Standard Decision Boundary in a support-domain of fuzzy classifier prediction for the task of imbalanced data classification

Pawel Ksieniewicz[0000−0001−9578−8395]

Wroclaw University of Science and Technology
Department of Systems and Computer Networks
pawel.ksieniewicz@pwr.edu.pl

Abstract. Many real classification problems are characterized by a strong disturbance in a prior probability, which for the most of classification algorithms leads to favoring majority classes. The action most often used to deal with this problem is oversampling of the minority class by the SMOTE algorithm. Following work proposes to employ a modification of an individual binary classifier support-domain decision boundary, similar to the fusion of classifier ensembles done by the *Fuzzy Templates* method to deal with imbalanced data classification without introducing any repeated or artificial patterns into the training set. The proposed solution has been tested in computer experiments, which results shows its potential in the *imbalanced data classification*.

Keywords: pattern recognition · classification · imbalanced data · fuzzy classifiers · standard normalization

1 Introduction

The base and the most important element of any *artificial intelligence* application is the decision module, most often being a trained *pattern recognition* model [4]. The development of such a solution requires the use of an algorithm capable of building knowledge around the specific type of training data.

In the case where training samples are only a set of non-described patterns, for example, to gather groups of objects based on *cluster analysis*, we are dealing with the problem of *unsupervised learning*. In most situations, however, we are not interested in identifying groups in the data set. The goal is preferably in assigning new objects, seen for the first time, to classes that we already have known there is a possibility to learn about their properties on the example of existing patterns. This type of learning is called *supervised learning*, and this specific task is *classification* [17].

In real classification problems, it is relatively rare for each class of a training set to be represented evenly. A significant disturbance in the proportions between classes is widely studied in the literature under the name of *imbalanced data classification* [5, 7].

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_8

Solutions for such problems are usually divided into three groups [9]. The first are *built-in methods* that try to modify the algorithm's principles or its decision process to take into consideration the disturbed prior probability of the problem [19, 24]. The second group, which is also the most popular in literature and applications, is based on data preprocessing aiming to balance the class counts in the training set. The most common solutions of this type are *under-* [20] and *oversampling* [18] together with methods for generating synthetic patterns such as SMOTE [22, 21, 6] or ADASYN [25, 1, 8]. The third group consists of *hybrid methods* [23], mainly feasting on achievements of *ensemble learning*, using a pool of diversified base classifiers [13, 12] and a properly constructed, imbalanced decision principle [11, 10].

Following work tries to propose a practical method from the *built-in methods* group of solutions, modifying the support-domain decision boundary of the *fuzzy classifier*. It is done using the knowledge acquired on the basis of support vectors obtained on the training set by the already built model, similarly to the propositions of *Fuzzy templates* [16, 15]. The second section describes how to adapt them to work with a single classification model and how to modify this approach to the proposed *Standard Decision Boundary* algorithm. The third chapter contains the design of computer experiments carried out and summarized in the fourth chapter, and the fifth one focuses on the overall conclusions drawn from the research.

2 Methods

The *feature space* of a problem in which the decision boundary of the classifier is drawn is the most often undertaken area of considering the construction of a classification method. However, its modification may also take place in the space of supports obtained by the model, which is the subject of the method proposed in this article.

*Regular Decision Boundary (*RDB*)* A fitting algorithm of every *fuzzy classifier* does not only provide bare prediction but also calculates the complementary (adding up to one) probability of belonging to each of the problem classes, which constructs the *support vector* of a predicted sample [3]. The classifier's decision, in the most popular approach, is taken in a favor of the class for which the highest support was obtained [14].

By simplifying the classification problem only for binary tasks, one may determine such a decision rule by the most straightforward equation of a straight line:

$$y = x, \tag{1}$$

where the x-axis represents support for the negative class and the y-axis is positive support. For the following paper, this rule will state as *Regular Decision Boundary* (RDB), and it is illustrated in Figure 1.

