographic Mass	961	0.025	0.6
6		Soybean	683	0.045	0.6
7		Tic-Tac-Toe Endgame	958	0.025	0.5
8		QSAR biodegradation	1055	0.045	0.4
9		Wilt	4839	0.001	0.3
10		Spambase	4601	0.014	0.1
11		Chess (KR vs. K-P)	3196	0.046	0.1
12		**Turkiye Student Eval.**	5819	0.025	0.1
13		Phishing Websites	11,055	0.003	0.1
14		Credit Approval	690	0.008	-1.2
15		Ecoli	366	0.045	-1.4
16		Solar Flare	323	0.045	-1.7
17		Dresses Attribute Sales	501	0.001	-4.4
18		Audiology	226	0.003	-4.8
19	synthetic	R15	600	0.005	0.7
20		D13'	150	0.045	-3.4

Table IV summarizes the best results comparing the evaluation of linear regression and model trees. We notice that the correlation between the diversity measures and the gain of stacking was weak or moderate for all induced models. The training set composed by 17 w/b G reached the best correlation and RRSE. However, in some cases where the gain of stacking was high, it was not statistically higher than the best base classifier.

In order to better understand the effects of diversity on the performed experiments, for each dataset where the gain of stacking passed T-test, i.e. where the stacking accuracy was statistically different from the best individual base classifier result, we plotted the values of the diversity measures used in the induced models. Figure 3 shows the relation between df and the gain of stacking. High values of both diversity and gain are plotted in green while low values are plotted in red. These points support the hypothesis that the greater the diversity, the greater the stacking accuracy. Other points plotted in blue go against this hypothesis. Similarly, Figures 4 and 5 present the relationship among KW, ρ and G.

Analyzing these figures we notice that only a few points are good values of diversity. Many of them are located near the center of abscissa axis, as well as the vast majority of low diversity values. The only exception where diversity are strong related to the increase of accuracy occurred with the

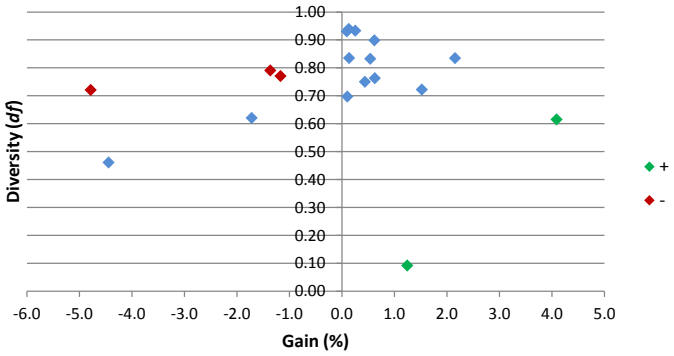

Fig. 3. Relation between double-fault and the gain of stacking.

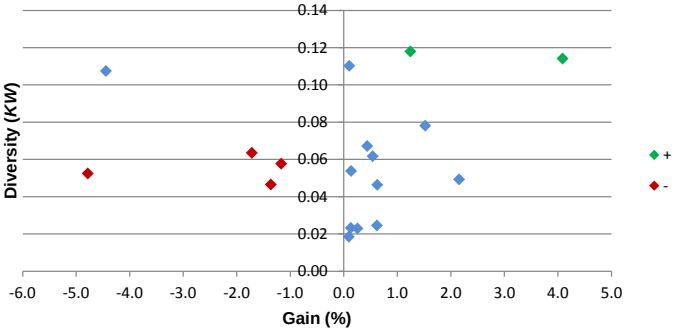

Fig. 4. Relation between Kohavi-Wolpert variance and the gain of stacking.

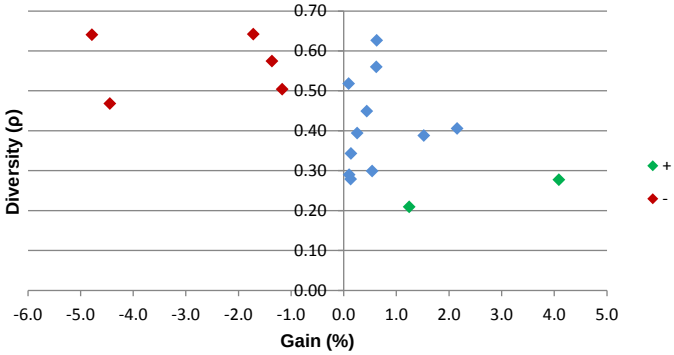

Fig. 5. Relation between correlation coefficient and the gain of stacking.

TABLE IV
EVALUATION OF THE REGRESSION MODELS.

	Datasets	Model	Correlation	RRSE (%)
1	17 b/w G	M5	0.5243	79.67
2	all 54	linear	0.4081	91.58
3	18 T-test	linear	0.3532	99.89

dataset Connect. Bench (S,M vs. R) which experiment reaches $df = 0.62$, $KW = 0.11$, $\rho = 0.28$ and $G = 4.1\%$.

V. CONCLUSION

This paper presented an analysis of the impact of diversity on stacking multiple classifiers. The experiments we have performed show some link between the studied diversity measures and the gain of stacking considering a lot of datasets.

