A Correction Method of a Base Classifier Applied to Imbalanced Data Classification

Pawel Trajdos[0000−0002−4337−6847] and Marek Kurzynski[0000−0002−0401−2725]

Wroclaw University of Science and Technology, Wroclaw, Poland,
pawel.trajdos@pwr.edu.pl, marek.kurzynski@pwr.edu.pl

Abstract. In this paper, the issue of tailoring the soft confusion matrix classifier to deal with imbalanced data is addressed. This is done by changing the definition of the soft neighbourhood of the classified object. The first approach is to change the neighbourhood to be more local by changing the Gaussian potential function approach to the nearest neighbour rule. The second one is to weight the instances that are included in the neighbourhood. The instances are weighted inversely proportional to the a priori class probability. The experimental results show that for one of the investigated base classifiers, the usage of the KNN neighbourhood significantly improves the classification results. What is more, the application of the weighting schema also offers a significant improvement.

Keywords: classification, probabilistic model, randomized reference classifier, soft confusion matrix, imbalanced data

1 Introduction

Imbalanced dataset, denoting the case when there is a significant difference between the prior probabilities for different classes, is a difficult problem for classification. It results from the fact that – on the one hand – for most such problems it is desirable to build classifiers with good performance on the minority class being the class of interest, but – on the other hand – in highly imbalanced datasets, the minority class is mostly sensitive to singular classification errors. Let's cite two practical classification problems as examples of such situation. The first example concerns fraud detection in online monetary transactions. Although fraud is becoming more common and this is a growing problem for banking systems, the number of fraudulent transactions is typically a small fraction of all financial transactions. So, we have here an imbalanced classification problem in which the classifier should correctly recognize objects from the minority class, i.e. detect all fraud transactions and at the same time it should not give false alarms. A similar situation is in the second example regarding computer-aided medical diagnosis. In the simple task of medical screening tests we have two classes: healthy people (majority class) and people suffering from a rare disease (minority class). Requirements for the diagnostic algorithm are the same as before: to successfully detect ill people.

DOI: 10.1007/978-3-030-50423-6_7

There are more negative consequences of imbalanced dataset that hinder correct classification. We can mention here [28]: overlapping classes (clusters of minority class are heavily contaminated with majority class), lack of density (learners do not have enough data to make generalization about the distribution of minority samples), noisy data (the presence of noise degrades the information capacity of minority class samples) and dataset shift (training and testing data follow the different distribution).

The difficulty in classifying imbalanced datasets has caused great interest among the pattern recognition research community in methods and algorithms that would effectively solve this problem. The proposed methods of classification of imbalanced datasets can be divided into two following categories [23,1]:

1. **Data level approach** (or external techniques) involves manipulating instances of the learning set to obtain a more balanced class distribution. This goal can be achieved through undersampling and/or oversampling procedures. In the first approach, instances are removed from the majority class, while in the second technique new artificial instances are added to the minority class. Different specified algorithms for both methods define the way of removing (adding) instances from the majority (to the minority) class. Random undersampling [17], ACOSampling [41], EUSBoost [10,19] for undersampling approach and SMOTE [3], ADASYN [14], SNOCC [42] for oversampling procedures are exemplary algorithms for this category of methods.
2. **Algorithm level approach** (or internal techniques) denotes classifiers which directly learn class characteristics from the imbalanced data. The leading approaches in this category of methods are:
   - **Improved algorithms** denote classifiers that are modified (improved) to fit their properties to the specifics of imbalanced classification. Support vector machines [15], artificial neural networks [8], k-nearest neighbours [25], decision tree [24], fuzzy inference system [7] and random forest [40] are the most popular methods which have been adapted to classification of imbalanced data.
   - **One-class learning** algorithms for imbalanced problem are trained on the representation of the minority class [32].
   - **Cost-sensitive learning** is based on a very-well known classification scheme in which the cost of misclassification depends on the kind of error made. For example, in the Bayes decision theory this cost is modeled by loss function (loss matrix), which practically can have any values [6]. Application of this scheme to the classification of imbalanced data denotes that first we define cost of misclassification of objects from the minority (class of interest) and majority class (e.g. using domain expert opinion) and then we build a classifier (learner) which takes into account different costs for different classification errors (e.g. minimizing the expected cost or risk) [18,20,31]
   - **Ensemble learning** – in this approach several base classifiers are trained and their predictions are combined to produce the final classification decision [9]. Ensemble methods applied to the imbalanced data classification combine ensemble learning algorithms and techniques dedicated

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_7

to imbalanced problems, e.g. undersampling/oversampling procedures [29,37] or cost-sensitive learning [31].

This paper is devoted to the new classifier for imbalanced data which belongs to the algorithm level category of methods. The algorithm developed is based on the author's method of improving the operation of any classifier called base classifier. In the method first the local class-dependent probabilities of misclassification and correct classification are determined. For this purpose two original concepts of randomized reference classifier (RRC) [39] and soft confusion matrix (SCM) [34] are used. Then, the determined probabilities are used for correction of the decision of the base classifier to increase the chance of correct classification of the recognized object. The developed method has already been successfully applied for the construction of multi-classifier systems [34], in multi-label recognition [35,36] and in the recognition of biosignals [22]. However, the algorithm is sensitive to imbalanced data distribution. In other words, its correction ability is lower when the class imbalance ratio is higher. To make the developed approach more practical, it is necessary to provide a mechanism of dealing with imbalanced class distribution. And this paper is aimed at dealing with this issue. In the proposed algorithm for imbalanced data, the classification functions have additional factors inversely proportional to the class size with the parameter experimentally tuned. This mechanism allows a controlled change in the degree of correction of the base classifier to highlight minority classes.

