Improved Sample Type Identification for Multi-Class Imbalanced Classification with Real-World Applications *

Jiawen Kong[1], Wojtek Kowalczyk[1], Kees Jonker[2], Stefan Menzel[3] and Thomas Bäck[1]

[1] Leiden University, Leiden, the Netherlands
{j.kong, w.j.kowalczyk, t.h.w.baeck}@liacs.leidenuniv.nl
[2] R&D, TATA Steel Europe, Ljmuiden, the Netherlands
kees.jonker@tatasteeleurope.com
[3] Honda Research Institute Europe GmbH, Offenbach, Germany
{stefan.menzel}@honda-ri.de

Abstract. Driven by studying the nature of imbalanced data, researchers proposed to consider different types of samples (safe, borderline, rare samples and outliers) in the minority class. The idea was first proposed and evaluated on binary imbalanced classification problems and then extended to multi-class scenarios. However, simply extending the identification rule in binary scenarios to multi-class scenarios results in several problems, for example, a higher percentage of unsafe samples in minority classes and a false identification of outliers. In this paper, we first show the drawbacks when extending this idea from binary to multi-class scenarios. Then, we propose a new identification rule for multi-class scenarios. In our experiments, we consider oversampling different types of samples before performing classification, where oversampling is a data-level approach to deal with the imbalance in the datasets. Experimental results on benchmark datasets indicate that the proposed rule can decrease the probability of false identification and improve the classification performance on minority class(es) on average by 7.4%. In addition, we apply our proposed rule on surface inspection data from the steel industry and confirm its effectiveness and potential usefulness in real-world applications.

Keywords: Multi-class imbalanced learning · Oversampling · Types of samples · Surface inspection data.

1 Introduction

Despite the progress for several years, learning from imbalanced data is still a challenging problem in machine learning. Solving imbalanced classification problems refers to the predictive modelling of data comprising a high or even extreme

* This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 766186.

imbalance in the sample distribution. Since machine learning models assume that the sample distribution is relatively balanced, the nature of imbalanced data violates this assumption, thus the class imbalance is commonly considered the determinant factor for the degradation of classification performance [5, 6]. However, several studies in the literature have pointed out that the data characteristics also play a crucial role in dealing with imbalanced problems [14, 17, 19]. Here, Napierala and Stefanowski proposed to consider minority samples consisting of four types of samples: *safe, borderline, rare* samples and *outliers* [16], the latter three are called *unsafe* samples. The types of minority samples in binary problems are determined by the neighborhood information:

- **safe** samples: the majority of the neighbors belongs to the same class;
- **borderline** samples: the proportion of the neighbors in both classes is approximately the same;
- **rare** samples: the majority of the neighbors belongs to a different class;
- **outliers**: all the neighbors belongs to a different class.

They studied the influence of these four types of samples on binary imbalanced classification, where the datasets are composed of two classes and one class significantly outnumbers the other. Other researchers then extended this idea to develop new techniques to improve imbalanced classification in both binary and multi-class scenarios [9, 12, 13]. However, the relationships among classes are more complicated in multi-class scenarios since there are more than two classes in the datasets. Simply extending the idea of four types of samples from binary to multi-class scenarios without changing the identification rule will cause several problems.

In this paper, we first introduce the identification rule for the four types of samples as proposed in the literature [16]. Then, we show the drawbacks when applying this identification rule to multi-class scenarios and emphasize the importance of proposing a new identification rule for multi-class scenarios. We find mainly two drawbacks: (1) a higher percentage of unsafe (*borderline, rare* and *outliers*) samples and (2) false identification of *outliers*. As a consequence, we propose a new identification rule for the four types of samples to handle the drawbacks mentioned above and validate the effectiveness of the new rule with benchmark datasets. In these experiments, we consider oversampling different types of samples before performing the classification, where oversampling is a data-level approaches to handle the imbalance in the datasets. Experimental results on benchmark and real-world data show that the proposed rule can significantly improve the classification performance on minority class(es) when a high imbalance exists in the datasets.

Class imbalance is present in many real-world classification tasks, for instance, medical diagnosis [15], email filtering [2], fault diagnosis [11], etc. Most of class imbalance applications in the literature have been devoted to binary classification problems. Most of the multi-class imbalanced benchmark datasets contain only a small number of attributes and a limited number of samples [1, 4]. Therefore, this paper makes an additional contribution by introducing a challenging industrial surface inspection dataset, with 172 attributes, 27 classes and

12496 samples. Experimental results on this industrial dataset also confirm the effectiveness and usefulness of our proposed rule for real-world applications.

