Assessing the performance of bipolar classifiers in three-class problems

1st Guillermo Villarino
Facultad de Estudios Estadísticos
Universidad Complutense de Madrid
Madrid, Spain
gvillari@ucm.es

2nd Daniel Gómez
Facultad de Estudios Estadísticos
Universidad Complutense de Madrid
Madrid, Spain
dagomez@estad.ucm.es

3rd J. Tinguaro Rodríguez
Facultad de Ciencias Matemáticas
Universidad Complutense de Madrid
Madrid, Spain
jtrodrig@mat.ucm.es

Abstract

In the context of supervised classification, several aspects already exist which need to be improved regarding the decision making step that any classifier passes through. Before providing the final assignment, many classification algorithms produce a soft score (either a probability, a fuzzy degree, a possibility, a cost, etc.) assessing the strength of the association between each item to be classified and each class. Usually, the final decision or classification step of these algorithms consists on assigning the item to the class with the highest soft score, a method typically known as the *maximum rule*. However, this procedure does not always take advantage of all the information contained in such soft scores. In other words, the final classification step of many algorithms may be improved through alternative procedures more sensible to the available soft information that the mentioned maximum rule.

To this aim, in this paper we propose a general bipolar approach that enables learning how to take advantage of the soft information provided by many classification algorithms in order to enhance the generalization power and accuracy of the classifiers. To show the suitability of the proposed approach, we also present some computational experiences for three-class classification problems, in which its application to some well-known classifiers as random forest and neural networks produce some improvements in performance.

***Index Terms*—Supervised classification models, bipolar models, Machine learning, Soft information**

I. Introduction

One of the most important topics in data science is classification, and particularly supervised classification tasks. In the literature, there exist a huge diversity of supervised classification algorithms, approaches and applications, depending on the specific tasks, type of data, characteristics or efficiency [7], [8]. Typically, in a supervised classification context the main aim is to be able to classify a set of items into classes based on a training sample or dataset that provides examples of association between items and classes, and that is used to train the classifiers in order to adequately generalize the observed associations, that is, to fit the classification models to the observed data.

Following the ideas presented in [12]–[15], in [17] classical supervised algorithms as CART [2], Random Forest (RF) [3] and Neural Networks [11], [16] were modelled as probabilistic classifiers, providing soft probabilistic assessments of the association of items with classes. In a second step, a bipolar probabilistic representation framework was developed by allowing some opposition or dissimilarity relationships between the classes to be introduced. In a third step, the more convenient structure of dissimilarity relationships was learned through an evolutionary algorithm. This more expressive representational model and the associated learning process permitted to improve the classification performance of the original classifiers in a binary classification context. In this paper we extend these results by addressing three-class classification problems instead of binary ones.

Moreover, in [18] we proposed a replication + aggregation scheme to obtain a fuzzy classifier from a probabilistic one as a robustness enhancing pre-process that permits developing a fuzzy bipolar model from any soft classification algorithm. The experimental results were also carried out in a binary classification context.

The remainder of the paper is organized as follows: Section II describes the preliminary concepts we will use along the work, including the differences between crisp and probabilistic classifiers, as well as some specific concepts regarding accuracy measures and Genetic Algorithms (GAs). Then, in Section III, we present the main idea of bipolar knowledge representation and the complete two-stage (learning and aggregating) process for constructing a bipolar classifier from a soft supervised one. Finally, the experimental framework along with the respective analysis of the results are presented in Sections IV and V. We summarize the paper with the main concluding remarks in Section VI.

II. Preliminaries

In this preliminary section, we introduce some concepts for a better understanding of the paper. We firstly introduce the main concepts of crisp and probabilistic classifiers as well as their differences and relationships to motivate one of the principal contribution of this paper: the importance of modelling the soft information of a classifier before making the final decision in a classification task.

A. Crisp and probabilistic classifiers

Let us denote by $\{C_1, \ldots, C_k\}$ the set classes of a classification problem, and by $X = \{x_1, \ldots, x_n\}$ the set of items that

has to be classified.