The paper is organized as follows. Section 2 introduces the formal notation used in the paper and provides a description of the proposed approach. The experimental setup is given in section 3. In section 4 experimental results are given and discussed. Section 5 concludes the paper.

2 Proposed Method

2.1 Preliminaries

Let be given pattern recognition problem in which x denotes d-dimensional feature vector of an object and j is its class number. Feature vector x belongs to the feature space $\mathcal{X} = \Re^d$ and class number j takes value in a finite set $\mathcal{M} = \{1, 2, 3, ..., M\}$. Let ψ be a trained classifier which assigns a class number to the recognized object. In other words, ψ_n maps the feature space to the set of class labels, viz. $\psi : \mathcal{X} \to \mathcal{M}$. Classifier ψ will be called base classifier. We suppose that ψ is described by the canonical model, i.e. for given object x it first produces values of normalized classification functions (supports) $g_i(x), i \in \mathcal{M}$ ($g_i(x) \in [0, 1], \sum g_i(x) = 1$) and then classifies object according to the maximum support rule:

$$\psi(x) = i \Leftrightarrow g_i(x) = \max_{k \in \mathcal{M}} g_k(x). \tag{1}$$

However, the base classifier ψ and formula (1) will not be used directly for classification. To classify object x a decision scheme will be used, which indirectly takes into account classification result of ψ and additionally uses the local

DOI: 10.1007/978-3-030-50423-6_7

(relative to x) properties of base classifier for correction of its decision to increase the chance of correct classification of the recognized object. The proposed decision scheme will be further modified in terms of imbalanced data classification. Source of information about the properties of the base classifier used in the correction procedure of ψ is a validation set:

$$\mathcal{V} = \{(x_1, j_1), (x_2, j_2), \ldots, (x_N, j_N)\}; \;\; x_k \in \mathcal{X}, \; j_k \in \mathcal{M} \tag{2}$$

containing pairs of feature vectors and their corresponding class labels.

The basis for the proposed method of classification is the probabilistic model meaning the assumption that x and j are observed values of random variables X and J, respectively.

2.2 Correction of Base Classifier

The corrected base classifier $\psi^{(Corr)}$, using the probabilistic model of the recognition task, acts according to the known Bayes scheme:

$$\psi^{(Corr)}(x) = i \Leftrightarrow P(i|x) = \max_{k \in \mathcal{M}} P(k|x). \tag{3}$$

Now, however, we will express *a posteriori* probabilities $P(j|x), j \in \mathcal{M}$ in a different way than in the classic Bayesian formula, making them dependent on the probabilistic properties of the base classifier, namely:

$$P(j|x) = \sum_{i \in \mathcal{M}} P(i, j|x) = \sum_{i \in \mathcal{M}} P(i|x) P(j|i, x), \tag{4}$$

where $P(i|x) = P(\psi(x) = i)$ and $P(j|i, x)$ denotes the probability that x belongs to the j-th class given that $\psi(x) = i$. Unfortunately, it should be noted that with both probabilities there is a serious problem. First, for the deterministic base classifier ψ probabilities $p(i|x), i \in \mathcal{M}$ are equal to 0 or 1. Secondly, probabilities $P(j|i, x)$ are class-dependent probabilities of the correct classification (for $i = j$) and the misclassification (for $i \neq j$) of ψ at the point x and estimating these probabilities would require many validation objects at this point.

To give the formula (4) a constructive character and calculate both probabilities we will use two concepts: the randomized reference classifier (RRC) and the soft confusion matrix (SCM). The RRC is randomized model of classifier ψ and with its help the probabilities $p(\psi(x) = i) \in [0, 1]$ will be calculated. In turn, the SCM will be used to determine the assessment of correct and misclassification of ψ at the point x, i.e. probabilities $P(j|\psi(x) = i), i, j \in \mathcal{M}$. The method defines the surrounding of the point x containing validation objects in terms of fuzzy sets allowing for flexible selection of membership functions and taking into account the case of imbalanced classes.

2.3 Randomized Reference Classifier

RRC is a probabilistic classifier which is defined by a probability distribution over the set of class labels $\mathcal{M}$. Its classifying functions $\{\delta_j(x)\}_{j \in \mathcal{M}}$ are observed

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_7

values of random variables $\{\varDelta_j(x)\}_{j\in\mathcal{M}}$ fulfilling the following conditions:

$$\varDelta_i(x) \in [0,1], \tag{5}$$

$$\sum_{i\in\mathcal{M}} \varDelta_i(x) = 1, \tag{6}$$

$$\mathbf{E}\left[\varDelta_i(x)\right] = g_i(x),\ i \in \{0,1\}, \tag{7}$$

where $\mathbf{E}$ is the expected value operator. Conditions (5) and (6) follow from the normalization properties of class supports, whereas condition (7) provides the equivalence of the randomized model $\psi^{(RRC)}$ and base classifier ψ. Based on the latter condition, the RRC can be used to provide a randomized model of any classifier that returns a vector of class-specific supports $g(x)$.