The remainder of this paper is organized as follows: In Section 2, the definition of the class imbalance problem and an oversampling technique, SMOTE, is described. In Section 3, the concept of pre-classifying data sample types, usually denoted as "types of samples", and corresponding existing work is presented, including the rule for identifying types of samples in binary scenarios and drawbacks when extending the rule without editing to multi-class scenarios. In Section 4, the improved identification rule for multi-class scenarios is presented. The experimental setup and results on benchmark datasets and one real-world dataset are introduced in Section 5. Section 6 concludes the paper and outlines further research.

2 Class Imbalance Learning

Strictly speaking, any dataset with an unequal class distribution can be considered imbalanced. However, in the imbalanced domain, a dataset is defined as imbalanced only when the samples in different classes have a significant or even extreme gap in the number of samples [5]. In other words, one or more classes significantly outnumbers the other class(es) in an imbalanced dataset. The classes with more samples are called *majority* classes, while the underrepresented classes are called *minority* classes. Class-imbalance problems have caught growing attention from both academic and industrial fields. Many real-world classification tasks suffer from this problem, for instance, medical diagnosis, email filtering and fault detection. It is much more important to identify the minority samples correctly in these problems. The price of misclassifying the minority samples would be a massive loss of money in fault diagnosis, an unqualified product in anomaly detection and a person's life in medical diagnosis.

Many techniques have been developed to improve the minority class accuracy in class imbalance problems. In our experimental design (see Section 5), we consider the oversampling technique, which is a data-level approach to balance the data distribution. Oversampling balances the class distribution by replicating existing samples in the minority class or generating new artificial samples for the minority class. One of the most representative oversampling approaches is the Synthetic Minority Oversampling TEchnique (SMOTE). SMOTE works by creating artificial minority samples to produce balanced data. The artificial samples are generated based on the randomly chosen minority samples and their K-nearest neighbors. A new synthetic sample $\boldsymbol{x_s}$ can be generated according to the following equation [8, 10]:

$$\boldsymbol{x_s} = \boldsymbol{x_i} + \delta \cdot (\hat{\boldsymbol{x}}_{\boldsymbol{i}} - \boldsymbol{x_i}); \tag{1}$$

where $\boldsymbol{x_i}$ is the minority sample to oversample, $\hat{\boldsymbol{x}}_{\boldsymbol{i}}$ is a randomly selected neighbor from its K-nearest minority class neighbors and δ is a random number, where $\delta \in [0, 1]$, as described in [3].

3 Types of Samples in the Imbalanced Learning Domain

In this section, we first introduce the existing rule for identifying types of samples in binary scenarios (Section 3.1). Then, we show the drawbacks when extending this idea from binary to multi-class scenarios (Section 3.2), which motivates our own research presented in Section 4.

3.1 Types of Samples in Binary Scenarios

Napierala and Stefanowski [16] proposed to distinguish four types of samples composing minority class distribution: *safe, borderline, rare* samples and *outliers*. The identification of the four types is done by analysing the local characteristics of the samples. Considering that many applications involve both nominal and continuous attributes, the HVDM metric (Heterogenous Value Difference Metric) [24] is applied to calculate the distance between different examples. Given the number of neighbors k, the rule to identify the four types is given in Table 1, where $R_{\frac{min}{all}}$ is the ratio of the number of its minority class neighbors to the total number of neighbors [9, 16].

Table 1. Identification rule to assign types for minority samples.

Type	Rule	Rule ($k = 5$)
Safe	$\frac{k+1}{2k} < R_{\frac{min}{all}} \leqslant 1$	$\frac{3}{5} < R_{\frac{min}{all}} \leqslant 1$
Borderline	$\frac{k-1}{2k} \leqslant R_{\frac{min}{all}} \leqslant \frac{k+1}{2k}$	$\frac{2}{5} \leqslant R_{\frac{min}{all}} \leqslant \frac{3}{5}$
Rare	$0 < R_{\frac{min}{all}} < \frac{k-1}{2k}$	$0 < R_{\frac{min}{all}} < \frac{2}{5}$
Outlier	$R_{\frac{min}{all}} = 0$	$R_{\frac{min}{all}} = 0$

3.2 Problems When Extending to Multi-class Scenarios

Since the types of minority class examples were introduced in the binary imbalanced learning domain, various researchers confirmed the occurrence of the different types of samples in real-world data. They studied the influence different types of minority samples on binary imbalanced classification [5], and concluded that the *unsafe* samples are the actual source of difficulty when learning from imbalanced problems [23]. As the importance of learning different types of samples has received more and more attention, some studies extended this idea to multi-class imbalanced classification without changing the identification rule for the four types of samples [12, 20, 22]. However, the relationships among classes in multi-class imbalanced scenarios are more complicated than in binary scenarios, resulting in two main drawbacks if we follow the identification rule for binary scenarios.