The regression models revealed connections between some measures and the quality of stacking. df, KW and ρ are related to the final classification accuracy, but low values of the correlation coefficients and high values of RRSE imply a weak relationship. So, as suggested by the literature for bagging and majority voting ensembles, predicting the improvement on the best base classifier accuracy using diversity measures is inappropriate.

As future work, we intend to conduct experiments with additional diversity measures and with more synthetic datasets, aiming to better understand the relations between data distribution, decision boundaries, classifiers diversity and the quality of stacking.

REFERENCES

[1] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.

[2] K. M. Ting and I. H. Witten, "Issues in stacked generalization," *Journal of Artificial Intelligence Research*, vol. 10, pp. 271–289, 1999.

[3] S. Dzeroski and B. Zenko, "Is combining classifiers with stacking better than selecting the best one?" *Machine Learning*, vol. 54, no. 3, pp. 255–273, 2004.

[4] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[5] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.

[6] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine learning*, vol. 51, no. 2, pp. 181–207, 2003.

[7] C. J. Merz, "Using correspondence analysis to combine classifiers," *Machine Learning*, vol. 36, no. 1-2, pp. 33–58, 1999.

[8] S. B. Kotsiantis and P. E. Pintelas, "A hybrid decision support tool-using ensemble of classifiers," in *Proceedings of the International Conference On Enterprise Information Systems*, 2004, pp. 448–453.

[9] S. Ali and A. Majid, "Can-evo-ens," *Journal of Biomedical Informatics*, vol. 54, no. C, pp. 256–269, 2015.

[10] L. Peppoloni, M. Satler, E. Luchetti, C. A. Avizzano, and P. Tripicchio, "Stacked generalization for scene analysis and object recognition," in *Proceedings of the IEEE International Conference on Intelligent Engineering Systems*. IEEE, 2014, pp. 215–220.

[11] N. Larios, J. Lin, M. Zhang, D. Lytle, A. Moldenke, L. Shapiro, and T. Dietterich, "Stacked spatial-pyramid kernel: An object-class recognition method to combine scores from random trees," in *Proceedings of the IEEE Workshop on Applications of Computer Vision*. IEEE, 2011, pp. 329–335.

[12] R. Ebrahimpour, N. Sadeghnejad, A. Amiri, and A. Moshtagh, "Low resolution face recognition using combination of diverse classifiers," in *Proceedings of the International Conference of Soft Computing and Pattern Recognition*. IEEE, 2010, pp. 265–268.

[13] S. R. Ness, A. Theocharis, G. Tzanetakis, and L. G. Martins, "Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2009, pp. 705–708.

[14] C. A. Shipp and L. I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," *Information Fusion*, vol. 3, no. 2, pp. 135 – 148, 2002.

[15] R. Dymitr and G. Bogdan, "Classifier selection for majority voting," *Information Fusion*, vol. 6, no. 1, pp. 63 – 81, 2005.

[16] B. Sluban and N. Lavrac, "Relating ensemble diversity and performance: A study in class noise detection," *Neurocomputing*, vol. 160, pp. 120 – 131, 2015.

[17] A. T. Muhammad and S. Jim, "Creating diverse nearest-neighbour ensembles using simultaneous metaheuristic feature selection," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1470–1480, 2010.

[18] M. Makhtar, L. Yang, D. Neagu, and M. Ridley, "Optimisation of classifier ensemble for predictive toxicology applications," in *14th International Conference on Computer Modelling and Simulation*, March 2012, pp. 236–241.

[19] S. Whalen and G. Pandey, "A comparative analysis of ensemble classifiers: Case studies in genomics," in *2013 IEEE 13th International Conference on Data Mining*, Dec 2013, pp. 807–816.

[20] F. A. Faria, J. A. dos Santos, A. Rocha, and R. da S. Torres, "A framework for selection and fusion of pattern classifiers in multimedia recognition," *Pattern Recognition Letters*, vol. 39, pp. 52 – 64, 2014, advances in Pattern Recognition and Computer Vision.

[21] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, USA: Prentice-Hall, Inc., 2007.

[22] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Annual Workshop on Computational Learning Theory*. ACM, 1992, pp. 144–152.

[23] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Proceedings of the Advances in Large Margin Classifiers*. MIT Press, 1999.

[24] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.

[25] W. W. Cohen, "Fast effective rule induction," in *Proceedings of the International Conference on Machine Learning*, 1995, pp. 115–123.

[26] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, USA: Morgan Kaufmann Publishers Inc., 1993.

[27] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] J. R. Quinlan, "Learning with continuous classes," in *Proceedings of the Australian Joint Conference on Artificial Intelligence*, vol. 92. Singapore: World Scientific, 1992, pp. 343–348.

[29] H. Chang and D.-Y. Yeung, "Robust path-based spectral clustering," *Pattern Recognition*, vol. 41, no. 1, pp. 191 – 203, 2008.

[30] C. J. Veenman, M. J. T. Reinders, and E. Backer, "A maximum variance cluster algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1273–1280, Sep 2002.

[31] C. Tomasini, L. Emmendorfer, E. N. Borges, and K. Machado, "A methodology for selecting the most suitable cluster validation internal indices," in *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2016, pp. 901–903.

[32] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

[33] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908.