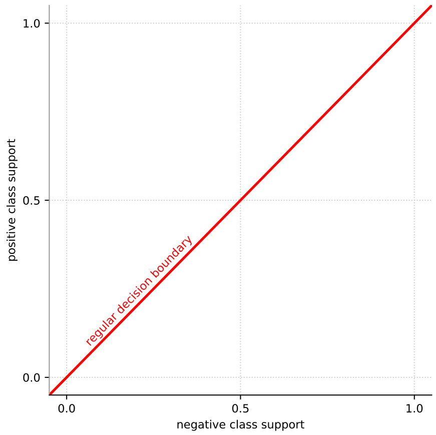


Fig. 1: Illustration of a *Regular Decision Boundary* (RDB).

*Fuzzy Templates Decision Boundary (*FTDB*)* A commonly perceived phenomenon that occurs in classification models build on an imbalanced training set is the general tendency to favor the majority class [5]. The support obtained for it receives a particular bonus, caused directly by the increased prior probability.

One of the possible counteractions to this phenomenon may be the modification of a decision rule in the support domain. Solutions of this type are quite common in the construction of *fusers* for the needs of classifier ensembles [15]. One of such approaches is the proposition of *Fuzzy Templates*, introducing the *Decision Profile*, being the matrix of *support vectors* obtained for all patterns from the training set by each classifier from the available pool [16]. To produce a prediction, algorithm determines class centroids of obtained supports, and the final decision is based on the *Nearest Mean* principle.

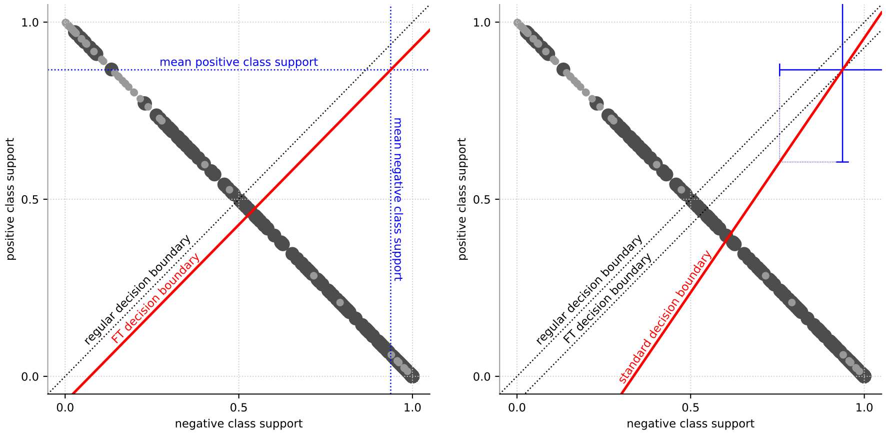
In the case of a single *fuzzy classifier*, in contrast to the ensemble products of *Decision Profiles*, each of the complementary support vectors obtained for the training set, by definition, must be on a diagonal of a support space perpendicular to the *Regular Decision Rule*. An attempt to employ the *Fuzzy Templates* approach in a single classification model may be described by the equation of a straight line parallel to the *Regular Decision Boundary*, but passing through a point determined by the mean support values calculated separately for the patterns of both the training set classes:

$$y = x + \mu_2 - \mu_1, \tag{2}$$

where μ_1 and μ_2 are mean supports of each class. For the purpose of the following paper this rule will state as *Fuzzy Templates Decision Boundary* (FTDB), and its example is illustrated in Figure 2a.

*Standard Decision Boundary (*SDB*)* The *Fuzzy Templates* method, is an additional, simple classifier, supplementing any fuzzy classification algorithm with

DOI: 10.1007/978-3-030-50423-6_8



(a) *Fuzzy Templates Decision Boundary* (b) *Standard Decision Boundary*

Fig. 2: Illustration of trainable decision boundaries.

the model learned from its answers. It is based on the calculation of the basic statistical measure (*mean value*) and its inclusion in the final prediction of the hierarchical ensemble. The following work proposes an enhancement of this approach by including into the decision process also the basic knowledge about the distribution of supports obtained by the base classifier, using a *standard deviation* measure.