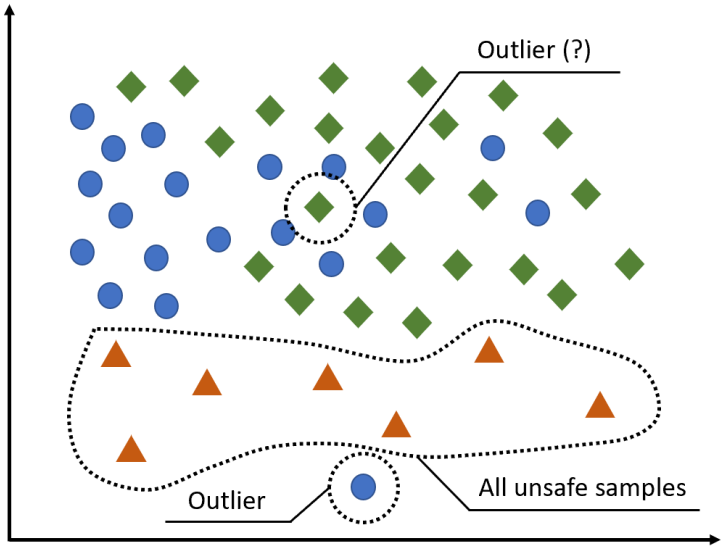

Fig. 1. An artificial 2-dimensional dataset showing the drawbacks when simply extending the identification rule in binary scenarios to multi-class scenarios. Suppose $k = 5$, then according to the identification rule in Table 1, the orange triangles (△) are all unsafe samples and the green diamond (⋄) marked with the dotted circle is an outlier.

- **A higher percentage of unsafe samples in minority classes.** In the identification rule in Table 1, the number of neighbors is set the same for all the classes when considering the neighborhood information. However, this setting neglects the fact that, in multi-class imbalanced classification, minority classes contain significantly fewer samples than in the majority classes. Hence, choosing the same k for all classes in multi-class scenarios will result in a higher percentage of unsafe samples (*borderline, rare, outliers*) in minority classes, see orange triangles (△) in Figure 1. The methods we propose to handle this problem are described will be shown later in Section 4.1.
- **False identification of outliers.** In the identification rule in Table 1, *outliers* refer to the isolated samples surrounded by different classes. For example, following this rule, the blue circle at the bottom (Figure 1) is classified as an outlier. However, the rule also distinguishes the green diamond (⋄) marked with the dotted circle (Figure 1) as an outlier. According to the geometric location of this sample, however, it is not an isolated sample far away from other samples in the same class. This indicates that the current rule leads to the false identification of some samples. In the case of multi-class problems, the relationships among classes are more complex, and our

proposed idea to reduce the probability of false identification is detailed in Section 4.2.

José et al. [20] analyzed the oversampling of different classes and types of samples with several benchmark multi-class imbalanced datasets. They calculate the percentage of each type of sample (safe/borderline/rare/outlier) using the identification rule for binary scenarios. Related information on selected datasets is given in Table 2. We can observe that if there is a significant gap between the number of minority and majority samples, over 60% of the minority class samples are considered outliers (see *C1* in *Balance* and *C1 & C2* in *Thyroid*). Hence, we confirm that the drawbacks above exist in multi-class benchmark datasets, and a new identification rule is required for distinguishing the types of samples in multi-class imbalanced scenarios.

Table 2. The number of samples of each class in the three selected datasets (detailed information on datasets shown in Table 6) and percentage of each type of sample (safe/borderline/rare/outlier) within the class (taken from José's work [20]). "Cj" indicates class j, percentages are rounded to integer values.

Dataset	C1	C2	C3
Balance	49 (0/0/4/**96**)	288 (74/26/0/0)	288 (73/27/0/0)
Thyroid	17 (0/12/6/**82**)	37 (0/11/24/**65**)	666 (97/3/0/0)
Wine	48 (98/2/0/0)	59 (100/0/0/0)	71 (85/14/1/0)

4 New Identification Rule for Multi-class Scenarios

In Section 3.2, we pointed out two main drawbacks when extending the identification rule from binary to multi-class scenarios. In this section, we propose a new identification rule for multi-class scenarios to overcome these drawbacks.

