Learning class-imbalanced problems from the perspective of data intrinsic characteristics

Kong, J.

CHAPTER 6

Improved Sample Type Identification for Multi-Class Imbalanced Classification

The idea of studying different types of samples 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. In this chapter, we introduce our proposed sample type identification for multi-class imbalanced classification. First, Section 6.1 shows the motivation and briefly introduces on our work. After that, in Section 6.2, the literature review and problems when extending to multi-class scenarios are presented. In Section 6.3, detailed information on the new identification rule is given. In Section 6.4, the information on the datasets, the experimental setup as well as the experimental results and discussion are introduced. In addition, a real-world application is described in Section 6.5. Section 6.6 concludes the chapter and outlines the further work.

6.1 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 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 (Fernández, García, Galar, Prati, Krawczyk, and Herrera, 2018; Ganganwar, 2012). However, several studies in the literature have pointed out that the data characteristics also play a crucial role in dealing with imbalanced problems (López, Fernández, García, Palade, and Herrera, 2013; Napierała, Stefanowski, and Wilk, 2010; Prati, Batista, and Monard, 2004). Here, Napierala and Stefanowski proposed to consider samples from minority class consisting of four types of samples: *safe, borderline, rare* samples and *outliers* (Napierala and Stefanowski, 2016). 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 (Kong, Kowalczyk, Menzel, and Bäck, 2020; Lango and Stefanowski, 2018; B. Liu and Tsoumakas, 2019). 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 chapter, we first recall the identification rule for the four types of samples as proposed in the literature (Napierala and Stefanowski, 2016). 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 approach 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 (Mazurowski, Habas, Zurada, Lo, Baker, and Tourassi, 2008), email filtering (Bermejo, Gámez, and Puerta, 2011), fault diagnosis (Krawczyk, Galar, Jeleń, and Herrera, 2016), 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 (Alcalá-Fdez, Fernández, Luengo, Derrac, García, Sánchez, and Herrera, 2011; D. Dua and Graff, 2017). Therefore, our work makes an additional contribution by introducing a challenging industrial surface defects 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.

6.2 Related Works

In this section, we first introduce the existing rule for identifying types of samples in binary scenarios from the related literature (Section 6.2.1). Then, we show the drawbacks when extending this idea from binary to multi-class scenarios (Section 6.2.2), which motivates our own research presented in Section 6.3.

6.2.1 Studies on Types of Samples in Binary Scenarios

It is essential to recall the identification of types of samples in binary scenarios. Napierala and Stefanowski first proposed the idea of identifying minority class samples in four categories: *safe, borderline, rare* samples and *outliers* (Napierala and Stefanowski, 2016), the latter three are called *unsafe* samples. The majority class samples can also be categorized into these four types. The general rule to identify the four types is as follows.

- a sample is considered to be **safe** if the majority of the neighbours belongs to the same class;
- a sample is considered to be **borderline** if the proportion of the neighbours in both classes is approximately the same;
- a sample is considered to be **rare** if the majority of the neighbours belongs to a different class;
- a sample is considered to be an **outlier** if all the neighbours belongs to a different class.

Since the idea was proposed, it has attracted widespread attention in the field of imbalanced learning, and more than 200 papers have cited the original paper so far. It appears in the citations of review papers as an important development in the imbalanced learning domain, and also in the citations of papers proposing new approaches as a source of inspiration. Various researchers confirmed the occurrence of the different types of samples in real-world data. They studied the influence of different types of minority class samples on binary imbalanced classification (Fernández, García, Galar, Prati, Krawczyk, and Herrera, 2018), and concluded that the *unsafe* samples are the actual source of difficulty when learning from imbalanced problems (S. Wang, Minku, and Yao, 2018). Studies also focus on investigating the influence of minority class samples on the performance of SMOTE (Skryjomski and Krawczyk, 2017). This idea is also evaluated in real-world applications. For example, authors in (García, Marqués, and Sánchez, 2019) explored the effects of sample types on credit risk and corporate bankruptcy prediction.

6.2.2 Problems When Extending to Multi-class Scenarios

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 (Lango and Stefanowski, 2018; Sáez, Krawczyk, and Woźniak, 2016; Sleeman IV and Krawczyk, 2021). 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.