It is obvious, that the probability of classifying an object x into the class i using the RRC is as follows:

$$P(\psi^{(RRC)}(x) = i) = P[\varDelta_i(x) > \varDelta_k(x), k \in \mathcal{M}\setminus i]. \tag{8}$$

The probability on the right side of (8) can be easily determined if we assume – as in the original work of Woloszynski and Kurzynski [39] – that $\varDelta_i(x)$ have the beta distribution.

Since $\psi^{(RRC)}$ acts – on average – as the modeled base classifier, the following approximation is fully justified:

$$P(\psi(x) = i) \approx P[\varDelta_i(x) > \varDelta_k(x), k \in \mathcal{M}\setminus i],\ x \in \mathcal{X}, i \in \mathcal{M}. \tag{9}$$

2.4 Soft Confusion Matrix

Classically, the confusion matrix is in the form of two-dimensional table, in which the rows correspond to the true classes while the columns match the outcomes of the classifier ψ, as it shown in Table 1.

Table 1. The multiclass confusion matrix of classifier ψ

		Classification by ψ			
		1	2	...	M
	1	$\varepsilon_{1,1}$	$\varepsilon_{1,2}$	...	$\varepsilon_{1,M}$
True	2	$\varepsilon_{2,1}$	$\varepsilon_{2,2}$	...	$\varepsilon_{2,M}$
class	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	M	$\varepsilon_{M,1}$	$\varepsilon_{M,2}$	...	$\varepsilon_{M,M}$

The value $\varepsilon_{i,j}$ is determined from validation set (2) as the following ratio ($|\cdot|$ is the cardinality of a set):

$$\varepsilon_{i,j} = \frac{|\bar{\mathcal{V}}_j \cap \bar{\mathcal{D}}_i|}{|\bar{\mathcal{V}}_j|}, \tag{10}$$

where $\bar{\mathcal{V}}_j = \{x_k \in \mathcal{V} : j_k = j\}$ (class set) denotes the set of validation objects from the j-th class and $\bar{\mathcal{D}}_i = \{x_k \in \mathcal{V} : \psi(x_k) = i\}$ (decision set) is the set of validation objects assigned by ψ to the i-th class.

The confusion matrix (10) gives an overall (for the whole feature space) image of the classifier properties, while our purpose is to assess the local probabilities $P(j|i,x)$. For this reason, we will generalize the term of confusion matrix enabling free shaping of the concept of "locality" and assigning weights to individual validation objects. Generalized confusion matrix, called the soft confusion matrix (SCM), referred to the recognized object $x \in \mathcal{X}$ is defined as follows:

$$\varepsilon_{i,j}(x) = \frac{|\mathcal{V}_j \cap \mathcal{D}_i \cap \mathcal{N}(x)|}{|\mathcal{V}_j \cap \mathcal{N}(x)|}, \tag{11}$$

where $\mathcal{V}_j, \mathcal{D}_i$ and $\mathcal{N}(x)$ are fuzzy sets specified in the validation set $\mathcal{V}$ and $|\cdot|$ denotes the cardinality of a fuzzy set [5].

Now we will define and give a practical interpretation of fuzzy sets that create the proposed SCM concept (11).

The class set $\mathcal{V}_j$. Identically as in (10), this set denotes the set of validation objects from the j-th class. Formulating the set $\mathcal{V}_j$ in terms of fuzzy sets theory it can be assumed that the grade of membership of validation object x_k to $\mathcal{V}_j$ is the class indicator which leads to the following definition of $\mathcal{V}_j$ as the fuzzy set:

$$\mathcal{V}_j = \{(x_k, \mu_{\mathcal{V}_j}(x_k))\}, \quad \text{where} \quad \mu_{\mathcal{V}_j}(x_k) = \begin{cases} 1 & \text{if } j_k = j, \\ 0 & \text{elsewhere.} \end{cases} \tag{12}$$

The decision set $\mathcal{D}_i$. For the confusion matrix (10) the crisp decision set $\bar{\mathcal{D}}_i$ includes validation objects x_k for which $\psi(x_k) = i$. The original concept of fuzzy decision set $\mathcal{D}_j$ is defined as follows:

$$\mathcal{D}_i = \{(x_k, \mu_{\mathcal{D}_i}(x_k)) : x_k \in \mathcal{V}, \mu_{\mathcal{D}_i}(x_k) = P(i|x_k)\}, \tag{13}$$

where $P(i|j, x_k)$ is calculated according to (9). Formula (13) demonstrates that now the membership of validation object x_k to the set $\mathcal{D}_i$ is not determined by the decision of classifier ψ. The grade of membership of validation object x_k to $\mathcal{D}_i$ depends on the potential chance of classifying object x_k to the i-th class by the base classifier. We assume, that this potential chance is equal to the probability $P(i|x) = P(\psi(x) = i)$ calculated using the randomized model RRC of base classifier ψ.