4.1 Adjusting k according to Imbalance Ratio

In the literature, the same k is used when assigning the types for samples in both majority classes and minority classes. However, considering the enormous gap between the sample size of minority and majority classes, choosing the same k will result in a higher percentage of unsafe samples in the minority class (stated in Section 3.2). Hence, to ensure a reasonable proportion of different types of samples in minority class(es), a smaller k should be used when analysing the

local characteristics of a minority sample. Here, we propose to adjust k to k_j according to the class distribution, as follows:

$$k_j = \left\lceil \sqrt{\frac{n_j}{N/C}} \times k \right\rceil, \tag{2}$$

where $j = 1, \ldots, C$ denotes the class index, n_j is the number of samples in class j, C is the number of classes and $N = \sum_{j=1}^{C} n_j$ is the total number of samples in the dataset. The results of adjusting k as shown in Table 3 indicate that Equation (2) meets our requirements for choosing a larger k for majority class(es) and a smaller k for minority class(es).

Table 3. The number of samples of each class in the three selected datasets and k_j for each class. k is preset to 5 and "Cj" indicates class j.

Dataset	C1	C2	C3
Balance	49	288	288
	$k_1 = 3$	$k_2 = 6$	$k_3 = 6$
Thyroid	17	37	666
	$k_1 = 2$	$k_2 = 2$	$k_3 = 9$
Wine	48	59	71
	$k_1 = 5$	$k_2 = 5$	$k_3 = 6$

4.2 Considering neighborhood Information of the neighbors

In Section 3.2, we illustrated that only considering neighbors of a sample is insufficient to identify the type because the neighborhood information might not adequately reflect the geometric location. Increasing k is a straightforward solution to expand neighborhood information. However, this will also decrease the number of safe samples for both minority and majority samples. For example, taking an extreme case, if k is large enough, all samples will be unsafe. Hence, we propose to consider neighborhood information of the neighbors additionally, i.e. we also find the k nearest neighbors for the neighbors. In our proposed approach, the importance of neighborhood information usually is higher than of neighborhood information of the neighbors. A definition of "type score (TS)" of data sample x is given below,

$$\begin{aligned} \mathrm{TS}(x) = & \overbrace{\alpha(x) \cdot \frac{n_x}{k_j}}^{\text{neighborhood}} + \underbrace{(1-\alpha(x)) \cdot \frac{N_x}{(k_j)^2}}_{\text{neighborhood of the neighbors}} \\ \alpha(x) = & \begin{cases} 1 - \frac{1}{k_j} & \text{if } k_j > 1 \\ 0.8 & \text{if } k_j = 1 \end{cases} \end{aligned} \tag{3}$$

where x belongs to class j, k_j is the number of nearest neighbors for sample x (see Section 4.1), n_x is the number of neighbors which share the same label with sample x, N_x is the number of neighbors of x's neighbors which share the same label with sample x, $\alpha(x)$ is the weight for the neighborhood information of sample x. If $k_j = 1$, we set $\alpha(x) = 0.8$ (to avoid $\alpha(x) = 1 - \frac{1}{k_j} = 0$) to ensure the higher importance of neighborhood information. Note that when considering the neighborhood information of the neighbors, we also use k_j. The proposed identification rule to assign the four types of samples in multi-class scenarios is given in Table 4. Following the proposed identification rule, the percentage of each type of sample is recalculated and shown in Table 5. For datasets with a significant gap between minority and majority sample sizes (*Balance* and *Thyroid*), the percentage of *outlier* type decreases from over 60% to less than 30% (compare with Table 2).

Table 4. Identification rule to assign types for samples in multi-class scenarios. Note that the thresholds can be adjusted.

Type	Safe	Borderline	Rare	Outlier
Rule	TS>0.75	0.5<TS≤0.75	0.05<TS≤0.5	TS≤0.05

Table 5. The number of samples of each class in the three selected datasets and percentage of each type of sample (safe/borderline/rare/outlier) within the class. "Cj" indicates class j, percentages are rounded to integer values.

Dataset	C1	C2	C3
Balance	49 (0/0/78/**22**)	288 (70/24/6/0)	288 (70/23/7/0)
Thyroid	17 (6/24/47/**23**)	37 (8/13/49/**30**)	666 (99/1/0/0)
Wine	48 (98/2/0/0)	59 (100/0/0/0)	71 (76/13/8/3)

5 Experiments

In this section, we first present our experimental setup and results on benchmark datasets (Section 5.1). After that, experiments on surface inspection data from industry are introduced in Section 5.2.