- **A higher percentage of unsafe samples in minority classes.** In the identification rule in Table 2.3, the number of neighbours is set the same for all the classes when considering the neighbourhood 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 ($\triangle$) in Figure 6.1. The methods we propose to handle this problem are described later in Section 6.3.1.

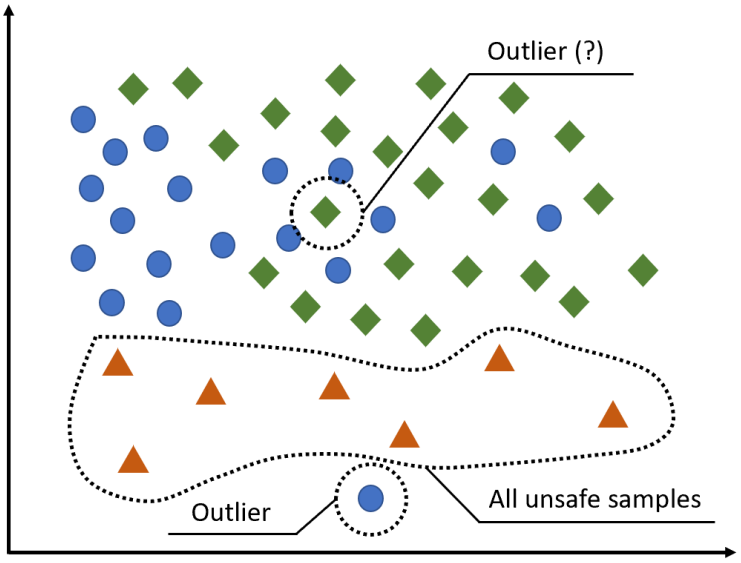

Figure 6.1: An artificial 2-dimensional dataset showing the drawbacks when simply extending the identification rule in binary scenarios to multi-class scenarios. Suppose $k = 5$, then according to the identification rule in Table 2.3, the orange triangles ($\triangle$) are all unsafe samples and the green diamond ($\diamond$) marked with the dotted circle is an outlier.

- **False identification of outliers.** In the identification rule in Table 2.3, *outliers* refer to the isolated samples surrounded by different classes. For example, following this rule, the blue circle at the bottom (Figure 6.1) is classified as an outlier. However, the rule also distinguishes the green diamond ($\diamond$) marked with the dotted circle (Figure 6.1) 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 6.3.2.

José et al. (Sáez, Krawczyk, and Woźniak, 2016) 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 6.1. We can observe that if there is a significant gap between the number of minority and majority class 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 6.1: The number of samples of each class in the three selected datasets (detailed information on datasets shown in Table 6.5) and percentage of each type of sample (safe/borderline/rare/outlier) within the class (taken from José's work (Sáez, Krawczyk, and Woźniak, 2016)). "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)

6.3 New Identification Rule for Multi-class Scenarios

In Section 6.2.2, 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.

6.3.1 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, where k is the k in k-NN within the sampling methods. 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 6.2.2). 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

class sample. Here, we propose to adjust k to k_j according to the class distribution, as follows:

$$k_j = \left\lceil \sqrt{\frac{n_j}{N/C}} \times k \right\rceil, \tag{6.1}$$

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 6.2 indicate that Equation (6.1) meets our requirements for choosing a larger k for majority class(es) and a smaller k for minority class(es).

Table 6.2: 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
	$k_1 = 5$	$k_2 = 5$	$k_3 = 6$

6.3.2 Considering neighbourhood Information of the neighbours

In Section 6.2.2, we illustrated that only considering neighbours of a sample is insufficient to identify the type because the neighbourhood information might not adequately reflect the geometric location. Increasing k is a straightforward solution to expand neighbourhood information. However, this will also decrease the number of safe samples for both minority and majority class samples. For example, taking an extreme case, if k is large enough, all samples will be unsafe. Hence, we propose to consider neighbourhood information of the neighbours additionally, i.e. we also find the k nearest neighbours for the neighbours. In our proposed approach, the importance of neighbourhood information usually is higher than of neighbourhood information of the neighbours. A definition of "type score (TS)" of data sample x is

given below,

$$\mathrm{TS}(x) = \overbrace{\alpha(x) \cdot \frac{n_x}{k_j}}^{\text{neighbourhood}} + \underbrace{(1-\alpha(x)) \cdot \frac{N_x}{(k_j)^2}}_{\text{neighbourhood of the neighbours}} \tag{6.2}$$