As we have pointed in the introduction, many classification users only takes into account the final output of the classification task. This is probably because they are only interested in the final solution provided by the classifier. This is the reason why in a general way, the classifiers are usually viewed as functions

$$C : X \longrightarrow \{C_1, \ldots, C_k\}, \tag{1}$$

that is, a procedure to assign one of the available classes to each of the items being classified.

Nevertheless, the classification process goes through many steps before to arrive to the final assignment, and it is in the intermediate steps that soft information usually appear as a natural way to model the information and the evidence being obtained. Particularly, it is very common that classification algorithms manage soft information for each item $x \in X$ about the probability that x belongs to each of the different classes, or in fuzzy classification models about the degree of membership of the item x in the set of classes.

Taking into account these considerations, in [17] we distinguished between crisp (classical) and probabilistic classifiers. A probabilistic classifier can be viewed as a function

$$C_P : X \longrightarrow [0,1]^k, \tag{2}$$

that assigns to each item x its probability of belonging to each of the available classes. Obviously, for any $x \in X$ it has to hold that $\sum_{i=1}^{k} (C_P(x))_i = 1$ because of the additivity of probability. We would like to remark that many classification algorithms (as for example neural networks, random forest or decision trees) could be viewed as probabilistic classifiers if we just look at the soft information provided by the algorithms before making the final decision or crisp assignment.

III. Probabilistic bipolar model

This section is devoted to present the underlying ideas of bipolar knowledge representation. Firstly, it merits to be stressed that the concept of dissimilarity assumes that the available classes are related through a certain opposition or dissimilarity structure informing of which classes provide negative evidence against the others. This dissimilar structure can be modelled through a dissimilarity matrix D, which contains the degree of dissimilarity for any pair of classes. Obviously, the main diagonal has to be composed by zero values.

It is clear that the dissimilarity matrix D plays a crucial role in this classification scheme since it determines how the negative evidence is derived from the initial evidence for each class. As a consequence, the performance of the resulting crisp classifier, as well as the effect of incorporating the bipolar representation framework and the aggregation method, are absolutely dependent on the choice of the matrix D.

Figure 3 shows a flow diagram of the proposed decision making stage, including the genetic search of the dissimilarity structure and its application to the test set.

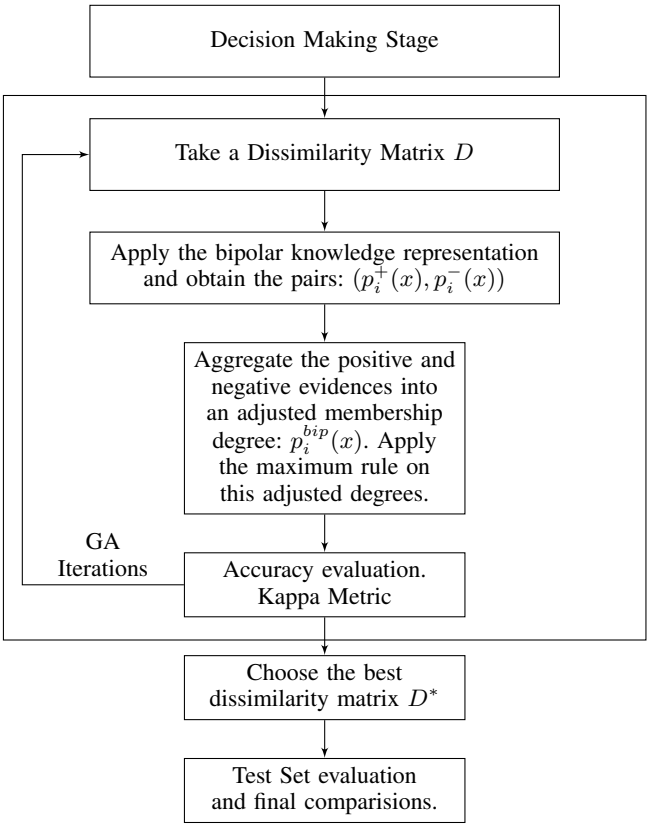

Fig. 1. Flow diagram of the proposed Decision Making Stage.