5.1 Experiments on Benchmark Datasets

Datasets. The experiments in this paper are based on 6 selected benchmark multi-class imbalanced datasets from the KEEL repository [1]. The descriptions of the each dataset is summarized in Table 6.

Table 6. Information on the benchmark datasets. AT, CL and NS indicate the number of attributes, the number of classes and the number of samples respectively.

Dataset	AT	CL	NS (in each class)
Balance	4	3	625 (49 / 288 / 288)
Contraceptive	9	3	1473 (333 / 511 / 629)
Glass	9	6	214 (9 / 13 / 17 / 29 / 70 / 76)
Thyroid	21	3	720 (17 / 37 / 666)
Wine	13	3	178 (48 / 59 / 71)
Winequality-red	11	6	1599 (10 / 18 / 53 / 199 / 638 / 681)

Setup. In this paper, we (1) improve the rule for identifying the four types of samples for multi-class imbalanced problems and (2) investigate how oversampling for different types of sample combinations affects the classification performance. Our experimental setup is illustrated in Figure 2. We consider $\binom{4}{3} + \binom{4}{2} + \binom{4}{1} = 15$ (excluding *None*) combinations of the four types of samples and SMOTE [3] to oversample these combinations in our experiments. Three classifiers (C5.0, SVM and Nearest Neighbor) are used as classification algorithms, and 5-fold stratified cross-validation is used to preserve the original class distribution [21].

Performance metrics. There is no standard performance metric to measure classifier performance in the multi-class imbalanced learning domain. The aim of studying the imbalanced problem is to improve the classification accuracy on minority class(es) while not losing too much accuracy on majority class(es). In our experiments, we use *MinAcc*, the average accuracy on minority class(es), to measure the performance on minority class(es) and *MAUC* (Multi-class Area Under the Curve) to measure the classification performance on the whole dataset.

$$\begin{aligned} \text{MinAcc} &= \sum_{i \in \text{C}_{min}} \text{TPR}_i / n_{\min}, \\ \text{MAUC} &= \frac{2}{c \cdot (c-1)} \sum_{j<k} \frac{A_{jk} + A_{kj}}{2}, \end{aligned} \tag{4}$$

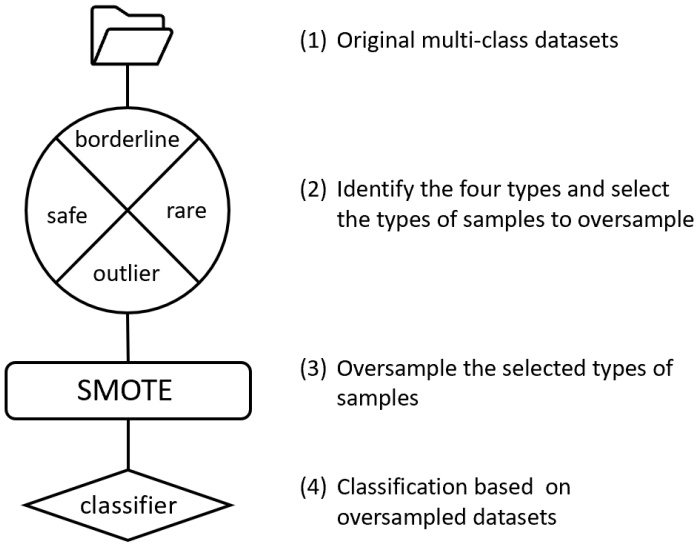

Fig. 2. Experimental setup to compare the effectiveness of the two different identification rules (inspired by [20]). The comparison is done via changing the identification rule in step (2).

C_{min} denotes the set of minority class indices, TPR_i is the true positive rate in class i, $n_{\min}$ denotes the number of minority classes, if there is more than one class being underrepresented in multi-class imbalanced classification, one should manually define the value of $n_{\min}$. A_{jk} indicates the probability that a sample randomly selected from class k has a lower probability for class j than randomly selected from class j, and A_{kj} is defined correspondingly. Detailed equation to compute A_{jk} can be found in [7].

Results and Discussions. Experimental results of C5.0 (average of 30 trials) on *Balance* and *Winequality-red* are given in Table 7 and Table 8. Note that there is one minority class in *Balance* and three minority classes in *Winequality-red*. Three main conclusions can be drawn from our experiments:

– Taking different types of sample combinations into account in the oversampling technique can significantly improve the classification performance on minority class(es). At the same time, improved or competitive classification performance on the whole dataset can also be achieved. Please refer to the bold numbers, the best performance in the 15 combinations, in Table 7 and Table 8. This improvement can be explained by the fact that, when considering different combinations, one or several types of samples will be discarded. This can be regarded as an informed undersampling to balance the class distribution.