This approach still assumes that the decision boundary goes through the intersection of mean supports, but its gradient is further modified. It depends directly on the ratio between standard deviations, so it also goes through the point designated as the difference between the expected values of the distribution and the standard deviations vector. The formula may represent the equation of the proposed decision boundary:

$$y = \frac{\sigma_2(x - \mu_1)}{\sigma_1} + \mu_2, \tag{3}$$

where σ_1 and σ_2 are standard deviations of both classes. Due to the employment of both statistical measures calculated for the needs of a standard normalization, this rule will state as *Standard Decision Boundary* (SDB), and its example is illustrated in Figure 2b.

Supposition Intuition suggests that changes in the prediction method implemented both by the FTDB and SDB models should increase the *precision* of the obtained decisions, although the linear nature of the used decision boundary in a presence of a such tendency must simultaneously lead to a worsening of the results of the *recall* metric. Using aggregate information about class distributions in a decision rule, ignoring the prior probabilities of the training set, may result in an increase in the overall quality of predictions in imbalanced data. So if the proposed method will obtain significantly better results in aggregate metrics, such as *F1-score*, *balanced accuracy score* or *geometric mean score*, it will be considered as promising.

3 Design of experiments

Datasets The problems considered in research during experiments are directly expressed by the selection of datasets that meets specific conditions. For the purposes of conducted experimental evaluation, it was decided to use data with a high degree of imbalance, exceeding the 1:9 ratio, with relatively low dimensionality (up to 20 features). The appropriate collection is contained in the KEEL data repository [2]. A summary of the datasets selected for testing, supplemented with information on the *imbalance ratio*, the count of features and patterns is presented in Table 1.

Compared approaches The basis of considerations taken in this work are the differences between approaches to draw a decision boundary in the *support space* and the effectiveness of this type of solutions in *imbalanced data classification* problems. For the purposes of evaluation, the three methods presented in Section 2 have been supplemented with the preprocessing method, being a *state-of-art* solution for this type of problems. Due to the very large *imbalance ratio*, it is often impossible to apply the SMOTE algorithm (with default parameterization it requires at least 5 minority class examples in the learning set), therefore *random oversampling* was chosen. The full list of compared algorithms presents as follows:

1. **RDB** — *Regular Decision Boundary* used in *Gaussian Naive Bayes* classifier,
2. **ROS-RDB** — *Regular Decision Boundary* used in *Gaussian Naive Bayes* classifier trained on datasets with *randomly oversampled* minority class,
3. **FTDB** — *Fuzzy Templates Decision boundary* used in *Gaussian Naive Bayes* classifier,
4. **SDB** — *Standard Decision boundary* used in *Gaussian Naive Bayes* classifier.

Evaluation methodology and metrics used During the experimental evaluation, a *stratified 5-fold cross validation* was used, for the non-deterministic ROS-RDB algorithm by performing an additional ten-time replication of the results. Both pair tests between the quality of classifiers for individual data sets and ranking tests, used for general assessment of the relations between them, were carried out using the Wilcoxon test using 5% significance level. Due to the imbalanced nature of the considered problems, in assessing the quality of solutions it was decided to use *precision* and *recall* metrics, supplemented with aggregated *F1-score*, *balanced accuracy score* and *geometric-mean-score* metrics. Full source code of the performed tests, along with the method implementations and a full report of results, are located on the publicly available Git repository[1].

[1] http://github.com/w4k2/sdb

DOI: 10.1007/978-3-030-50423-6_8

Table 1: Overview of imbalanced classification datasets selected for experimental evaluation.