The neighbourhood set $\mathcal{N}(x)$. As it seems, this set play the crucial role in the proposed concept of SCM, because it decide which validation objects x_k and with which weights will be taken into account in the procedure of determining the local properties of $\psi(x)$. Formally, $\mathcal{N}(x)$ is also a fuzzy set:

$$\mathcal{N}(x) = \{(x_k, \mu_{\mathcal{N}(x)}(x_k)) : x_k \in \mathcal{V}\}, \tag{14}$$

but its membership function is not defined univocally because it depends on the adopted concept of "locality" (relative to x). There are two typical methods of determining the set $\mathcal{V}(x)$. In the first approach the neighbourhood of x is precisely defined and only validation objects belonging to this neighbourhood are used to calculate (11). In the second method all validation points are members of the set $\mathcal{N}(x)$ and its membership functions is equal to 1 for $x_k = x$ and decreases with increasing the distance between x_k and x. In the further experimental investigations two forms of the fuzzy set $\mathcal{N}(x)$ were used as representative of both approaches.

1. ***K*NN Neighborhood**. Let first define the K-neighbourhood of the test object x as the set of K nearest validation objects, viz.

   $$\mathcal{K}_K(x) = \{x_{n1}, \ldots, x_{nK} \in \mathcal{V} : \max_{l=1,2,\ldots,K} \mathrm{dist}(x_{nl}, x)^2 \leq \min_{x_k \notin \mathcal{K}_K(x)} \mathrm{dist}(x_k, x)^2\}, \tag{15}$$

   where $\mathrm{dist}(x_k, x)^2$ denotes the Euclidean distance in the feature space $\mathcal{X}$. The KNN-related membership function of $\mathcal{N}(x)$ is defined as follows:

   $$\mu^{(K)}_{\mathcal{N}(x)}(x_k) = \begin{cases} 1 \text{ if } x_k \in \mathcal{K}(x), \\ 0 \text{ otherwise.} \end{cases} \tag{16}$$

   This kind of neighbourhood should be more fragile to the local properties of the data since it completely ignores the instances that are not in $\mathcal{K}$.
2. **Gaussian Neighborhood**. In this method the Gaussian membership function was applied for defining the set $\mathcal{N}(x)$:

   $$\mu^{(G)}_{\mathcal{N}(x)}(x_k) = \exp(-\beta \mathrm{dist}(x, x_k)^2), \tag{17}$$

   where $\beta \in \mathbb{R}_+$ is parametr of μ. The Gaussian-based neighbourhood was originally proposed to use with the SCM classifier in [34].

2.5 Dealing with Imbalanced data

In this paper, the issue of imbalanced class distribution is dealt with via modification of the membership function of the neighbourhood set $\mathcal{N}(x)$. We propose to add a new factor (weight) to original membership function $\mu_{\mathcal{N}(x)}(x_k)$ which is inversely proportional to the *a priori* probability of class $P(j), j \in \mathcal{M}$. Assuming that the minority class is the class of interest, such a method relatively enhances class proportionally to its importance. The proposed approach also means, that the neighbourhood set $\mathcal{N}(x)$ is now dependent on the class j to which the validation objects used to calculate $\varepsilon_{i,j}(x)$ in (11) belong. Thus, the membership function of the neighbourhood set (14) that includes imbalanced classes is as follows:

$$\mu_{\mathcal{N}_j(x)}(x_k) = c\mu_{\mathcal{N}(x)}(x_k)P(j)^{-\gamma}. \tag{18}$$

$\gamma \in \mathbb{R}_+$ is the coefficient that controls weighting intensity and c is normalized coefficient.

Finally, from (14) and modification (18) we get the following approximation:

$$P(j|i, x) \approx \frac{\varepsilon_{i,j}(x)}{\sum_{j\in\mathcal{M}} \varepsilon_{i,j}(x)}, \tag{19}$$

DOI: 10.1007/978-3-030-50423-6_7

which together with (9), (4) and (3) give the corrected base classifier $\psi^{(Corr)}(x)$ in the version tailored for the case of imbalanced data.

3 Experimental Setup

To validate the classification quality obtained by the proposed approaches the experimental evaluation, which setup is described below, is performed.

The following base classifiers were employed:

- ψ_{NB} – Naive Bayes classifier with kernel density estimation [13].
- ψ_{J48} – Weka version of the C4.5 algorithm [27] with Laplace smoothing [26]
- ψ_{NC} – nearest centroid (Nearest Prototype) [21]

The classifiers implemented in WEKA framework [12] were used. If not stated otherwise, the classifier parameters were set to their defaults. For each base classifier, the training dataset is resampled with weights inversely proportional to the *a priori* probability of instance-specific class. This is to make base classifiers robust against imbalanced data.

During the experimental evaluation the following classifiers were compared:

1. ψ_{R} – unmodified base classifier,
2. ψ_{G} – SCM classifier with unmodified Gaussian neighbourhood,
3. ψ_{Gw} – SCM classifier with weighted Gaussian neighbourhood,
4. ψ_{K} – SCM classifier with unmodified KNN neighbourhood,
5. ψ_{Kw} – SCM classifier with weighted KNN neighbourhood.

The size of the neighbourhood, expressed as β coefficient, the number of nearest neighbours K and the weighting coefficient γ, were chosen using a fivefold cross-validation procedure and the grid search technique. The following values of β, K and γ were considered: $\beta \in \left\{2^{-2}, 2^{-1}, 2^{1}, \cdots, 2^{6}\right\}$, $K \in \{1, 3, 5, 7, \cdots, 15\}$, $\gamma \in \left\{0, 2^{-6}, 2^{-5}, 2^{-4}, \cdots 2^{2}\right\}$. The values were chosen in such a way that minimizes macro-averaged kappa coefficient.