- From the performance comparison between two identification rules ($R_{min/all}$ and TS), it can be concluded that our proposed identification rule provide significantly better performance on classifying minority class(es). Moreover, there are less "–" in the experiments using the proposed identification rule, where "–" means that there are not enough samples to execute the k-nearest-neighbor algorithm in the oversampling step. Both points confirm the appropriateness of and improvement provided by the proposed rule.
- Only informative experimental results are shown in the paper, and results on other datasets are omitted. The relationship between imbalance ratio and *MinAcc* is shown in Figure 3. The imbalance ratio (IR) for multi-class classification in this paper is defined as the average majority sample size to the average minority sample size. It is worth mentioning that if the imbalanced ratio is not significant (< 4), oversampling different combinations of types of samples will not bring a significant improvement on minority classification performance. However, no linear relationship between the imbalance ratio and *MinAcc* can be concluded (see linear regression equation and R^2 in Figure 3). This is because the improvement is not also determined by the imbalance ratio, but also depends on the separability of classes.

Table 7. Performance results of C5.0 on the dataset *Balance*.
"1 0 1 0" represents "safe(1) borderline(0) rare(1) outlier(0)", i.e. only safe and rare samples are oversampled. $R_{min/all}$ and TS indicate the different rules for identifying types of samples. "–" means that there are not enough samples to execute the k-nearest-neighbor algorithm in the oversampling step.

Combination	MinAcc		MAUC	
	$R_{min/all}$	TS	$R_{min/all}$	TS
1 1 1 1	0.0129	0.0129	0.7449	0.7449
1 1 1 0	–	0.1590	–	0.8179
1 1 0 1	**0.1546**	0.1374	0.8119	0.7712
1 0 1 1	0.1386	0.1600	**0.8138**	0.8216
0 1 1 1	0.0535	0.0676	0.7894	0.7934
1 1 0 0	0	0.0222	0.7534	0.7470
1 0 1 0	–	**0.1907**	–	**0.8219**
0 1 1 0	–	0.1301	–	0.8101
1 0 0 1	0.1151	0.1037	0.8092	0.7764
0 1 0 1	0.0474	0.0823	0.7825	0.7810
0 0 1 1	–	0	–	0.7348
1 0 0 0	0	0	0.7489	0.7537
0 0 1 0	–	0	–	0.7303
0 1 0 0	–	0	–	0.7481
0 0 0 1	–	–	–	–

Table 8. Performance results of C5.0 on the dataset *Winequality-red*. The huge difference in the corresponding positions of the two columns in *MinAcc* is caused by the significant difference between the four types of samples under the two identification rules, i.e., data distribution in different combinations varies a lot.

Combination	MinAcc		MAUC	
	$R_{min/all}$	TS	$R_{min/all}$	TS
1 1 1 1	0.0819	0.0819	**0.6751**	0.6751
1 1 1 0	–	0.0771	–	0.6581
1 1 0 1	0.0281	0.1219	0.6571	0.6637
1 0 1 1	0.0520	0.0588	0.6600	0.6627
0 1 1 1	0.0466	0.1170	0.6541	0.6534
1 1 0 0	–	–	–	–
1 0 1 0	–	0.0498	–	0.6576
0 1 1 0	–	0.0394	–	0.6548
1 0 0 1	**0.1305**	0.0444	0.6518	0.6584
0 1 0 1	0.0511	0.1140	0.6553	0.6601
0 0 1 1	0.0851	0.0680	0.6615	0.6637
1 0 0 0	–	0.0698	–	**0.6782**
0 0 1 0	–	0.0875	–	0.6616
0 1 0 0	–	–	–	–
0 0 0 1	0.0563	**0.1485**	0.6461	0.6453

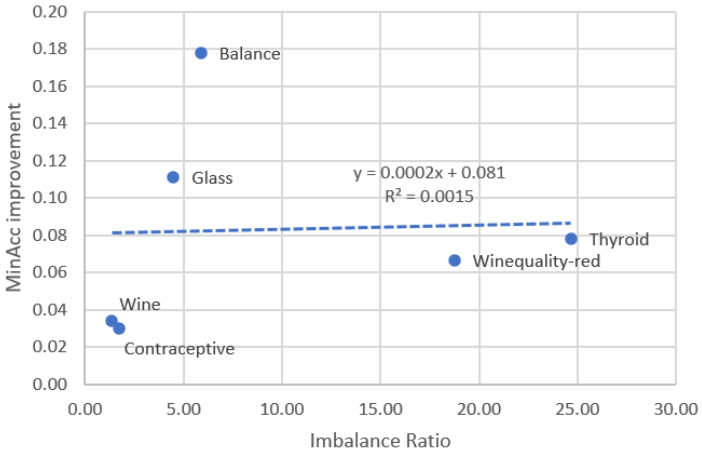

Fig. 3. Relationship between imbalance ratio and *MinAcc*. The imbalance ratio (IR) for multi-class classification in this paper is defined as the average majority sample size to the average minority sample size.