A. Learning the dissimilarity matrix

Ideally, in real situations the adequate structure of dissimilarity between classes should be proposed by application domain experts based on his knowledge. However, in many cases it may be more practical to obtain the matrix D through a learning process carried out once the base soft classifier has been trained. When this learning process is driven by a measure of performance focused on the generalization accuracy of the adjusted crisp classifier, the resulting matrix tends to fix some of the misassignments committed by the base classifier on the training sample, hopefully also improving its predictive accuracy on new queries or a test sample. Therefore, this learning approach allows that any probabilistic classifier may benefit from introducing a dissimilarity structure in the set of classes, aiding the decision rule of the classifiers to better adapt to the specific features of each dataset or application context.

Here we propose that the learning process of the dissimilarity matrix D is performed by means of a genetic algorithm (GA). The specific parameters of the applied GA are given in Section IV-C.

B. Obtaining the paired structure (p^+, p^-)

In this section we show the application of the dissimilarity matrix already learned by the GA to obtain the paired structure containing the positive and negative evidences.

To do so, we depart from the soft information (estimated probabilities) given by the base algorithm for an item x, $p_i(x) = p_i^+(x)$, treating it as our positive probability of class C_i membership. Then, we apply the bipolar knowledge representation approach to get the negative evidence in the following way:

$$p_i^-(x) = \sum_{j \neq i} d_{ij}\ p_j^+(x) = \sum_{j=1}^{k} d_{ij}\ p_j^+(x) = D_i p^+(x), \quad (3)$$

Once the bipolar paired structure has been obtained, one of the possibilities we have is to aggregate this positive and negative evidences into a bipolar adjusted degree of evidence by applying any kind of aggregation operator.

Let us stress this is only one among the wide spectrum of possibilities for dealing with paired structures.

C. Aggregating bipolar evidence: the additive and logistic cases

Let us now address the question of how to aggregate, for a given class C_i and an item x, the pair of positive and negative evidence degrees $p_i^+(x)$ and $p_i^-(x)$ in order to obtain a single adjusted degree $p_i^{adj}(x)$. Obviously, different aggregation choices will lead to different adjusted classifiers. In this work we have studied two different kinds of aggregation, that are defined below.

Let $p_i^+(x)$, $p_i^-(x)$ be the positive and negative probabilities of item x into class C_i. The additive adjusted degree of x into class C_i is defined as

$$p_i^{add}(x) = \max\{0, p_i^+(x) - p_i^-(x)\}. \quad (4)$$

Notice that the previous definition can be interpreted as the Lukasiewicz t-norm $W(a,b) = \max\{a + b - 1, 0\}$ of the positive and non-negative degrees, that is, $p_i^{add}(x) = W(p_i^+(x), n(p_i^-(x)))$, where n stands for the standard negation $n(a) = 1 - a$. In this way, the positive evidence $p_i^+(x)$ initially provided by the soft classifier is adjusted by subtracting from it the negative evidence $p_i^-(x)$. Particularly, the initial degrees are not modified when no class is dissimilar to C_i, that is, when $D_i = 0$.

Thus, an adjusted degree $p_i^{add}(x) > 0$ represents the existence of a positive gap between the support for class C_i and the support for class dC_i, that is, for the classes considered dissimilar to C_i. In this situation, the strength of the association of item x with class C_i may have been reduced from its initial assessment, but it is still perfectly possible that item x is finally assigned to C_i. On the other hand, a zero value of $p_i^{add}(x)$ represents a situation in which there exist more evidence for the dissimilar class dC_i than for C_i, and thus the adjusted classifier should not assign the item to class C_i.

In the following definition, we propose an alternative way to aggregate the positive and negative information into a single adjusted degree.