The experimental code was implemented using WEKA framework. The source code of the algorithms is available online [1].

To evaluate the proposed methods the following classification-loss criteria are used [30]: Macro-averaged FDR (1- precision), FNR (1-recall), F_1, Matthews correlation coefficient (MCC) ;Micro-averaged F_1, MCC. More quality measures from the macro-averaging group are considered because this kind of measures is more sensitive to the performance for minority classes.

Following the recommendations of [4] and [11], the statistical significance of the obtained results was assessed using the two-step procedure. The first step is to perform the Friedman test [4] for each quality criterion separately. Since the multiple criteria were employed, the familywise errors (FWER) should be controlled [2]. To do so, the Bergman-Hommel [2] procedure of controlling FWER of the conducted Friedman tests was employed. When the Friedman

[1] `https://github.com/ptrajdos/rrcBasedClassifiers/tree/develop`

test shows that there is a significant difference within the group of classifiers, the pairwise tests using the Wilcoxon signed-rank test [38,4] were employed. To control FWER of the Wilcoxon-testing procedure, the Bergman-Hommel approach was employed [2]. For all tests the significance level was set to $\alpha = 0.05$.

The experimental evaluation was conducted on the collection of the 78 benchmark datasets taken from the Keel repository containing imbalanced datasets with imbalance ratio higher than 9 [2].

During the preprocessing stage, the datasets underwent a few transformations. First, all nominal attributes were converted into a set of binary variables. The transformation is necessary whenever the distance-based algorithms are employed [33]. To reduce the computational burden and remove irrelevant information, the PCA procedure with the variance threshold set to 95% was applied [16]. The features were also normalized to have zero mean value and zero unit variance.

4 Results and Discussion

To compare multiple algorithms on multiple benchmark sets the average ranks approach [4] is used. To provide a visualization of the average ranks, the radar plots are employed. In the plots, the data is visualized in such way that the lowest ranks are closer to the centre of the graph. The radar plots related to the experimental results are shown in figures 1a – 1c.

Due to the page limit, the full results are published online [3]

The numerical results are given in Tables 2 – 4. Each table table is structured as follows. The first row of each section contains names of the investigated algorithms. Then the table is divided into six sections – one section is related to a single evaluation criterion. The first row of each section is the name of the quality criterion investigated in the section. The second row shows the p-value of the Friedman test. The third one shows the average ranks achieved by algorithms. The following rows show p-values resulting from pairwise Wilcoxon test. The p-value equal to 0.00 informs that the p-values are lower than 10^{-3} and p-value equal to 1.00 informs that the value is higher than 0.999.

4.1 Macro Averaged Criteria

Let us begin with the analysis of the results related to KNN neighbourhood. For the Naive Bayes and J48 classifiers, there are no significant differences between the Gaussian neighbourhood and KNN neighbourhood. For the Nearest Centroid classifier, on the other hand, the KNN neighbourhood gives better results in terms of FNR, F_1 and MCC. For the FDR criterion, there is no significant difference. It means that for ψ_{NC} classifier applying KNN neighbourhood

[2] `https://sci2s.ugr.es/keel/imbalanced.php#subB`

[3] `https://github.com/ptrajdos/MLResults/blob/master/RandomizedClassifiers/RRC_Imbalanced_CLDD2020.zip`

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_7

improves recall without affecting precision what results in better overall performance. What is more, for the J48 classifier, only the classifiers based on the KNN neighbourhood offers a significant improvement in terms of F_1 criterion.

Now the impact of applying the weighting scheme is assessed. Generally speaking, the application of the weighting scheme results in improving recall at the cost of reducing precision. However, in general, the reduction of precision is not significant (except for ψ_{NC} classifier and KNN approach). As a consequence, the overall classification quality, measured in terms of F_1 criterion remains unchanged (no significant difference). This kind of change is the expected consequence of applying the weighting scheme. On the other hand, in cases of J48 and NC classifiers, there are significant improvements in terms of MCC criterion. What is more, for the J48 classifier, only the classifiers based on the weighted neighbourhood offers a significant improvement in terms of MCC criterion.

