Clustering and Weighted Scoring Algorithm Based on Estimating the Number of Clusters

Jakub Klikowski[0000−0002−3825−5514] and Robert Burduk[0000−0002−3506−6611]

Department of Systems and Computer Networks
Wroclaw University of Science and Technology
Wroclaw, Poland
{jakub.klikowski, robert.burduk}@pwr.edu.pl

Abstract. Imbalanced datasets are still a big method challenge in data mining and machine learning. Various machine learning methods and their combinations are considered to improve the quality of the classification of imbalanced datasets. This paper presents the approach with the clustering and weighted scoring function based on geometric space are used. In particular, we proposed a significant modification to our earlier algorithm. The proposed change concerns the use of automatic estimating the number of clusters and determining the minimum number of objects in a particular cluster. The proposed algorithm was compared with our earlier proposal and state-of-the-art algorithms using highly imbalanced datasets. The performed experiments show that the proposed modification is statistically better for a larger number of reference classifiers than the original algorithm.

Keywords: Imbalanced Data · Ensemble of classifiers · Class imbalance · Decision boundary · Scoring function.

1 Introduction

Machine learning methods can be divided into several groups, which include, among others, supervised learning, unsupervised learning, association rules, or time series analysis. It is vital for supervised learning to have data for which the class labels are known. In real problems, the number of objects in each class rarely is the same. If the imbalanced ratio expressed as the majority class's quotient to the minority class is much greater than 1, we deal with imbalanced data. Such data is also called skew data. Many practical problems concern imbalanced data because they arise directly from the problem's characteristics and the available training data [10]. Examples of practical applications where there are skews include: network intrusion detection [2, 14], source code fault detection [6], or in general fraud detection [1].

In the supervised classification of skew data, methods belonging to two main trends are used. These are data-level [7, 13, 25] and algorithm-level [28] methods. The data-level methods use a resampling process that can be performed by

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_4

oversampling, undersampling, and hybrid in nature. The algorithm-level methods concern the modification of known machine learning algorithms to increase minority class classification performance [9, 12, 16].

In this article, we consider the algorithm-level approach. In particular, we present a significant modification of our previous algorithm presented in [17]. The proposed modification uses the Silhouette Value [23] to estimate the number of clusters automatically. Additionally, we took into account the number of necessary objects to designate one cluster. Taking the above into account, the main objectives of this work are summarized as follows:

- A proposal of a new clustering and weighted scoring algorithm based on estimating the number of clusters.
- The proposed algorithm has considered the minimum number of objects in each cluster.
- A new experimental setup on highly imbalanced datasets compares the proposed algorithm with the previous one and other state-of-the-art algorithms for supervised classification.

The paper is structured as follows: the Section 2 introduces the base concept of ensemble of classifiers and presents the proposed algorithm. In the Section 3 the experiments that were carried out are presented, while results and the discussion appear in the Section 4. Finally, we conclude the paper in the Section 5.

2 Proposed Method

2.1 Ensemble of classifiers

The ensemble of classifiers (EoC) is widely discussed in the literature to solve the problem of skew data [9, 15, 20]. The use of the EoC belongs to the algorithm-level approach to solving the imbalanced data problem.

In general, the idea of EoC determination is to build a predictive model by integrating multiple base classification models $\Psi_1, \Psi_2, \ldots, \Psi_L$, where L is the number of classifiers in the EoC. The procedure for creating an EoC can be divided into three major steps: generation – a phase where individual classifiers are trained, selection – a phase where only a few (or even one) individual models from the previous step are selected for inclusion in the EoC and combining the base classifier outputs.

The idea presented in the following article is the construction of the EoC, diversified by the disjoint division of problem classes into clusters, introducing the integration rule based on the geometric characteristics of its components.

The first of the two steps necessary to build an effective EoC is the selection of models for its pool [19], required to make as independent decisions as possible. The strategy adopted in the discussed method is the use of the homogeneous [27] ensemble, built on the basis of linear classifiers, where each model is learned from a combination of [8] class clusters determined using the *K-Means* [5] algorithm.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_4

2.2 Clustering and Weighted Scoring Algorithm Based on Estimating the Number of Clusters

This paper presents a certain extension of the CWS method developed by Ksieniewicz and Burduk [17]. The implemented changes focus on extending this approach's main idea and expanding the research to a larger pool of datasets. To better justify the new proposed algorithm, the whole procedure will be described step by step.