$$\alpha(x) = \begin{cases} 1 - \frac{1}{k_j} & \text{if } k_j > 1 \\ 0.8 & \text{if } k_j = 1 \end{cases}$$

where x belongs to class j, k_j is the number of nearest neighbours for sample x (see Section 6.3.1), n_x is the number of neighbours which share the same label with sample x, N_x is the number of neighbours of x's neighbours which share the same label with sample x, $\alpha(x)$ is the weight for the neighbourhood 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 neighbourhood information. Note that when considering the neighbourhood information of the neighbours, we also use k_j. The proposed identification rule to assign the four types of samples in multi-class scenarios is given in Table 6.3. Following the proposed identification rule, the percentage of each type of sample is recalculated and shown in Table 6.4. For datasets with a significant gap between minority and majority class sample sizes (*Balance* and *Thyroid*), the percentage of *outlier* type decreases from over 60% to less than 30% (compare with Table 6.1).

Table 6.3: Identification rule to assign types for samples in multi-class scenarios. Note that the thresholds can be adjusted (hand-tuned) depending on the given datasets.

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

6.4 Experiments

In this section, we introduce the information on the datasets used in our experiments. Then, the experimental setup is described. After that, the experimental results and discussions are given.

Table 6.4: 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)

6.4.1 Information on the Datasets

The experiments in this chapter are based on 6 selected benchmark multi-class imbalanced datasets from the KEEL repository (Alcalá-Fdez, Fernández, Luengo, Derrac, García, Sánchez, and Herrera, 2011). The descriptions of the datasets are summarized in Table 6.5.

Table 6.5: 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)

6.4.2 Experimental Setup

In this chapter, 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 6.2. We consider $\binom{4}{4}+\binom{4}{3}+\binom{4}{2}+\binom{4}{1}=15$ (excluding *None*) combinations of the four types of samples and SMOTE (Chawla, Bowyer, Hall, and Kegelmeyer, 2002) to oversample these combinations in our experiments. To be specific, $\binom{4}{4}$ means we choose all four types of samples to be oversampled, $\binom{4}{3}$ means we choose three out of four types of samples to be oversampled, $\binom{4}{2}$ means we choose two out of four types of samples to be oversampled and $\binom{4}{1}$ means we choose only one type of samples to be oversampled. Three classifiers (C5.0, SVM and Nearest Neighbour) are used as classification algorithms, and 5-fold stratified cross-validation is used to preserve the original class distribution (M. S. Santos, Soares, Abreu, Araujo, and J. Santos, 2018).

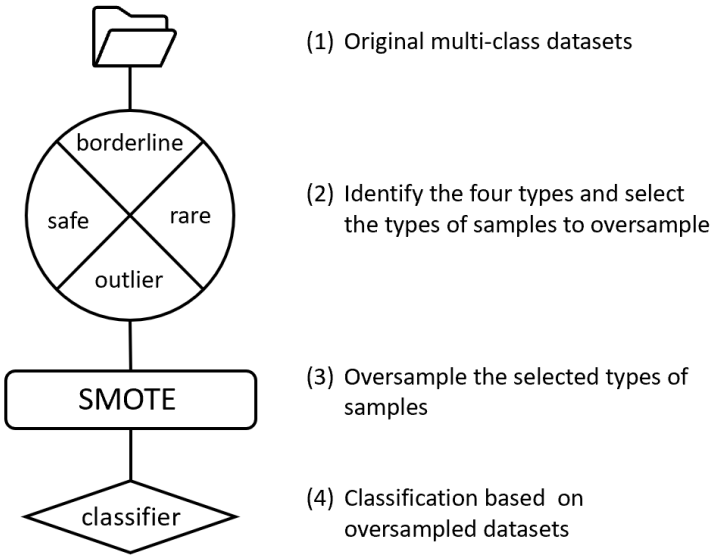

Figure 6.2: Experimental setup to compare the effectiveness of the two different identification rules (inspired by (Sáez, Krawczyk, and Woźniak, 2016)). The comparison is done via changing the identification rule in step (2).