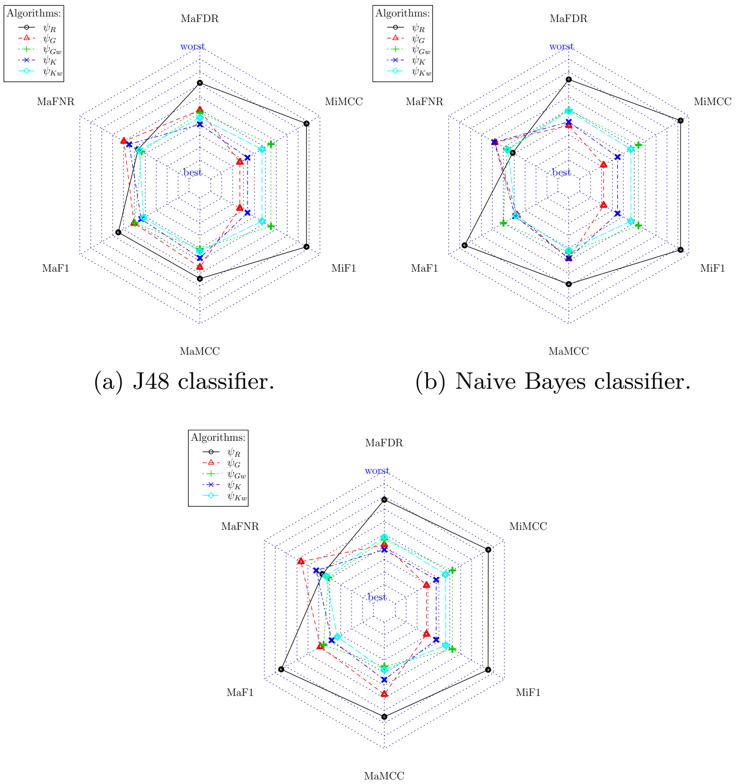
Now the correction ability of the SCM classifier is investigated. As it was said above, for the J48 base classifier, the overall correction ability depends on the type of the neighbourhood applied. For ψ_{NB} and ψ_{NC} classifiers, on the other hand, there is always a significant improvement in terms of F_1 and MCC. criteria. In general, the application of SCM classifier, when compared to the base classifier, improves the precision at the cost of decreasing recall. The recall-decrease is lower for the SCM classifiers using the weighted neighbourhood approach. So, applying the weighting scheme eliminates the main drawback of the SCM classifier used to the imbalanced data.

4.2 Micro Averaged Criteria

For the micro-averaged criteria, the statistical tests show that all differences are significant. Consequently, all investigated approaches improve the overall majority-class-performance in comparison to the unmodified base classifier. However, the classifiers with weighted neighbourhood show lower classification quality compared with classifiers that use no weights. This is an obvious consequence of trying to improve the performance for the minority class.

Table 2. Statistical evaluation. Wilcoxon test results for J48 classifier.

	Ψ_R	Ψ_G	Ψ_{Gw}	Ψ_K	Ψ_{Kw}	Ψ_R	Ψ_G	Ψ_{Gw}	Ψ_K	Ψ_{Kw}	Ψ_R	Ψ_G	Ψ_{Gw}	Ψ_K	Ψ_{Kw}
Nam.	MaFDR					MaFNR					MaF1				
Frd.	4.079e-06					3.984e-02					3.189e-03				
Rank	3.846	2.981	2.897	2.532	2.744	2.865	3.385	2.737	3.192	2.821	3.590	3.013	2.987	2.750	2.660
Ψ_R		.002	.000	.000	.000		.128	1.00	.128	1.00		.217	.217	.009	.003
Ψ_G			.933	.491	.491			.003	.952	.007			.985	.572	.206
Ψ_{Gw}				.491	.491				.022	1.00				.572	.217
Ψ_K					.491					.000					.217
Nam	MaMCC					MiF1					MiMCC				
Frd	1.494e-03					1.412e-24					1.412e-24				
Rnk	3.564	3.205	2.622	2.910	2.699	4.506	2.064	3.205	2.346	2.878	4.506	2.064	3.205	2.346	2.878
Ψ_R		.603	.009	.129	.026		.000	.004	.000	.000		.000	.004	.000	.000
Ψ_G			.010	.535	.045			.000	.005	.000			.000	.005	.000
Ψ_{Gw}				.173	.717				.000	.003				.000	.003
Ψ_K					.026					.000					.000



(a) J48 classifier.

(b) Naive Bayes classifier.

(c) Nearest Centroid classifier.

Fig. 1. Radar plots for the investigated classifiers.

5 Conclusions

This paper addresses the issue of tailoring the soft confusion matrix classifier to dealing with imbalanced data. Two concepts based on the change of the neighbourhood were proposed. The experimental results show that, in some circumstances, these approaches can improve the obtained classification quality. It shows that classifiers based on the RRC concept and SCM concept, in particular, are robust tools that can deal with various types of data. The other way of tailoring the SCM classifier to imbalanced data may be the modification of the $P(i|x)$ probability distribution. This aspect should be studied carefully.

Acknowledgments. This work was supported by the statutory funds of the Department of Systems and Computer Networks, Wroclaw University of Science and Technology.

Table 3. Statistical evaluation. Wilcoxon test results for NB classifier.

	Ψ_R	Ψ_G	Ψ_{Gw}	Ψ_K	Ψ_{Kw}	Ψ_R	Ψ_G	Ψ_{Gw}	Ψ_K	Ψ_{Kw}	Ψ_R	Ψ_G	Ψ_{Gw}	Ψ_K	Ψ_{Kw}
Nam.	MaFDR					MaFNR					MaF1				
Frd.	2.116e-08					1.697e-02					7.976e-18				
Rank	3.955	2.494	2.987	2.609	2.955	2.654	3.308	2.872	3.327	2.840	4.417	2.494	2.994	2.564	2.532
Ψ_R		.000	.000	.000	.000		.005	.392	.020	.392		.000	.000	.000	.000
Ψ_G			.103	.612	.501			.001	.511	.006			.197	.664	.664
Ψ_{Gw}				.096	.501				.019	.833				.091	.041
Ψ_K					.074					.006					.749
Nam.	MaMCC					MiF1					MiMCC				
Frd.	1.224e-04					1.226e-31					1.226e-31				
Rank	3.737	2.929	2.744	2.923	2.667	4.699	1.878	3.154	2.391	2.878	4.699	1.878	3.154	2.391	2.878
Ψ_R		.004	.000	.000	.000		.000	.000	.000	.000		.000	.000	.000	.000
Ψ_G			.145	.894	.300			.000	.002	.000			.000	.002	.000
Ψ_{Gw}				.397	.939				.000	.015				.000	.015
Ψ_K					.397					.000					.000

Table 4. Statistical evaluation. Wilcoxon test results for NC classifier.