Let $p_i^+(x)$, $p_i^-(x)$ be the positive and negative evidence degrees of item x into class C_i. The logistic adjusted membership degree of x into class C_i is defined as

$$p_i^{log}(x) = \begin{cases} 1 - e^{-\frac{p_i^+(x)}{p_i^-(x)}} & if\ p_i^-(x) > 0 \\ 1 & otherwise \end{cases} \quad (5)$$

Unlike the additive logic of the previous aggregation, this logistic aggregation focuses on the ratio between positive and negative information, adjusting it to range in the [0,1] interval through a logistic transformation. This permits a somehow more flexible behaviour of the adjusted degrees, in the sense that the choice of the dissimilarity matrix D may have an even greater influence in the adjustment of the initial positive evidence provided by the base soft classifier, up to the point that class $p_i^{log}(x) = 1$ whenever no evidence is gathered for the dissimilar class dC_i, that is, when $p_i^-(x) = 0$.

As mentioned above, once one of these two aggregation methods has been applied and the adjusted degrees $p_i^{adj}(x)$ has been obtained for each class (either $p_i^{adj}(x) = p_i^{add}(x)$ or $p_i^{adj}(x) = p_i^{log}(x)$), the final decision on the classification of item x is made by applying the maximum rule to such adjusted degrees. Therefore, the item x is finally assigned to the class C_h with a maximum adjusted degree $p_h^{adj}(x)$, that is, $h = arg\max_{i \in \{1,\ldots,k\}} p_i^{adj}(x)$.

IV. Experimental framework

This section is devoted to present a computational experience aimed to assess the performance of our dissimilarity - based bipolar knowledge representation approaches (additive and logistic) when applied on recognized classifiers such as Random Forest [3] and Neural Networks [11], [16].

A. Experimental setting details

As just mentioned, the base classifiers used in this experiment are Random Forest (RF) and Neural Networks (NNet). This experience is designed to compare the benchmark performance of these classifiers with those obtained from the later ones by means of the proposed dissimilarity learning process and the additive and logistic adjustments.

The results for each classifier in each experiment will be obtained following a 5-fold cross validation scheme for each dataset. In each folder, that is, for each training set, the optimal base classifier parametric configuration is approximated using a grid P on the space of parameters of the algorithms considered. In order to evaluate the performance of each specific parametric configuration $p \in P$, 25 bootstrap samples of the training set are generated, in such a way that the base classifiers are fit to each of these bootstrap samples and then tested on a hold-out sample (composed by the non selected instances in the bootstrapping process) using the kappa statistic.

At each folder, the genetic dissimilarity learning process is carried out departing from the vectors of estimated probabilities $p(x)$ of the items x in the training sample in the way shown in III.

The train and test performance measures of each of the 3 classifiers in each dataset considered in each experiment are

Id.	Data-set	#Ex.	#Atts.	(R/I/N)
Aut	Autos	159	25	(15/0/10)
Car	Car	159	25	(15/0/10)
Wnq	Winequality-red	1599	11	(11/0/0)
Pen	Penbased	10992	16	(0/16/0)
Pag	Page-blocks	5472	10	(4/6/0)
Der	Dermathology	366	34	(0/34/0)
Eco	ecoli	336	7	(7/0/0)
Fla	flare	1066	25	(15/0/10)
Gla	Glass	214	9	(9/0/0)
Shu	Shuttle	2175	9	(0/9/10)
Yea	Yeast	1484	8	(8/0/0)
Lin	Lymphography	148	18	(3/0/15)
Bal	Balance	625	4	(4/0/0)
Win	Wine	178	13	(13/0/0)
Nty	Newthyroid	215	5	(4/1/0)
Hay	Hayes-Roth	160	4	(0/4/0)
Con	Contraceptive	1473	9	(6/0/3)
Thy	Thyroid	720	21	(6/0/15)

TABLE I
SUMMARY DESCRIPTION FOR THE EMPLOYED DATASETS.

finally computed by respectively averaging the train and test accuracy rates of the $F = 5$ different folders.