The Clustering and Weighted Scoring with Estimating the Number of Clusters *CWS-ENC* is based on an approach that uses the original procedure to determine objects' score function. The value obtained depends heavily on the position of object x in the geometric space. The scoring function [17] is expressed by the equation 1:

$$wsf_l(x) = 1 - \frac{sf_l(x)}{\sum_{l=1}^{L} sf_l(x)}, \tag{1}$$

Where one of the main components is $sf_l(x)$. This is a function used to determine the distance from the decision boundaries of models (Ψ_l) forming EoC to clusters' centroids. Clusters are created with the K-means algorithm. The procedure to calculate the distance is described by the equation 2, where $\|\Psi_l(x)\|$ is result of base classifier (Ψ_l) decision function on object x, C is the number of clusters, and d_c means the distance from the object x to clusters' centroids expressed by any distance metric. Preliminary experiments have shown that the method gets the best results using the "Manhattan" distance metric.

$$sf_l(x) = \|\Psi_l(x)\| + \sum_{c=1}^{C} d_c, \tag{2}$$

An essential new change of this algorithm is forming clusters (Algorithm 1). The original idea [17] assumed that the number of clusters is chosen arbitrarily or experimentally. However, this requires additional preparation before starting a new experiment or assumes that it is optimal for all datasets. The proposed solution implies that the number of clusters should be selected dynamically based on the clusters' evaluation consistency using the Silhouette Coefficient [11] metric. This is the maximum value of the mean Silhouette Value $sv(x)$ [23] over the entire dataset and is described by the equation 3:

$$sv(x) = \frac{b(x) - a(x)}{max\{a(x), b(x)\}}, \tag{3}$$

Where x is an object from cluster C, $b(x)$ is a mean dissimilarity of x to cluster C and $a(x)$ is a mean dissimilarity of x to other clusters. In practice, this means that the different number of clusters in the range from 2 to K_{max} are created and evaluated using the metric Silhouette Coefficient. Then the most optimal option is selected. In the performed preexperiments, the value of K_{max} was set as 5.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_4

Moreover, a particular heuristic was introduced for minority class clusters. This rule blocks the possibility of creating new clusters when the currently divided set of samples does not reach the threshold for the minimum number of objects. This avoids situations where the algorithm tries to divide several objects from the minority class into multiple clusters in the strongly imbalanced dataset. Such an enforced segmentation only increases the imbalance ratio, which results in the deterioration of the classifier's predictive ability. This threshold's value is a parameter of the method and was set at 25 samples in the preexperiments.

The procedure of the ENC-CWS method is described in more detail by the Algorithm 1. In the first step, the whole dataset is divided into subsets D_i composed by objects from one class. Then for each subset D_i the quantity of the objects is checked, when this number is less than the predetermined threshold S_{min} then the algorithm creates only one cluster for this data. When the number of objects is more, then follows the procedure to determine the correct number of clusters. Using the K-means method, the data is clustered where the cluster number M_i changes from 2 to the predetermined K_{max}. For each set of clusters, consistency is measured using the Silhouette Coefficient metric. Then, the setup with K^{M_i} clusters that obtains the best score is selected and these clusters are stored. Next, the centroids of the created clusters are determined. After this the learning of new models with base classifier is performed. It is done using combinations of clusters in one-to-one manner, but from different classes. Then a weighted scoring function is computed for each sample.

3 Experimental evaluation

The experimental analysis aims to verify the predictive perfomance of the *CWS-ENC* method for imbalance problems. In the following, the research questions are formulated:

RQ1: How does a method employing the idea of weighted score function based on objects geometric position handle datasets with varying imbalance levels in a comparison to the selected approaches?

RQ2: Do the modifications made bring improvements over the original version of the method?