DATASET	SAMPLES	FEATURES	IR
ecoli-0-3-4-vs-5	200	7	1:9
yeast-2-vs-4	514	8	1:9
ecoli-0-6-7-vs-3-5	222	7	1:9
ecoli-0-2-3-4-vs-5	202	7	1:9
glass-0-1-5-vs-2	172	9	1:9
yeast-0-3-5-9-vs-7-8	506	8	1:9
yeast-0-2-5-6-vs-3-7-8-9	1004	8	1:9
yeast-0-2-5-7-9-vs-3-6-8	1004	8	1:9
ecoli-0-4-6-vs-5	203	6	1:9
ecoli-0-1-vs-2-3-5	244	7	1:9
ecoli-0-2-6-7-vs-3-5	224	7	1:9
glass-0-4-vs-5	92	9	1:9
ecoli-0-3-4-6-vs-5	205	7	1:9
ecoli-0-3-4-7-vs-5-6	257	7	1:9
yeast-0-5-6-7-9-vs-4	528	8	1:9
vowel0	988	13	1:10
ecoli-0-6-7-vs-5	220	6	1:10
glass-0-1-6-vs-2	192	9	1:10
ecoli-0-1-4-7-vs-2-3-5-6	336	7	1:11
led7digit-0-2-4-5-6-7-8-9-vs-1	443	7	1:11
glass-0-6-vs-5	108	9	1:11
ecoli-0-1-vs-5	240	6	1:11
glass-0-1-4-6-vs-2	205	9	1:11
glass2	214	9	1:12
ecoli-0-1-4-7-vs-5-6	332	6	1:12
ecoli-0-1-4-6-vs-5	280	6	1:13
shuttle-c0-vs-c4	1829	9	1:14
yeast-1-vs-7	459	7	1:14
glass4	214	9	1:15
ecoli4	336	7	1:16
page-blocks-1-3-vs-4	472	10	1:16
glass-0-1-6-vs-5	184	9	1:19
shuttle-c2-vs-c4	129	9	1:20
yeast-1-4-5-8-vs-7	693	8	1:22
glass5	214	9	1:23
yeast-2-vs-8	482	8	1:23
yeast4	1484	8	1:28
yeast-1-2-8-9-vs-7	947	8	1:31
yeast5	1484	8	1:33
ecoli-0-1-3-7-vs-2-6	281	7	1:39
yeast6	1484	8	1:41

4 Experimental evaluation

4.1 Results

Scores and paired tests Table 2 contains the results achieved by each of the considered algorithms for the aggregate, *F1-score* metric. The ROS-RDB method, being a typical approach to deal with imbalanced data using single model, looks the worst in the pool, which not only does not improve RDB results, but also often leads to statistically significant worse results. The FTDB method, although sporadically, leads to a significant improvement over RDB, never achieving results significantly inferior to it. Definitely the best in this competition is the SDB method proposed in this paper, which in eleven cases is statistically significantly better than each of the other methods, and in fourteen cases better than RDB.

Table 2: Results achieved by analyzed methods for all considered datasets with *F1-score* metric. Bold values shows dependency to the best classifier in a competition and the numbers below scores show classifier significantly worse than the one in the column.