B. Data sets

We have selected a benchmark of 18 datasets from the KEEL dataset repository [1]. Particularly, we have used the 5-folder cross-validation datasets provided by KEEL in the different experiments. Table I summarizes the properties of the selected datasets, showing for each dataset the number of examples (#Ex.), the number of attributes (#Atts.) and type (Real/Integer/Natural) To transform multi-class datasets into trhee-class ones, we have taken as class C_0 and C_1 the originals closest to 20% of instances and as class C_2 the union of the remainder classes.

C. Genetic algorithm details

Finally, regarding the GA used at the evolutionary tuning of the dissimilarity structures, we have used the default GA for real-coded chromosomes implemented in the *genalg* R package. It is a standard GA, with usual crossover and mutation operators, the details of which can be consulted at [20]. The GA has been run with the following configuration, that provided satisfying solutions in a feasible amount of time:

- Population Size: 50 individuals.
- Number of iterations: 20
- Mutation Chance: 0.01.
- Elitism: About 20% of the population size.

Let us note at this point that we have tried a more complex configuration for the GA used in number of iterations, specifically we have used a 40 iterations and 100 individuals with no improvements.

D. Statistical test for performance comparison

In this paper, we use some hypothesis validation techniques in order to give statistical support to the analysis of the results.

Specifically, we employ the Wilcoxon rank test [19] as a non-parametric statistical procedure for making pairwise comparisons between two algorithms. For multiple comparisons, we use the Friedman aligned ranks test, which is recommended in the literature [4], [5] to detect statistical differences among a group of results. Finally, the Holm post-hoc test [6] has been used to find the algorithms that reject the equality hypothesis with respect to a selected control method. A complete description of these tests, with many considerations and recommendations and even the software used to run this analysis can be found on the website http://sci2s.ugr.es/sicidm/.

V. Experimental results

This section is aimed to present the results of the computational experience described above, and carried out in order to study the capacity of enhancement of our bipolar adjusted classifiers with respect to the reference base classifier to which the proposed final decision tuning method is applied.

Results are grouped, for each base algorithm, in pairs for training and test, where the best global result for each considered dataset is stressed in **bold-face**. None is stressed in case of ties.

The experimental study has been obtained using R Software. Specifically, we used the *caret* package [21] for the classifiers training, fitting them through the underlying classical packages *random forest* and *nnet*, and finally the *genalg* package [20] to assess the GA.

For performing all the analysis presented in this paper we have used a computer AMD A10-6700 3.94GHz, 8GB RAM, Windows 8.1.

We can observe from the results of tables II and III the general good behaviour of the bipolar tuning method, at least regarding one of the bipolar adjustment methods, since it allows the improvement in performance of the reference algorithms.

	RF					
	Ref		bipAdd		bipLog	
	Train	Test	Train	Test	Train	Test
Aut	1.000	0.716	1.000	**0.719**	1.000	0.706
Car	0.996	**0.867**	1.000	0.854	1.000	0.857
Wnq	1.000	0.515	1.000	0.489	1.000	**0.525**
Pen	1.000	**0.903**	1.000	0.895	1.000	0.892
Pag	1.000	0.831	1.000	**0.832**	1.000	**0.832**
Der	1.000	**0.995**	1.000	0.993	1.000	0.992
Eco	1.000	0.758	1.000	**0.775**	1.000	0.764
Fla	0.796	0.783	0.805	**0.787**	0.807	0.784
Gla	1.000	0.672	1.000	0.658	1.000	**0.677**
Shu	1.000	**0.996**	1.000	**0.996**	1.000	0.995
Yea	1.000	0.377	1.000	0.366	1.000	**0.378**
Lin	0.981	0.672	0.996	0.675	0.996	**0.710**
Bal	0.612	**0.556**	0.615	0.523	0.617	0.513
Win	1.000	**0.979**	1.000	0.954	1.000	0.973
Nty	1.000	**0.935**	1.000	0.912	1.000	0.895
Hay	0.885	0.703	0.886	**0.715**	0.886	**0.715**
Con	0.788	0.280	0.807	**0.286**	0.807	0.279
Thy	1.000	0.895	1.000	**0.897**	1.000	0.891
Mean	0.948	**0.746**	0.950	0.740	0.951	0.743

TABLE II
RESULTS IN TRAIN AND TEST ACHIEVED BY THE GENETIC BIPOLAR APPROACHES APPLIED TO THE RF ALGORITHM.