3.1 Setup

The experimental evaluation was implemented in the Python programming language. Some elements from the *scikit-learn* [21] and *stream-learn* [18] libraries were used to perform the experiments. The project implementation with the results is available on the GitHub code repository [1]. The conducted experimental analysis aims to verify whether the introduced modifications will improve classification quality compared to the previous method variant. In the following is the list of approaches that were compared with the proposed method:

[1] Repository link: https://github.com/w4k2/cws-enc

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_4

Algorithm 1: CWS-ENC – for binary problem

Input: D – Learning set
x – object
K_{max} – maximum number of clusters
S_{min} – minimum number of samples

Output: The ensemble classifier decision

1 Divide D into D_i subsets of data, where $i \in 1, 2$ is the number of class labels.
2 If size of D_i is greater than S_{min} determine the best number of clusters M_i from 2 to K_{max} using the Silhouette Coefficient metric for each class. Otherwise M_i is equal 1.
3 Divide D_i into K^{M_i} clusters using the K-means clustering algorithm separately for each $i-th$ class.
4 Find the cluster centroids $C_1^i, \ldots, C_{M_i}^i$ as the means of the points in the respective clusters.
5 Train base classifier $\Psi_1, \ldots, \Psi_L$ using each combination of clusters from different class labels, i.e. one cluster from each class label, $L = M_1 * M_2$.
6 Calculate weighted scoring functions for the object x:

$$wsf_l(x) = 1 - \frac{sf_l(x)}{\sum_{l=1}^{L} sf_l(x)},$$

where

$$sf_l(x) = \|\Psi_l(x)\| + \sum_{c=1}^{2} d_c.$$

7 The ensemble classifier decision:

$$\widehat{\Psi}(x) = sign\left(\sum_{l=1}^{L} wsf_l(x)\Psi_l(x)\right),$$

where $\Psi(x)$ is the prediction returned by base classifier $\Psi(x) \in \{-1, 1\}$.

- **CWS-ENC** (*Clustering and Weighted Scoring with Estimating the Number of Clusters*) — *EoC* proposed in this work and explained in Section 2.
- **CWS** (*Clustering and Weighted Scoring*) — *EoC* with the pool diversified by pairs of clusters and integrated geometrically by the rules proposed by Ksieniewicz and Burduk [17].
- **SVC** (*Support Vector Machine*) — the base model with the scaled gamma and linear kernel [22].
- **CMV** (*Clustering and Majority Vote*) — *EoC* identical with *CWS* but integrated using the majority vote [24].
- **CSA** (*Clustering and Support Accumulation*) — *EoC* identical with *CWS* and *CMV* but integrated using the support accumulation rule [26].

The testing procedure consist in evaluating datasets with the *Stratified K-fold Crossvalidation*, where the K is equal 5. The classification quality was expressed

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_4

in the form of six metrics - *balanced accuracy score* (BAC), *F1-score* (F-1), *G-mean* (GMN), *precision* (PRE), *recall* (REC) and *specificity* (SPE). Next, for the obtained results, statistical analysis was performed using the *Wilcoxon rank test* with the significance level $alpha = 0.05$ [4]. 58 imbalanced binary datasets were used to conduct the study, which are described in Table 1.

Table 1. Overview of real datasets used in experimental evaluation (KEEL [3])