5.2 Experiments on Surface Inspection Dataset

The surface of a steel product is one of the major quality aspects. Therefore, surface anomalies should be avoided or at least known. A camera-based Surface Inspection Systems (SIS) is used in various process lines to identify those anomalies in the industry [18]. Grey value images taken from the surface by the SIS contains information on the anomalies. These images of various anomalies occurring in production are assessed and gathered in defined classes within a defect library. The defect library is used to train and test classifiers (classification algorithms), and these classifiers are finally used to identify the new surface anomalies from production. Thus, a stable, accurate and high classification performance is a must in the quality check procedure. However, the imbalance in the number of various defect types makes it challenging to obtain a stable and accurate classification performance.

Dataset and preprocessing. The surface inspection dataset used in this paper is taken from a defect library after a certain selection (for privacy reasons). The dataset initially contains 12496 samples along with 173 attributes. Samples with missing values were first removed, then we performed the feature correlation analysis and deleted highly-correlated attributes. The dataset after preprocessing contains 12456 samples with 62 attributes. The information on surface inspection data for experiments is given in Table 9.

Table 9. Information on the *surface inspection* dataset after preprocessing. NS and "class" indicate the number of samples and class label respectively.

class	NS	class	NS	class	NS	class	NS
25	2012	1	385	11	282	20	134
17	1666	10	382	19	255	23	121
24	1211	12	379	22	243	6	71
15	1205	16	357	9	215	4	39
18	937	7	354	21	201		
3	623	5	312	27	165	Total	
2	457	13	296	8	154	25	12456

Results and Discussions. Experimental results on the industrial surface inspection dataset are given in Table 10. This real-world dataset is a multi-class imbalanced dataset with an extreme imbalance ratio. Significant improvements on both minority and overall classification performance can be observed in Table 10. This is consistent with our conclusions from the experiments on benchmark datasets in Section 5.1. Furthermore, the best performances out of 15 combinations are contributed mainly by "no outliers (1 1 1 0)", which also shows that the outlier type has a significant influence on the classification performance in real-world imbalanced problems. In addition, the proposed identification rule (TS)

outperforms the other one on classifying minority samples. This confirms that the proposed rule can better recognise the outliers in this real-world problem.

Table 10. Performance results of C5.0 in *surface inspection* dataset."1 0 1 0" represents "safe(1) borderline(0) rare(1) outlier(0)", i.e. only safe and rare samples are oversampled. $R_{min/all}$ and TS indicate the different rules for identifying types of samples. "–" means that there are not enough samples to execute the k-nearest-neighbor algorithm in the oversampling step.

Combination	MinAcc		MAUC	
	$R_{min/all}$	TS	$R_{min/all}$	TS
1 1 1 1	0.5256	0.5256	0.8748	0.8748
1 1 1 0	**0.5361**	**0.5468**	0.8900	**0.8917**
1 1 0 1	0.4927	0.4780	**0.8924**	0.8881
1 0 1 1	0.5022	0.4994	0.8879	0.8880
0 1 1 1	0.5040	0.4923	0.8759	0.8746
1 1 0 0	–	–	–	–
1 0 1 0	–	0.5430	–	0.8914
0 1 1 0	0.5190	0.5301	0.8796	0.8794
1 0 0 1	0.4806	0.4754	0.8871	0.8857
0 1 0 1	0.4903	0.4671	0.8803	0.8758
0 0 1 1	0.4891	0.4944	0.8668	0.8679
1 0 0 0	–	–	–	–
0 0 1 0	–	–	–	–
0 1 0 0	–	–	–	–
0 0 0 1	–	–	–	–

6 Conclusions

This paper introduces the types of samples (safe, borderline, rare and outlier) in binary imbalanced literature and the drawbacks of extending this idea to multi-class imbalanced scenarios. We proposed a new identification rule to deal with these drawbacks and evaluated the effectiveness of this proposed rule on six benchmark datasets and a real-world application. According to our experimental results, the following conclusions can be derived:

- Oversampling different combinations of types of samples can provide better or competitive performance in classifying minority class(es) while not losing too much classification performance on majority class samples.
- The proposed identification rule for types of samples makes the percentage of each type of sample within the class more reasonable (avoiding all samples in the minority class considered as outliers).