Dataset	**1** **RDB**	**2** **ROS-RDB**	**3** **FTDB**	**4** **SDB**
ecoli-0-3-4-vs-5	0.340 2	0.268 —	0.396 2	**0.670** ***all***
yeast-2-vs-4	**0.295** —	0.269 —	**0.334** **2**	**0.454** **2**
ecoli-0-6-7-vs-3-5	**0.190** —	**0.298** —	**0.218** —	**0.338** —
ecoli-0-2-3-4-vs-5	0.332 —	0.260 —	0.383 —	**0.659** ***all***
glass-0-1-5-vs-2	**0.218** —	**0.183** —	**0.232** —	**0.239** —
yeast-0-3-5-9-vs-7-8	**0.269** —	**0.212** —	**0.252** —	**0.229** —
yeast-0-2-5-6-vs-3-7-8-9	**0.262** —	**0.478** **3**	0.401 —	**0.469** —
yeast-0-2-5-7-9-vs-3-6-8	0.201 2	0.165 —	0.272 1, 2	**0.381** ***all***
ecoli-0-4-6-vs-5	**0.629** —	**0.584** —	**0.629** —	**0.736** —
ecoli-0-1-vs-2-3-5	**0.217** —	**0.244** —	**0.217** —	**0.387** —
ecoli-0-2-6-7-vs-3-5	**0.208** —	**0.186** —	**0.208** —	**0.256** —
glass-0-4-vs-5	**0.960** —	**0.960** —	**0.960** —	**0.760** —
ecoli-0-3-4-6-vs-5	0.312 2	0.247 —	0.350 2	**0.669** ***all***
ecoli-0-3-4-7-vs-5-6	0.356 —	0.251 —	0.489 —	**0.665** ***all***
yeast-0-5-6-7-9-vs-4	0.174 —	0.173 —	0.195 —	**0.362** ***all***
vowel0	**0.709** **2**	0.562 —	**0.697** **2**	**0.676** **2**
ecoli-0-6-7-vs-5	**0.633** —	**0.663** —	**0.660** —	**0.688** —
glass-0-1-6-vs-2	0.199 —	**0.231** —	**0.218** —	**0.236** **1**
ecoli-0-1-4-7-vs-2-3-5-6	**0.324** —	**0.384** —	**0.357** —	**0.384** —
led7digit-0-2-4-5-6-7-8-9-vs-1	**0.640** —	**0.622** —	**0.640** —	**0.646** —
glass-0-6-vs-5	**0.867** —	**0.867** —	**0.867** —	**0.733** —
ecoli-0-1-vs-5	**0.632** —	**0.582** —	**0.638** —	**0.823** —
glass-0-1-4-6-vs-2	**0.229** —	**0.260** —	**0.230** —	**0.240** —
glass2	**0.169** —	**0.195** —	**0.179** —	**0.187** **1**
ecoli-0-1-4-7-vs-5-6	**0.538** —	**0.662** —	**0.570** —	**0.688** —
ecoli-0-1-4-6-vs-5	**0.709** —	**0.664** —	**0.723** —	**0.764** —
shuttle-c0-vs-c4	**0.980** —	**0.980** —	**0.980** —	**0.980** —
yeast-1-vs-7	0.141 —	0.136 —	0.153 1, 2	**0.223** ***all***
glass4	**0.190** —	**0.481** —	**0.233** —	**0.233** —
ecoli4	**0.696** —	0.602 —	**0.696** —	**0.787** **2**
page-blocks-1-3-vs-4	**0.511** —	**0.521** —	**0.524** —	**0.540** —
glass-0-1-6-vs-5	**0.760** —	**0.760** —	**0.760** —	**0.667** —
shuttle-c2-vs-c4	**0.813** —	**0.800** —	**0.813** —	**1.000** —
yeast-1-4-5-8-vs-7	0.086 —	**0.088** —	**0.085** —	**0.103** **1**
glass5	**0.768** —	**0.768** —	**0.768** —	**0.693** —
yeast-2-vs-8	**0.254** —	0.190 —	**0.262** **2**	**0.202** —
yeast4	0.073 —	0.071 —	0.086 1, 2	**0.117** ***all***
yeast-1-2-8-9-vs-7	0.067 —	0.066 —	0.068 —	**0.098** ***all***
yeast5	0.154 2	0.120 —	0.165 1, 2	**0.642** ***all***
ecoli-0-1-3-7-vs-2-6	**0.434** —	**0.388** —	**0.434** —	**0.490** —
yeast6	0.066 2	0.060 —	0.066 2	**0.169** ***all***

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_8

Table 3: Results achieved by analyzed methods for all considered datasets with *recall* metric. Bold values shows dependency to the best classifier in a competition and the numbers below scores show classifier significantly worse than the one in the column.