6.4.3 Experimental Results and Discussion

Experimental results of the decision tree C5.0 (average of 30 trials) on *Balance* and *Winequality-red* are given in Table 6.6 and Table 6.7. 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:

Table 6.6: Performance results of decision tree (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-neighbour 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	–	–	–	–

- 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 6.6 and Table 6.7. 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-neighbour algorithm in the oversampling step. Both points confirm the appropriateness of and improvement provided by the proposed rule.

- Only experimental results on the dataset *Winequality* are shown in this chapter. Experimental results on other datasets can be found in Appendix A. The relationship between imbalance ratio and *MinAcc* is shown in Figure 6.3. The imbalance ratio (IR) for multi-class classification in this chapter is defined as the average majority sample size to the average minority class 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 6.3). This is because the improvement is not only determined by the imbalance ratio, but also depends on the separability of classes.

Table 6.7: 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

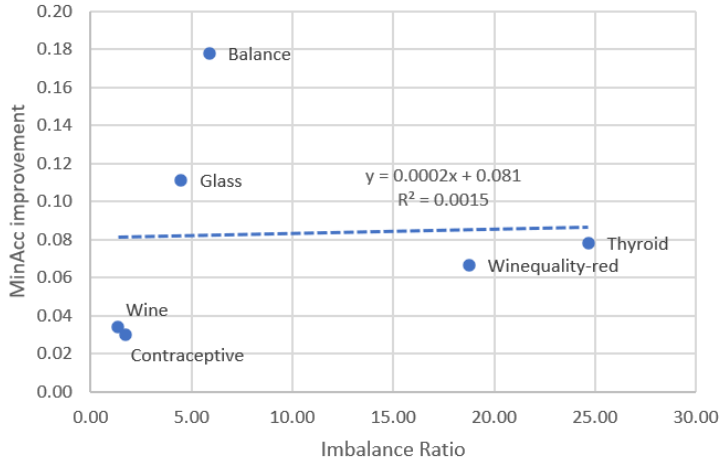

Figure 6.3: Relationship between imbalance ratio and *MinAcc*. The imbalance ratio (IR) for multi-class classification in this chapter is defined as the average majority sample size to the average minority class sample size.

6.5 Applications on the Detection of Surface Defects

In this section, we report our study on an imbalanced application for detecting surface defects. We first introduce the industrial problem. Then, the information on the surface defects dataset is given in Section 6.5.1. After that, the visualisation and preprocessing step on this high-dimensional dataset is described in Section 6.5.2. In Section 6.5.3, we evaluate our proposed sample identification rule on the surface defects 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 (Neogi, Mohanta, and Dutta, 2014). 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. Figure 6.4 shows a diagram of how to capture the defects images. 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.

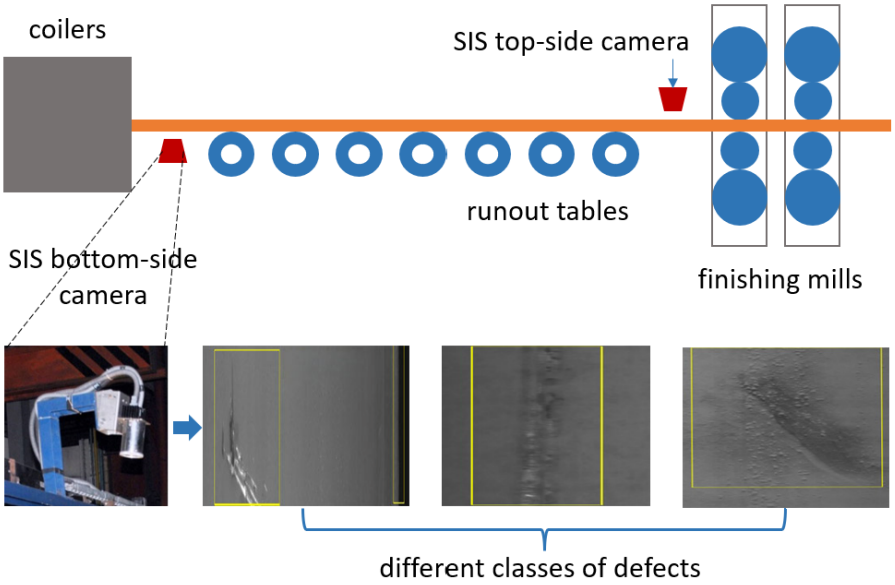

Figure 6.4: A diagram of how to capture the defects images. Defects images are from TATA steel official website[1], for example purpose.