Dataset name	IMB. RATIO	SAMPLES	FEATURES	Dataset name	IMB. RATIO	SAMPLES	FEATURES
abalone-21_vs_8	40	581	8	*glass4*	15	214	9
abalone-3_vs_11	32	502	8	*glass5*	23	214	9
abalone9-18	16	731	8	*glass6*	6.4	214	9
cleveland-0_vs_4	13	177	13	*led7digit-0-2-4-5-6-7-8-9_vs_1*	11	443	7
dermatology-6	17	358	34	*lymphography-normal-fibrosis*	24	148	18
ecoli-0-1-3-7_vs_2-6	39	281	7	*new-thyroid1*	5.1	215	5
ecoli-0-1-4-6_vs_5	13	280	6	*newthyroid2*	5.1	215	5
ecoli-0-1-4-7_vs_2-3-5-6	11	336	7	*poker-9_vs_7*	30	244	10
ecoli-0-1-4-7_vs_5-6	12	332	6	*shuttle-6_vs_2-3*	22	230	9
ecoli-0-1_vs_2-3-5	9.2	244	7	*shuttle-c2-vs-c4*	20	129	9
ecoli-0-1_vs_5	11	240	6	*vowel0*	10	988	13
ecoli-0-2-3-4_vs_5	9.1	202	7	*winequality-red-3_vs_5*	68	691	11
ecoli-0-2-6-7_vs_3-5	9.2	224	7	*winequality-red-8_vs_6*	35	656	11
ecoli-0-3-4-6_vs_5	9.2	205	7	*winequality-red-8_vs_6-7*	46	855	11
ecoli-0-3-4-7_vs_5-6	9.3	257	7	*winequality-white-3_vs_7*	44	900	11
ecoli-0-3-4_vs_5	9	200	7	*winequality-white-9_vs_4*	33	168	11
ecoli-0-4-6_vs_5	9.2	203	6	*yeast-0-2-5-6_vs_3-7-8-9*	9.1	1004	8
ecoli-0-6-7_vs_3-5	9.1	222	7	*yeast-0-2-5-7-9_vs_3-6-8*	9.1	1004	8
ecoli-0-6-7_vs_5	10	220	6	*yeast-0-3-5-9_vs_7-8*	9.1	506	8
ecoli2	5.5	336	7	*yeast-0-5-6-7-9_vs_4*	9.4	528	8
ecoli3	8.6	336	7	*yeast-1-2-8-9_vs_7*	31	947	8
ecoli4	16	336	7	*yeast-1-4-5-8_vs_7*	22	693	8
glass-0-1-4-6_vs_2	11	205	9	*yeast-1_vs_7*	14	459	7
glass-0-1-5_vs_2	9.1	172	9	*yeast-2_vs_4*	9.1	514	8
glass-0-1-6_vs_2	10	192	9	*yeast-2_vs_8*	23	482	8
glass-0-1-6_vs_5	19	184	9	*yeast3*	8.1	1484	8
glass-0-4_vs_5	9.2	92	9	*yeast4*	28	1484	8
glass-0-6_vs_5	11	108	9	*yeast5*	33	1484	8
glass2	12	214	9	*yeast6*	41	1484	8

4 Results

The obtained results are presented in a Table 2, on which the exact values of the mean ranks and advantages with statistical significance are printed. Some numbers indicate that the method performance is statistically better under the rank values than the other methods. It can be easily seen that the proposed approach obtains statistical superiority over the methods *SVC*, *CMV* and *CSA* for most of the metrics. The exceptions are precision and specificity.

Much better readability of the average results is presented by the radar plot showing the graphical form results. The advantage in the scores obtained for the *CWS-ENC* method is easily seen here. All rankings for this method except the specificity metric are more or less better. There is also a noticeable improvement in quality over the method without the proposed modifications. Unfortunately, this advantage does not achieve statistical significance.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_4

Table 2. Results for mean ranks and statistical significance

	CWS-ENC (1)	**CWS (2)**	**SVC (3)**	**CMV (4)**	**CSA (5)**
BAC	3.724	3.379	2.974	2.241	2.681
	3,4,5	4,5	4	–	–
F-1	3.603	3.319	3.034	2.336	2.707
	3,4,5	4,5	4	–	–
GMN	3.655	3.336	2.810	2.397	2.802
	3,4,5	3,4	–	–	–
REC	3.586	3.457	2.586	2.595	2.776
	3,4,5	3,4,5	–	–	–
PRE	3.267	3.138	3.198	2.578	2.819
	4	4	4	–	–
SPE	2.716	2.681	3.776	2.672	3.155
	–	–	*all*	–	4

The proposed method performs poorly for the specificity metric compared to the others. The *SVC* dominates in this metric and has the best result with a statistically significant advantage over the rest of the methods. However, it is essential to note that the strong ability to classify majority class data is associated with a high decrease in the recall metric and slightly for BAC, F-1, and GMN metrics. This is a typical performance of a method that predicts too much bias toward the majority class when dealing with imbalanced data. Overall, the *CMV* approach received the weakest result.