Regarding the bipolar method applied to the RF classifier, in Table II we show the results and the following brief description of its behaviour.

- There is no improvement by kappa means when comparing the additive bipolar model against reference.
- The additive bipolar classifier outperforms the classification of the remainder approaches in 8 out of 18 datasets and the logistic one does so in 6 of them.
- Reference wins in 6 out of 18 datasets.
- There is a tie between the additive bipolar approach and the reference in the Shuttle dataset.

Thus we can see that we have reached improvements or ties in 12 out of 18 datasets when comparing. It is important to note the variable behaviour of the additive bipolar method in this case. Despite being the winner method in number of datasets, we can see that its mean is not the best because of the lower kappa value obtained in several of the remainder datasets.

	Nnet					
	Ref		bipAdd		bipLog	
	Train	Test	Train	Test	Train	Test
Aut	0.504	0.382	0.533	**0.385**	0.532	**0.385**
Car	1.000	0.997	1.000	0.997	1.000	0.997
Wnq	0.359	0.341	0.399	**0.356**	0.399	**0.356**
Pen	0.954	0.855	0.964	**0.866**	0.966	0.856
Pag	0.853	0.753	0.874	0.755	0.887	**0.774**
Der	1.000	0.987	1.000	**0.991**	1.000	**0.991**
Eco	0.753	**0.697**	0.779	0.680	0.777	0.688
Fla	0.785	**0.788**	0.794	0.782	0.795	0.777
Gla	0.660	0.507	0.688	**0.517**	0.687	0.513
Shu	0.991	0.976	0.993	**0.977**	0.994	**0.977**
Yea	0.440	0.360	0.473	0.379	0.473	**0.381**
Lin	0.896	0.667	0.922	0.671	0.925	**0.678**
BAl	0.600	0.586	0.603	0.562	0.603	**0.563**
Win	0.945	0.911	0.959	**0.915**	0.960	0.901
Nty	0.986	0.957	0.995	0.957	0.997	0.957
Hay	0.811	**0.615**	0.850	0.600	0.845	0.588
Con	0.356	0.334	0.383	0.336	0.383	**0.338**
Thy	0.859	0.738	0.904	0.770	0.925	**0.803**
Mean	0.764	0.692	0.784	0.694	0.786	**0.696**

TABLE III
RESULTS IN TRAIN AND TEST ACHIEVED BY THE GENETIC BIPOLAR APPROACHES APPLIED TO THE NNET ALGORITHM.

Considering the NNet classifier, the bipolar method reaches the results shown in Table III that could be interpreted as follows:

- There is an improvement by kappa means of 0.004 when comparing the logistic bipolar model against reference, being of 0.002 in case of the additive one.
- Both additive and logistic bipolar classifiers outperform the classification of the remainder approaches in 7 and 10 out of 18 datasets respectively.
- Reference wins in 3 out of 18 datasets.
- There two ties in these results.

On balance we have reached improvements or ties in 14 out of 18 datasets when comparing the bipolar approaches against the reference.

In order to detect significant differences among the results of the different approaches, we carry out the Friedman aligned rank test. This test obtains a low p-value for all the three algorithms, which implies that there are significant differences between the results provided by each method.

For this reason, we can apply a post-hoc test to compare our methodology against the remaining approaches. Specifically, a Holm test is applied using the best approach (the one with lower ranking) as control method and computing the adjusted p-value (APV) for the one with the highest ranking.