	Ψ_R	Ψ_G	Ψ_{Gw}	Ψ_K	Ψ_{Kw}	Ψ_R	Ψ_G	Ψ_{Gw}	Ψ_K	Ψ_{Kw}	Ψ_R	Ψ_G	Ψ_{Gw}	Ψ_K	Ψ_{Kw}
Nam.	MaFDR					MaFNR					MaF1				
Frd.	3.149e-10					2.684e-04					4.529e-17				
Rank	4.090	2.667	2.827	2.506	2.910	2.865	3.654	2.647	3.096	2.737	4.378	2.942	2.827	2.532	2.321
Ψ_R		.000	.000	.000	.000		.057	.651	1.00	.651		.000	.000	.000	.000
Ψ_G			1.00	.470	1.00			.000	.004	.000			1.00	.043	.044
Ψ_{Gw}				.171	1.00				.225	1.00				.044	.043
Ψ_K					.043					.056					1.00
Nam.	MaMCC					MiF1					MiMCC				
Frd.	5.340e-11					4.793e-20					4.001e-20				
Rank	3.994	3.282	2.397	2.821	2.506	4.404	2.147	3.096	2.500	2.853	4.410	2.147	3.096	2.500	2.846
Ψ_R		.000	.000	.000	.000		.000	.000	.000	.000		.000	.000	.000	.000
Ψ_G			.000	.028	.010			.000	.010	.000			.000	.010	.000
Ψ_{Gw}				.310	.979				.001	.010				.001	.010
Ψ_K					.310					.001					.001

References

1. Ali, A., Shamsuddin, S.M., Ralescu, A.L., et al.: Classification with class imbalance problem: A review. Int. J. Advance Soft Compu. Appl **7**(3), 176–204 (2015)
2. Bergmann, B., Hommel, G.: Improvements of general multiple test procedures for redundant systems of hypotheses. In: Multiple Hypothesenprüfung / Multiple Hypotheses Testing, pp. 100–115. Springer Berlin Heidelberg (1988). https://doi.org/10.1007/978-3-642-52307-6_8
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. jair **16**, 321–357 (Jun 2002). https://doi.org/10.1613/jair.953
4. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research **7**, 1–30 (2006)
5. Dhar, M.: On cardinality of fuzzy sets. IJISA **5**(6), 47–52 (May 2013). https://doi.org/10.5815/ijisa.2013.06.06
6. Duda, R.: Pattern classification. Wiley, New York (2001)
7. Fernández, A., del Jesus, M.J., Herrera, F.: Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. Int. J. Approximate Reasoning **50**(3), 561–577 (Mar 2009). https://doi.org/10.1016/j.ijar.2008.11.004
8. Fu, K., Cheng, D., Tu, Y., Zhang, L.: Credit card fraud detection using convolutional neural networks. In: International Conference on Neural Information Processing. pp. 483–490. Springer (2016)
9. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst., Man, Cybern. C **42**(4), 463–484 (Jul 2012). https://doi.org/10.1109/tsmcc.2011.2161285
10. Galar, M., Fernández, A., Barrenechea, E., Herrera, F.: EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern Recognit. **46**(12), 3460–3471 (Dec 2013). https://doi.org/10.1016/j.patcog.2013.05.006

11. Garcia, S., Herrera, F.: An extension on“statistical comparisons of classifiers over multiple data sets”for all pairwise comparisons. Journal of Machine Learning Research **9**, 2677–2694 (Dec 2008)
12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software. SIGKDD Explor. Newsl. **11**(1), 10 (Nov 2009). https://doi.org/10.1145/1656274.1656278
13. Hand, D.J., Yu, K.: Idiot's bayes: Not so stupid after all? International Statistical Review / Revue Internationale de Statistique **69**(3), 385 (Dec 2001). https://doi.org/10.2307/1403452
14. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). pp. 1322–1328. IEEE, IEEE (Jun 2008). https://doi.org/10.1109/ijcnn.2008.4633969
15. Hwang, J.P., Park, S., Kim, E.: A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function. Expert Syst. Appl. **38**(7), 8580–8585 (Jul 2011). https://doi.org/10.1016/j.eswa.2011.01.061
16. Jolliffe, I.T., Cadima, J.: Principal component analysis: A review and recent developments. Phil. Trans. R. Soc. A **374**(2065), 20150202 (Apr 2016). https://doi.org/10.1098/rsta.2015.0202
17. Kaur, H., Pannu, H.S., Malhi, A.K.: A systematic review on imbalanced data challenges in machine learning. CSUR **52**(4), 1–36 (Aug 2019). https://doi.org/10.1145/3343440
18. Khan, S.H., Hayat, M., Bennamoun, M., Sohel, F.A., Togneri, R.: Cost-sensitive learning of deep feature representations from imbalanced data. IEEE Trans. Neural Netw. Learning Syst. **29**(8), 3573–3587 (Aug 2018). https://doi.org/10.1109/tnnls.2017.2732482
19. Krawczyk, B., Galar, M., Jelen, L., Herrera, F.: Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Appl. Soft Comput. **38**, 714–726 (Jan 2016). https://doi.org/10.1016/j.asoc.2015.08.060
20. Krawczyk, B., Woźniak, M., Schaefer, G.: Cost-sensitive decision tree ensembles for effective imbalanced classification. Appl. Soft Comput. **14**, 554–562 (Jan 2014). https://doi.org/10.1016/j.asoc.2013.08.014
21. Kuncheva, L., Bezdek, J.: Nearest prototype classification: clustering, genetic algorithms, or random search? IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) **28**(1), 160–164 (1998). https://doi.org/10.1109/5326.661099
22. Kurzynski, M., Krysmann, M., Trajdos, P., Wolczowski, A.: Multiclassifier system with hybrid learning applied to the control of bioprosthetic hand. Comput. Biol. Med. **69**, 286–297 (Feb 2016). https://doi.org/10.1016/j.compbiomed.2015.04.023
23. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences **250**, 113–141 (Nov 2013). https://doi.org/10.1016/j.ins.2013.07.007
24. Park, Y., Ghosh, J.: Ensembles of (α)-trees for imbalanced classification problems. IEEE Trans. Knowl. Data Eng. **26**(1), 131–143 (Jan 2014). https://doi.org/10.1109/tkde.2012.255
25. Patel, H., Thakur, G.: A hybrid weighted nearest neighbor approach to mine imbalanced data. In: Proceedings of the International Conference on Data Mining (DMIN). pp. 106–110. The Steering Committee of The World Congress in Computer Science, Computer ...(2016)