4.1 Lessons learned

In summary, the research questions stated above will be answered:

RQ1: How does a method employing the idea of weighted score function based on objects geometric position handle datasets with varying imbalance levels?

Tests performed on 58 datasets whose imbalance level varies between 5 and 68 allows for a substantial study of binary imbalanced problems. The obtained results and their statistical analysis show that the weighted score function based on objects' geometric position is the right solution for imbalanced data classification. The presented method improves the classification quality expressed in different metrics compared to the *CMV* or *CSA* methods. For minority class and aggregate metrics, this advantage is statistically significant.

RQ2: Do the modifications made bring improvements over the original version of the method?

The implemented changes bring a visible improvement in the obtained results. The analysis of ranking tests shows that the proposed modification to dynamically select the number of clusters and threshold for the minimum number of samples achieves predictive performance better than the original approach.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_4

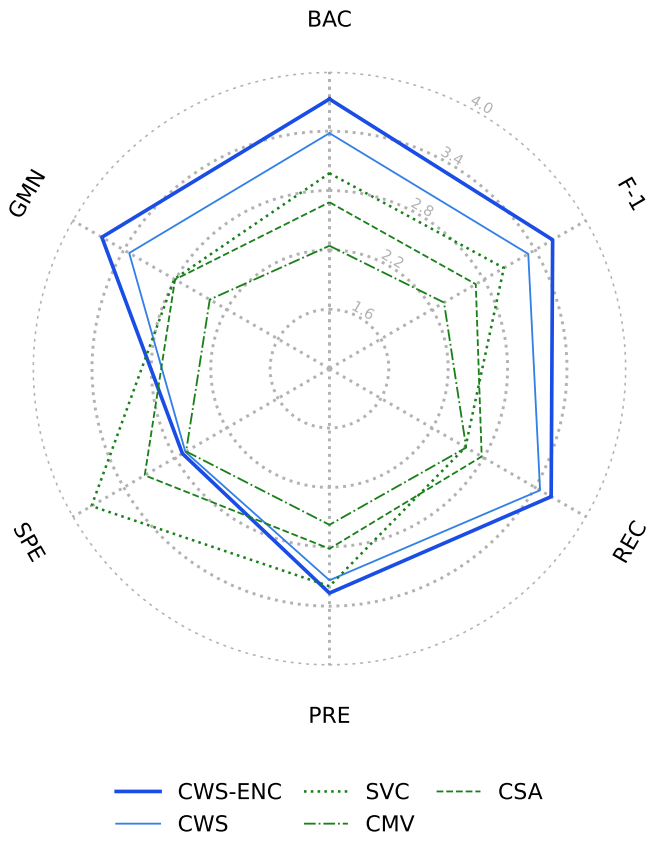

Fig. 1. Radar plot of mean ranks obtained by the Friedman test.

Unfortunately, the advantage does not have statistical significance, which may be because, despite some adjustments, the methods have many common traits.

5 Conclusions

In this article was proposed a new method, based on an existing approach [17]. Introduced changes brought a noticeable performance improvement. Extended testing on a larger collection of imbalanced datasets and statistical analysis showed the presented method's good classification quality. The proposed algorithm modification significantly statistically improves minority class classification performance. In the datasets used, the majority class was marked in the confusion matrix as a negative class and the minority class as a positive class. The algorithm proposed in the article increases the value of the classification quality measure, which is REC, and reduces the value of SPE. Changes in these two measures' values, expressed as the statistical test's mean ranks indicate that

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_4

the proposed algorithm identifies objects from the minority class more accurately. The results obtained has significant potential for further development and broader research on the imbalanced dataset.

Future work in the following directions is worth considering:

- Perform experiments for multi-class problems.
- Use more and different linear base classifiers for testing.