Dataset	**1** **RDB**	**2** **ROS-RDB**	**3** **FTDB**	**4** **SDB**
ecoli-0-3-4-vs-5	**0.850** —	**0.850** —	**0.850** —	**0.850** —
yeast-2-vs-4	**0.902** —	**0.922** —	**0.902** —	**0.825** —
ecoli-0-6-7-vs-3-5	**0.170** —	**0.260** —	**0.210** —	**0.360** —
ecoli-0-2-3-4-vs-5	**0.850** —	**0.850** —	**0.850** —	**0.850** —
glass-0-1-5-vs-2	**0.633** —	**0.733** —	**0.633** —	**0.633** —
yeast-0-3-5-9-vs-7-8	**0.880** —	**0.880** —	**0.800** —	**0.760** —
yeast-0-2-5-6-vs-3-7-8-9	**0.307** —	**0.557** **3**	0.436 —	**0.505** —
yeast-0-2-5-7-9-vs-3-6-8	**0.917** —	**0.854** —	**0.897** —	**0.897** —
ecoli-0-4-6-vs-5	**0.650** —	**0.650** —	**0.650** —	**0.850** —
ecoli-0-1-vs-2-3-5	**0.160** —	**0.200** —	**0.160** —	**0.440** —
ecoli-0-2-6-7-vs-3-5	**0.190** —	**0.190** —	**0.190** —	**0.310** —
glass-0-4-vs-5	**1.000** —	**1.000** —	**1.000** —	**0.800** —
ecoli-0-3-4-6-vs-5	**0.850** —	**0.850** —	**0.850** —	**0.850** —
ecoli-0-3-4-7-vs-5-6	**0.760** —	**0.760** —	**0.760** —	**0.920** —
yeast-0-5-6-7-9-vs-4	**0.960** —	**0.960** —	**0.920** —	**0.864** —
vowel0	**0.811** —	**0.844** —	**0.811** —	**0.811** —
ecoli-0-6-7-vs-5	**0.700** —	**0.850** —	**0.750** —	**0.800** —
glass-0-1-6-vs-2	**0.683** —	**0.800** —	**0.683** —	**0.683** —
ecoli-0-1-4-7-vs-2-3-5-6	**0.267** —	**0.333** —	**0.300** —	**0.333** —
led7digit-0-2-4-5-6-7-8-9-vs-1	**0.757** —	**0.832** —	**0.757** —	**0.786** —
glass-0-6-vs-5	**0.900** —	**0.900** —	**0.900** —	**0.800** —
ecoli-0-1-vs-5	**0.600** —	**0.650** —	**0.650** —	**0.850** —
glass-0-1-4-6-vs-2	**0.650** —	**0.700** —	**0.650** —	**0.650** —
glass2	**0.733** —	**0.833** —	**0.733** —	**0.733** —
ecoli-0-1-4-7-vs-5-6	**0.560** —	**0.760** —	**0.600** —	**0.760** —
ecoli-0-1-4-6-vs-5	**0.800** —	**0.850** —	**0.850** —	**0.850** —
shuttle-c0-vs-c4	**0.984** —	**0.984** —	**0.984** —	**0.984** —
yeast-1-vs-7	**0.933** —	**0.933** —	**0.933** —	**0.833** —
glass4	**0.200** —	**0.567** —	**0.267** —	**0.267** —
ecoli4	**0.950** —	**0.950** —	**0.950** —	**0.900** —
page-blocks-1-3-vs-4	**0.593** —	**0.667** —	**0.627** —	**0.667** —
glass-0-1-6-vs-5	**0.900** —	**0.900** —	**0.900** —	**0.800** —
shuttle-c2-vs-c4	**1.000** —	**1.000** —	**1.000** —	**1.000** —
yeast-1-4-5-8-vs-7	**0.967** **4**	**1.000** **4**	**0.933** **4**	0.800 —
glass5	**0.900** —	**0.900** —	**0.900** —	**0.800** —
yeast-2-vs-8	**0.950** —	**1.000** —	**0.900** —	**0.650** —
yeast4	**0.962** —	**0.982** —	**0.962** —	**0.904** —
yeast-1-2-8-9-vs-7	**1.000** **4**	**1.000** **4**	**1.000** **4**	0.733 —
yeast5	**1.000** —	**1.000** —	**1.000** —	**0.886** —
ecoli-0-1-3-7-vs-2-6	**0.800** —	**0.800** —	**0.800** —	**0.800** —
yeast6	**1.000** —	**0.971** —	**0.971** —	**0.914** —