26. Provost, F., Domingos, P.: Tree induction for probability-based ranking. Machine Learning **52**(3), 199–215 (Sep 2003). https://doi.org/10.1023/a:1024099825458
27. Quinlan, J.R.: C4.5 : Programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
28. Ramyachitra, D., Manikandan, P.: Imbalanced dataset classification and solutions: A review. International Journal of Computing and Business Research (IJCBR) **5**(4), 186–194 (2014)
29. Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Trans. Syst., Man, Cybern. A **40**(1), 185–197 (Jan 2010). https://doi.org/10.1109/tsmca.2009.2029559
30. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4) (Jul 2009). https://doi.org/10.1016/j.ipm.2009.03.002
31. Sun, Y., Kamel, M.S., Wong, A.K., Wang, Y.: Cost-sensitive boosting for classification of imbalanced data. Pattern Recognit. **40**(12), 3358–3378 (Dec 2007). https://doi.org/10.1016/j.patcog.2007.04.009
32. Sun, Y., Wong, A.K.C., Kamel, M.S.: Classification of imbalanced data: A review. Int. J. Patt. Recogn. Artif. Intell. **23**(04), 687–719 (Jun 2009). https://doi.org/10.1142/s0218001409007326
33. Tian, Y., Deng, N.: Support vector classification with nominal attributes. In: Hao, Y., Liu, J., Wang, Y., Cheung, Y.m., Yin, H., Jiao, L., Ma, J., Jiao, Y.C. (eds.) Computational Intelligence and Security, pp. 586–591. Springer Berlin Heidelberg (2005). https://doi.org/10.1007/11596448_86
34. Trajdos, P., Kurzynski, M.: A dynamic model of classifier competence based on the local fuzzy confusion matrix and the random reference classifier. International Journal of Applied Mathematics and Computer Science **26**(1) (jan 2016). https://doi.org/10.1515/amcs-2016-0012
35. Trajdos, P., Kurzynski, M.: A correction method of a binary classifier applied to multi-label pairwise models. Int. J. Neur. Syst. **28**(09), 1750062 (Sep 2018). https://doi.org/10.1142/s0129065717500629
36. Trajdos, P., Kurzynski, M.: Weighting scheme for a pairwise multi-label classifier based on the fuzzy confusion matrix. Pattern Recognit. Lett. **103**, 60–67 (Feb 2018). https://doi.org/10.1016/j.patrec.2018.01.012
37. Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. In: 2009 IEEE Symposium on Computational Intelligence and Data Mining. pp. 324–331. IEEE, IEEE (Mar 2009). https://doi.org/10.1109/cidm.2009.4938667, `https://doi.org/10.1109/cidm.2009.4938667`
38. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6), 80 (Dec 1945). https://doi.org/10.2307/3001968
39. Woloszynski, T., Kurzynski, M.: A probabilistic model of classifier competence for dynamic ensemble selection. Pattern Recognition **44**(10-11), 2656–2668 (Oct 2011). https://doi.org/10.1016/j.patcog.2011.03.020
40. Wu, Q., Ye, Y., Zhang, H., Ng, M.K., Ho, S.S.: ForesTexter: An efficient random forest algorithm for imbalanced text categorization. Knowledge-Based Systems **67**, 105–116 (Sep 2014). https://doi.org/10.1016/j.knosys.2014.06.004
41. Yu, H., Ni, J., Zhao, J.: ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. Neurocomputing **101**, 309–318 (Feb 2013). https://doi.org/10.1016/j.neucom.2012.08.018
42. Zheng, Z., Cai, Y., Li, Y.: Oversampling method for imbalanced classification. Computing and Informatics **34**(5), 1017–1037 (2016)