Acknowledgements

This work was supported by the Polish National Science Centre under the grant No. 2017/25/B/ST6/01750 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. Abdallah, A., Maarof, M.A., Zainal, A.: Fraud detection system: A survey. Journal of Network and Computer Applications **68**, 90–113 (2016)
2. Abdulhammed, R., Faezipour, M., Abuzneid, A., AbuMallouh, A.: Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic. IEEE sensors letters **3**(1), 1–4 (2018)
3. 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 & Soft Computing **17** (2011)
4. Alpaydin, E.: Introduction to machine learning. MIT press (2014)
5. Basu, S., Banerjee, A., Mooney, R.: Semi-supervised clustering by seeding. In: In Proceedings of 19th International Conference on Machine Learning (ICML-2002. Citeseer (2002)
6. Choraś, M., Pawlicki, M., Kozik, R.: Recognizing faults in software related difficult data. In: International Conference on Computational Science. pp. 263–272. Springer (2019)
7. Fotouhi, S., Asadi, S., Kattan, M.W.: A comprehensive data level analysis for cancer diagnosis on imbalanced data. Journal of biomedical informatics **90**, 103089 (2019)
8. Fred, A., Lourenço, A.: Cluster ensemble methods: from single clusterings to combined solutions. In: Supervised and unsupervised ensemble methods and their applications, pp. 3–30. Springer (2008)
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 Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **42**(4), 463–484 (2011)
10. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications **73**, 220–239 (2017)
11. Kaufmann, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis. New York: Jonh Wiley (1990)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_4

12. Klikowski, J., Ksieniewicz, P., Woźniak, M.: A genetic-based ensemble learning applied to imbalanced data classification. In: International Conference on Intelligent Data Engineering and Automated Learning. pp. 340–352. Springer (2019)
13. Koziarski, M., Woźniak, M., Krawczyk, B.: Combined cleaning and resampling algorithm for multi-class imbalanced data with label noise. arXiv preprint arXiv:2004.03406 (2020)
14. Kozik, R., Choras, M., Keller, J.: Balanced efficient lifelong learning (b-ella) for cyber attack detection. J. UCS **25**(1), 2–15 (2019)
15. Krawczyk, B., Woźniak, M.: Leveraging ensemble pruning for imbalanced data classification. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 439–444. IEEE (2018)
16. Krawczyk, B., Woźniak, M., Schaefer, G.: Cost-sensitive decision tree ensembles for effective imbalanced classification. Applied Soft Computing **14**, 554–562 (2014)
17. Ksieniewicz, P., Burduk, R.: Clustering and weighted scoring in geometric space support vector machine ensemble for highly imbalanced data classification. In: International Conference on Computational Science. pp. 128–140. Springer (2020)
18. Ksieniewicz, P., Zyblewski, P.: stream-learn–open-source python library for difficult data stream batch analysis. arXiv preprint arXiv:2001.11077 (2020)
19. Kuncheva, L.I.: Combining pattern classifiers: methods and algorithms. John Wiley & Sons (2004)
20. Lopez-Garcia, P., Masegosa, A.D., Osaba, E., Onieva, E., Perallos, A.: Ensemble classification for imbalanced data based on feature space partitioning and hybrid metaheuristics. Applied Intelligence **49**(8), 2807–2822 (2019)
21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
22. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: ADVANCES IN LARGE MARGIN CLASSIFIERS. pp. 61–74. MIT Press (1999)
23. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
24. Ruta, D., Gabrys, B.: Classifier selection for majority voting. Information fusion **6**(1), 63–81 (2005)
25. Szeszko, P., Topczewska, M.: Empirical assessment of performance measures for preprocessing moments in imbalanced data classification problem. In: IFIP International Conference on Computer Information Systems and Industrial Management. pp. 183–194. Springer (2016)
26. Woźniak, M.: Hybrid classifiers: methods of data, knowledge, and classifier combination, vol. 519. Springer (2013)
27. Woźniak, M., Graña, M., Corchado, E.: A survey of multiple classifier systems as hybrid systems. Information Fusion **16**, 3–17 (2014)
28. Zhang, C., Bi, J., Xu, S., Ramentol, E., Fan, G., Qiao, B., Fujita, H.: Multi-imbalance: An open-source software for multi-class imbalance learning. Knowledge-Based Systems **174**, 137–143 (2019)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_4