Obviously, it would be desirable for the reference to reach the highest or, at least, not the lowest ranking since it is usually associated with worse results.

Algorithm	Rank RF	Rank NNet
"Ref"	22.22	31.5
"BipAdd"	31.83	26.44
"BipLog"	28.44	24.55
p-val	0.00097	0.000913
APV	**0.1336**	**0.371**

TABLE IV
AVERAGE RANKINGS OF THE ALGORITHMS (ALIGNED FRIEDMAN), ASSOCIATED P-VALUES AND HOLM TEST APV FOR EACH ALGORITHM WITH THE MAX AGGREGATION.

Table IV, reflects that there are statistical significant differences between the three classifiers for both RF and NNet algorithms. However, in case of RF this differences and the respective statistical analysis should be carefully interpreted because of the lower ranking value obtained by the reference algorithm. In fact, the reference (RF without applying any bipolar approach) seems to reach the best results regarding the Friedman aligned rank test in spite of not being the best in number of datasets outperformed. Therefore there is no statistical evidence of the superiority of any method compared in case of RF.

Regarding the base Nnet classifier, Table IV shows the superiority of both bipolar approaches in ranking values, however the Holm post-hoc test reflects that there is not enough evidence to ensure that both bipolar approaches outperform the reference.

Comparison	R^+	R^-	p-val
RFbipAdd vs. RFRef	115.0	56.0	0.1913
RFbipLog vs. RFRef	100.0	71.0	0.5135
NNetbipAdd vs. NNetRef	100.0	53.0	0.2559
NNetbipLog vs. NNetRef	95.0	58.0	0.3684

TABLE V
WILCOXON TEST TO COMPARE THE BIPOLAR TUNING APPROACHES (R^+) AGAINST THE BASE CLASSIFIER (R^-).

The statistical analysis of the pairwise comparisons of methods, which is carried out by means of a Wilcoxon test, Table V, reflects the weak superiority of the proposed methodology when it is applied to the RF and Nnet algorithms with not so high p-values in case of additive bipolar model. Again, the application of the methodology on the RF and NNet algorithm does not reach significant improvements.

VI. Discussion and final remarks

In this paper we have studied the extension of probabilistic supervised classifiers into a bipolar knowledge representation framework by means of the introduction of a dissimilarity

structure in the set of classes. These structures enable considering different opposition or dissimilarity relationships between the available classes, that otherwise are by default considered as independent, unrelated objects. These relationships provide further information of the underlying structure of the classification problems being addressed, which can be used at the final decision or classification stage to better exploit the soft scores provided by any classifier to assess the association between each item and each class. Therefore, the introduction of dissimilarity structures may allow to strengthen the adaptation of the classifiers to each specific application context, in which classes acquire a particular semantics, thus also improving the classifier performance.

In this sense, the proposed approach can be understood as a general post processing method to fine tune the maximum decision rule usually applied to make the decision on the class assignment of each item to be classified.

To study the feasibility of the proposed approach, and particularly to remark that it can be applied to any soft classifier despite how it is obtained, we have applied it to two of the most powerful supervised classifiers, random forests and artificial neural networks. A rigorous and extensive computational experience has been conducted to analyse whether the proposed additive and logistic bipolar approaches enabled an statistically significant improvement of the base probabilistic classifiers.

Along this experimental study, we have reached several lessons learned:

- The bipolar framework improved the results of the two base machine learning algorithms considered in this work in number of datasets outperformed.
- Both the additive and logistic adjustment methods did not significantly outperform the results of the base classifier. However, they reached not so high p-values in the Wilcoxon test, specially the additive one.
- Comparing both the additive and the logistic proposed classifiers, we found there is no clear winner. In fact, this question seems to be somehow dependent on the base algorithm considered as well as on the dataset of application.

These results lead us to conclude that the proposed approach provides a suitable solution to confront three-class classification problems and improve the decision rule that manages how the intermediate soft information gathered by many classifiers is exploited.