6.5.1 Information on Surface Defects Dataset

The images captured by the SIS cameras will be processed in the feature extraction module. Relevant defect features, e.g. geometrical, textural and moment features, are extracted for the purpose of classification. Both the images and information after extraction will be stored in the defect library. The surface defects dataset used in this chapter is taken from a defect library after a certain selection (for privacy reasons). The dataset is after extraction and contains 12496 samples along with 173 attributes. After removing samples with missing values, there are 12456

[1] https://automation.tatasteel.com/products/rolling-mills/squins-surface-quality-inspection-system/

samples in total. The information on surface defects data for experiments is given in Table 6.8.

Table 6.8: Dimension of each record in the *surface defects* dataset after preprocessing. NS and “class" indicate the number of samples and class label respectively. There are 25 classes and 12456 samples in total

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

6.5.2 Visualisation and Preprocessing

Visualisation is an important step when dealing with real-world applications. It can provide some general information on the datasets, e.g. clusters. In the data-preprocessing step, missing values and redundant attributes are usually removed to provide high-quality data for future experiments.

Visualisation with t-SNE

Before experimenting with this real-world application dataset, we visualise the data to get some general information on the data. *T-distributed Stochastic neighbourhood Embedding* (t-SNE) (Van der Maaten and G. Hinton, 2008), a variation of *Stochastic neighbourhood Embedding* (SNE) (G. E. Hinton and Roweis, 2002), is a statistical technique for visualising high-dimensional data. It first converts high-dimensional Euclidean distance into conditional probability to characterise similarity among data points. Then, t-SNE models the similarity distribution among data points in the low-dimensional map. After that, it minimises the Kullback-Leibler divergence (KL divergence) between the joint distributions in high-dimensional and low-dimensional space.

t-SNE has been used for visualisation in various applications, consisting of medical research (Esteva, Kuprel, Novoa, Ko, Swetter, Blau, and Thrun, 2017), music analysis (Van den Oord, Dieleman, and Schrauwen, 2013), bioinformatics

(Baxevanis, Bader, and Wishart, 2020), etc. In this chapter, we use t-SNE to visualise the surface defects data from industry. As we discussed in Section 6.2.2, the relationships among classes in multi-class scenarios are more complicated than in binary scenarios. It is very intuitive from Figure 6.5 that as the number of classes increases, it gets more and more difficult to visualise the boundaries of different classes.