– Our experimental results do not show significant improvement on datasets that are not highly imbalanced. Therefore, it is recommended to analyse the types of samples only when the dataset is highly imbalanced.
– The proposed identification rule can be applied to real-world multi-class imbalanced datasets and significantly improve the classification performance. When dealing with real-world problems, much attention should be paid to the sample type "outlier".

In future work, it is worth studying the relationship between imbalance ratio, separability of classes and performance improvement while analysing the four types of samples in the imbalanced learning domain. In addition, further study on applying the proposed identification rule to more real-world applications is encouraged. However, real-world data available in the machine learning community is rare due to confidentiality and the time-consuming generation. We also would like to explore how these four types of samples can be used for interacting with and benefiting from the feedback of human experts in real-world applications. One scenario is, for example, the rule identifies some outlier samples and plans to delete these samples in future analysis. Then, the human experts check whether these are real outliers and provide feedback to the algorithm training process.

References

1. 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)
2. Bermejo, P., Gámez, J.A., Puerta, J.M.: Improving the performance of naive bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets. Expert Systems with Applications **38**(3), 2072–2080 (2011)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
4. Dua, D., Graff, C.: UCI machine learning repository (2017), `http://archive.ics.uci.edu/ml`
5. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from imbalanced data sets, vol. 10. Springer (2018)
6. Ganganwar, V.: An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering **2**(4), 42–47 (2012)
7. Hand, D.J., Till, R.J.: A simple generalisation of the area under the roc curve for multiple class classification problems. Machine learning **45**(2), 171–186 (2001)
8. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on knowledge and data engineering **21**(9), 1263–1284 (2009)
9. Kong, J., Kowalczyk, W., Menzel, S., Bäck, T.: Improving imbalanced classification by anomaly detection. In: International Conference on Parallel Problem Solving from Nature. pp. 512–523. Springer (2020)

10. Kong, J., Kowalczyk, W., Nguyen, D.A., Menzel, S., Bäck, T.: Hyperparameter optimisation for improving classification under class imbalance. In: 2019 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE (2019)
11. Krawczyk, B., Galar, M., Jeleń, Ł., Herrera, F.: Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Applied Soft Computing **38**, 714–726 (2016)
12. Lango, M., Stefanowski, J.: Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data. Journal of Intelligent Information Systems **50**(1), 97–127 (2018)
13. Liu, B., Tsoumakas, G.: Synthetic oversampling of multi-label data based on local label distribution. arXiv preprint arXiv:1905.00609 (2019)
14. 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 (2013)
15. Mazurowski, M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A., Tourassi, G.D.: Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks **21**(2-3), 427–436 (2008)
16. Napierala, K., Stefanowski, J.: Types of minority class examples and their influence on learning classifiers from imbalanced data. Journal of Intelligent Information Systems **46**(3), 563–597 (2016)
17. Napierała, K., Stefanowski, J., Wilk, S.: Learning from imbalanced data in presence of noisy and borderline examples. In: International conference on rough sets and current trends in computing. pp. 158–167. Springer (2010)
18. Neogi, N., Mohanta, D.K., Dutta, P.K.: Review of vision-based steel surface inspection systems. EURASIP Journal on Image and Video Processing **2014**(1), 1–19 (2014)
19. Prati, R.C., Batista, G.E., Monard, M.C.: Class imbalances versus class overlapping: an analysis of a learning system behavior. In: Mexican international conference on artificial intelligence. pp. 312–321. Springer (2004)
20. Sáez, J.A., Krawczyk, B., Woźniak, M.: Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. Pattern Recognition **57**, 164–178 (2016)
21. Santos, M.S., Soares, J.P., Abreu, P.H., Araujo, H., Santos, J.: Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]. IEEE Computational Intelligence Magazine **13**(4), 59–76 (2018)
22. Sleeman IV, W.C., Krawczyk, B.: Multi-class imbalanced big data classification on spark. Knowledge-Based Systems **212**, 106598 (2021)
23. Wang, S., Minku, L.L., Yao, X.: A systematic study of online class imbalance learning with concept drift. IEEE transactions on neural networks and learning systems **29**(10), 4802–4821 (2018)
24. Wilson, D.R., Martinez, T.R.: Improved heterogeneous distance functions. Journal of artificial intelligence research **6**, 1–34 (1997)