However, we must improve the results in statistical terms so that we could ensure the superiority of our proposed methodology when applied in three-class classification problems by enlarging the benchmark of datasets, and considering several different parametric configuration for training the base classifier as well as the evolutionary search of the dissimilarity structure among the set of classes.

Regarding future research on this approach, a main line of work will be devoted to study further mechanisms than the additive and logistic aggregations for exploiting the bipolar pairs of positive and negative evidence. A particularly appealing possibility is to use these bipolar pairs as the base information of a multivalued para-consistent logic, as those proposed in [9], [10], [12]. This would allow an even more expressive representational framework to take advantage of all the information contained in the soft scores provided by classifiers.

Acknowledgement

This research has been partially supported by the Government of Spain, grant TIN2015-66471-P and the FPU fellowship grant 2015/06202 from the Ministry of Education of Spain.

References

[1] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, (2011) KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework, *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17:2-3, pp. 255-287.

[2] Breiman L. (1984) Classification and Regression Trees. New York, NY: Kluwer Academic Publishers;

[3] Breiman L. (2001) Random Forests. Mach. Learn.vol.40 5-32.

[4] Demsar J., (2006) Statistical comparisons of classifiers over multiple datasets, J. Mach. Learn. Res. 7 1-30.

[5] García S., Fernandez A., Luengo J. and Herrera F., (2010) Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, Inform. Sciences 180(10) 2044-2064.

[6] Holm S., (1979) A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 65-70.

[7] Kumar, R. and Verma, R. (2012) Classification algorithms for data mining: A survey, International Journal of Innovations in Engineering and Technology, 2 7-14.

[8] Lim, TS., Loh, WY. and Shih, YS. (2000) A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms, Machine Learning, 40 203-228.

[9] Ozturk,M., Tsoukiàs, A. (2007) Modeling uncertain positive and negative reasons in decision aiding. Decis. Support Syst. 43, 1512-1526

[10] Turunen, E., Ozturk, M., Tsoukiàs, A. (2010) Paraconsistent semantics for Pavelka style fuzzy sentential logic. Fuzzy Sets Syst. 161, 1926-1940

[11] Ripley, B. D. (1996) Pattern Recognition and Neural Networks. Cambridge.

[12] Rodríguez, J.T., Turunen, E.,Ruan,D.,Montero, J. (2014) Another paraconsistent algebraic semantics for Lukasiewicz-Pavelka logic. Fuzzy Sets Syst. 242, 132-147

[13] Rodríguez, J. T., Vitoriano, B., Montero, J. (2011) Rule-based classification by means of bipolar criteria. 2011 IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making (MDCM), 197-204.

[14] Rodríguez, J. T., Vitoriano, B., Montero, J. (2012) A general methodology for data-based rule building and its application to natural disaster management. Computers & Operations Research, 39 (4) 863-873.

[15] Rodríguez JT, Vitoriano B, Gómez D, Montero, J. (2013) Classification of Disasters and Emergencies under Bipolar Knowledge Representation. In: Vitoriano B, Montero J and Ruan D (eds), Decision Aid Models for Disaster Management and Emergencies, vol. 7, Atlantis Computational Intelligence Systems, 209-232.

[16] Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.

[17] Villarino, G., Gómez, D., Rodríguez, J. T. (2017). Improving Supervised Classification Algorithms by a Bipolar Knowledge Representation. In Advances in Fuzzy Logic and Technology 2017 (pp. 518-529).

[18] Villarino, G., Gómez, D., Rodríguez, J.T. et al. Soft Comput (2018). https://doi.org/10.1007/s00500-018-3320-9

[19] Wilcoxon F., (1945) Individual comparisons by ranking methods, Biometrics 1 80-83.

[20] Willighagen E., (2005) genalg: R Based Genetic Algorithm. http://cran.r-project.org/

[21] Kuhn M., (2008) Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5) 1-26. doi:http://dx.doi.org/10.18637/jss.v028.i05