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_8

For both the *precision metric* and the other aggregate measures (*balanced accuracy score* and *geometric mean score*), the observations are identical to those drawn from the *F1-score*, so the relevant result tables are not attached directly to the article, while still being public in the repository indicated in the previous section.

The aggregate metrics, such as *F1-score*, allows to draw some binding conclusions, but does not give a full picture of interpretation. As expected, with the *recall* metric (Table 3), the FTDB and RDB algorithms give some deterioration relative to both the base method and the ROS-RDB approach. Statistical significance occurs in this difference, however, only once for DTDB and twice for RDB.

Rank tests The final comparison of the considered solutions was carried out by ranking tests, included in Table 4. The ROS-RDB method obtains a small, but statistically significant advantage in the ranking over all other methods for the *recall* metric, but in all other measures it stands out very negatively, which leads to suggestions about its overall uselessness in the considered task of highly imbalanced data classification. If the goal of counteracting the tendency of favoring in the prediction of the majority class (which was stated as the basic problem in the classification of imbalanced data) is to equalize the impact of both classes, on the example of the considered data sets, the ROS method must be rejected because it leads to the reverse tendency. In the case of *precision* and each of the aggregate metrics the same statistically significant relation is observed. The RDB method is better than ROS-RDB, the FTDB method is better than both RDB methods, and the SDB proposed in this paper is significantly better than all the competitors in the considered pool of solutions.

Table 4: Results for mean ranks according to all considered metrics.

	1	2	3	4
Metric	**RDB**	**ROS-RDB**	**FTDB**	**SDB**
F1-score	2.215	1.927	2.607	3.251
	2	–	1,2	*all*
precision	2.271	1.837	2.646	3.246
	2	–	1,2	*all*
recall	2.463	2.800	2.446	2.290
	–	*all*	–	–
balanced accuracy	2.224	1.963	2.595	3.217
	2	–	1,2	*all*
geometric mean score	2.205	1.868	2.641	3.285
	2	–	1,2	*all*

5 Conclusions

Following paper, considering the binary classification of imbalanced data, proposed the application of the *Fuzzy Templates* method in the construction of the support-domain decision boundary for a single model in order to balance the impact of classes of different counts on the prediction of the decision system. The proposal was further developed to use both *standard normalization* metrics, introducing the *Standard Decision Boundary* method. Both solutions were tested in computer experiments on the example of a highly imbalanced dataset collection and compared to both the base method and the *state-of-art* preprocessing method.

Both proposed solutions seem to improve the quality of imbalanced data classification in relation to the regular support-domain decision boundary, in contrast to oversampling, without leading to overweight of the predictive towards the minority class. Modification of the use of *Fuzzy Templates* in the form of *Standard Decision Boundary* is also more effective than the simple use of a class support prototype and may be considered a recommendable solution to the problem of binary classification of imbalanced data. Due to the promising results achieved for individual models, the next works will attempt to generalize the SDB method for *classifier ensembles*.

Acknowledgements

This work was supported by the Polish National Science Centre under the grant No. 2017/27/B/ST6/01325 as well as by the statutory funds of the Department of Systems and Computer Networks, Faculty of Electronics, Wroclaw University of Science and Technology.