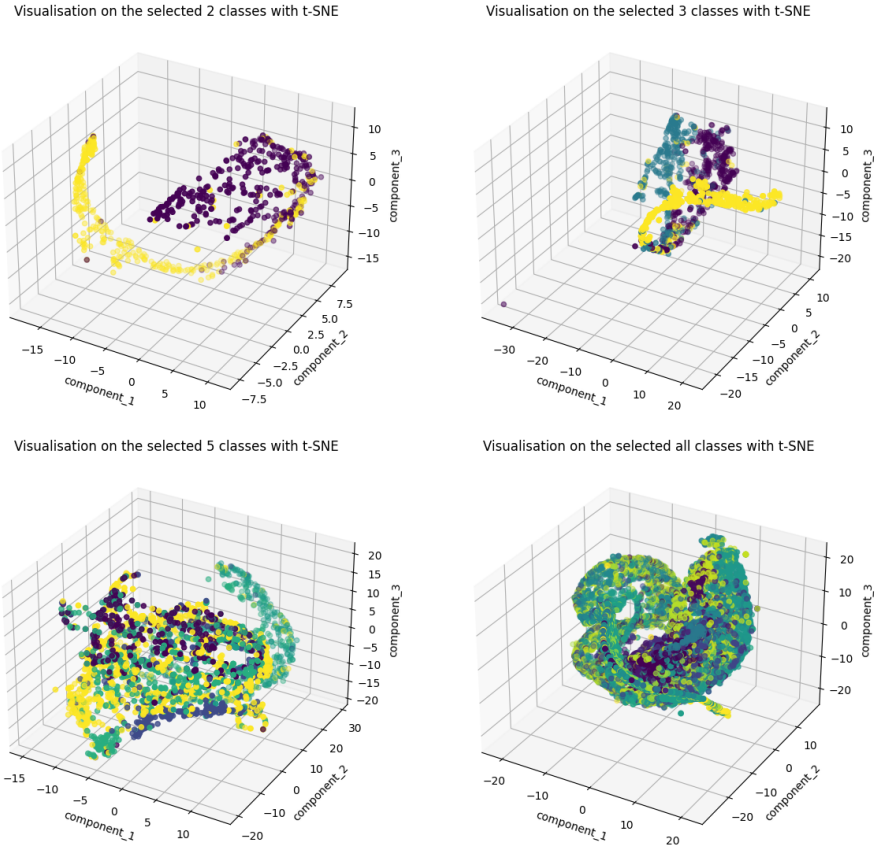

Figure 6.5: Visualisation on *surface defects* dataset with 2/3/5/all classes (top-left/top-right/bottom-left/bottom-right).

Data Preprocessing

As we mentioned in Section 6.5.1, we have already deleted the missing values. Therefore, in this chapter, we focus on reducing dimensionality via feature

correlation analysis, i.e. feature selection. *Correlation* is a statistical term to describe the linear relationship between two or more variables. When correlation happens in features (attributes), we call this *feature correlation*. In other words, if two features have a high correlation, we can predict one from the other. When training a predictive model based on a certain dataset, correlated features are considered redundant and we can delete one of them for simplification. As per the *Occam's razor*, "entities should not be multiplied beyond necessity" (Schaffer, 2015). (In Latin, *Entia non sunt multiplicanda praeter necessitatem* (Bauer, 2007).)

According to the information from the industry (which provides the surface defects data), the first 20 attributes in the surface defects dataset are only for internal recording, such as image number, date, top camera or bottom camera, etc. These features provide no information on the defects and can be directly deleted. After that, we calculate the feature correlation through *Pearson* correlation. From Figure 6.6, we can observe that many features are highly correlated. For our surface defects dataset, if the correlation between two features is higher than 0.7 (this number is suggested by the industrial expert in TATA company), one of them will be deleted. After removing the redundant features, there are 62 features left for future experiments.

6.5.3 Experiments on Surface Defects Dataset

Experimental results on the industrial surface defects dataset are given in Table 6.9. 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 6.9. This is consistent with our conclusions from the experiments on benchmark datasets in Section 6.4.3. 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 class samples. This confirms that the proposed rule can better recognise the outliers in this real-world problem.

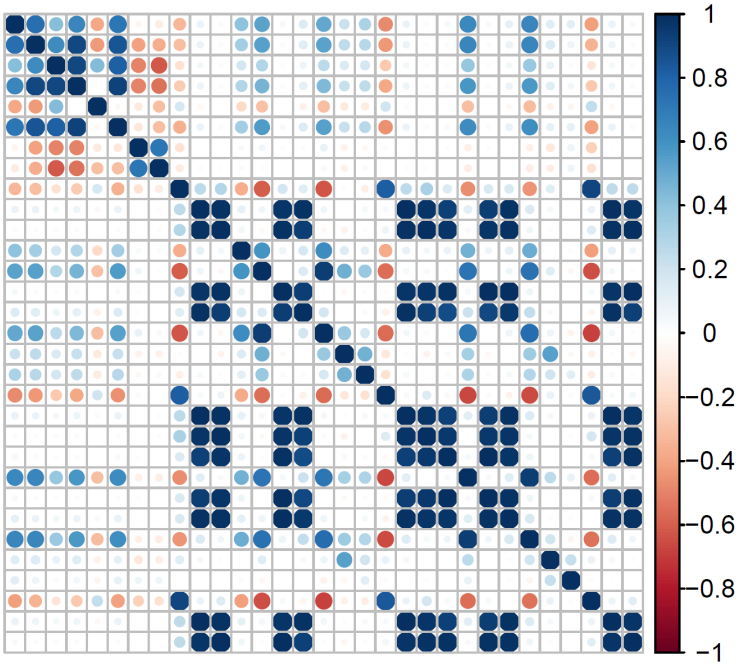

Figure 6.6: Visualisation of correlation matrix on 31 selected features. Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients.

6.6 Conclusions and Future Work

The idea of introducing four types of samples (safe, borderline, rare and outlier) in binary imbalanced literature has been done already. This chapter introduces 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

Table 6.9: Performance results of C5.0 in *surface defects* 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-neighbour 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	–	–	–	–

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.