References

1. Aditsania, A., Adiwijaya, Saonard, A.L.: Handling imbalanced data in churn prediction using ADASYN and backpropagation algorithm. In: Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017 (2017)
2. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic and Soft Computing (2011)
3. del Amo, A., Montero, J., Cutello, V.: On the principles of fuzzy classification. Annual Conference of the North American Fuzzy Information Processing Society - NAFIPS (1999)
4. Bishop, C.M.: Pattern recognition and machine learning. springer (2006)
5. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from Imbalanced Data Sets (2018)
6. Fernández, A., García, S., Herrera, F., Chawla, N.V.: SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary (2018)
7. Ganganwar, V.: An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering (2012)

DOI: 10.1007/978-3-030-50423-6_8

8. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of the International Joint Conference on Neural Networks (2008)
9. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions (2016)
10. Ksieniewicz, P.: Undersampled Majority Class Ensemble for highly imbalanced binary classification. In: Second International Workshop on Learning with Imbalanced Domains: Theory and Applications. pp. 82–94 (2018)
11. Ksieniewicz, P.: Combining Random Subspace Approach with smote Oversampling for Imbalanced Data Classification. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 11734 LNAI, pp. 660–673 (2019)
12. Ksieniewicz, P., Woźniak, M.: Imbalanced data classification based on feature selection techniques. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 11315 LNCS, pp. 296–303 (2018)
13. Ksieniewicz, P., Wozniak, M., Torgo, L., Krawczyk, B., Branco, P., Moniz, N.: Dealing with the task of imbalanced, multidimensional data classification using ensembles of exposers. Proceedings of Machine Learning Research (2017)
14. Kuncheva, L.: Fuzzy classifier design, vol. 49. Springer Science & Business Media (2000)
15. Kuncheva, L.I., Bezdek, J.C., Duin, R.P.: Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition (2001)
16. Kuncheva, L.I., Bezdek, J.C., Sutton, M.A.: On combining multiple classifiers by fuzzy templates. In: Annual Conference of the North American Fuzzy Information Processing Society - NAFIPS (1998)
17. Mitchell, T.M.: The Discipline of Machine Learning. Machine Learning (2006)
18. Moreo, A., Esuli, A., Sebastiani, F.: Distributional random oversampling for imbalanced text classification. In: SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (2016)
19. Ohsaki, M., Wang, P., Matsuda, K., Katagiri, S., Watanabe, H., Ralescu, A.: Confusion-matrix-based kernel logistic regression for imbalanced data classification. IEEE Transactions on Knowledge and Data Engineering (2017)
20. Prusa, J., Khoshgoftaar, T.M., DIttman, D.J., Napolitano, A.: Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data. In: Proceedings - 2015 IEEE 16th International Conference on Information Reuse and Integration, IRI 2015 (2015)
21. Rodriguez-Torres, F., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: Deterministic oversampling methods based on SMOTE. Journal of Intelligent and Fuzzy Systems (2019)
22. Wang, Q., Luo, Z.H., Huang, J.C., Feng, Y.H., Liu, Z.: A novel ensemble method for imbalanced data learning: Bagging of extrapolation-SMOTE SVM. Computational Intelligence and Neuroscience (2017)
23. Woźniak, M., Graña, M., Corchado, E.: A survey of multiple classifier systems as hybrid systems. Information Fusion (2014)
24. Xu, Y., Yang, Z., Zhang, Y., Pan, X., Wang, L.: A maximum margin and minimum volume hyper-spheres machine with pinball loss for imbalanced data classification. Knowledge-Based Systems (2016)
25. Zhang, Y.: Deep Generative Model for Multi-Class Imbalanced Learning. ProQuest Dissertations and Theses (2018)

DOI: 10.1007/978-3-030-50423-6_8