Identification of Different Types of Minority Class Examples in Imbalanced Data

Krystyna Napierala and Jerzy Stefanowski

Institute of Computing Sciences, Poznań University of Technology,
60–965 Poznań, Poland
{krystyna.napierala,jerzy.stefanowski}@cs.put.poznan.pl

Abstract. The characteristics of the minority class distribution in imbalanced data is studied. Four types of minority examples – safe, borderline, rare and outlier – are distinguished and analysed. We propose a new method for identification of these examples in the data, based on analysing the local neighbourhoods of examples. Its application to UCI imbalanced datasets shows that the minority class is often scattered without too many safe examples. This characteristics of data distributions is also confirmed by another analysis with Multidimensional Scaling visualization. We examine the influence of these types of examples on 6 different classifiers learned over various real-world datasets. Results of experiments show that the particular classifiers reveal different sensitivity to the type of examples.

Keywords: Imbalanced data, Classifiers, MDS Visualisation.

1 Introduction

Learning classifiers from imbalanced data has been receiving a growing research interest. Although several methods have already been introduced (see, e.g., a review in [2,4]), it is still worth asking a question about the nature of the class imbalance problem and about the properties of data distribution which make it so difficult. Some earlier studies, mainly based on experiments with artificial data, showed that simple class imbalance ratio was not the main difficulty. The degradation of classification performance is also related to other factors, e.g. to *decomposition* of the minority class into many sub-concepts with very few examples, which correspond to the *small disjuncts* [5]. Moreover, *overlapping* between classes strongly deteriorates the recognition of the minority class [3,9].

Following these related studies one could still look for other factors characterizing the data distribution. In our earlier papers [8] we hypothesized that some minority class examples could be located deeper inside the majority class. They could be treated as *outliers* or *rare cases* (if they are not single ones). We think that they should not be considered as a noise, as they are too rare and too precious for the minority class.

The role of the above mentioned data factors has been preliminary studied by us in the experiments with special artificial datasets [8]. Related research

E. Corchado et al. (Eds.): HAIS 2012, Part II, LNCS 7209, pp. 139–150, 2012.

was also mainly focused on experimenting with artificial datasets [5,3,9]. By introducing a certain type of disturbance (e.g. overlapping or small disjuncts) and manipulating with its degree, the influence on the recognition of minority classes and on the abilities of particular classifiers were analysed.

In this study we direct our interest to the real-world imbalanced datasets. We would like to verify how often these factors actually occur in the data and to study their impact on the performance of different popular classifiers.

Our first aim is to analyse the distribution of examples in 19 real imbalanced datasets, mainly coming from the UCI repository[1] and often used in various experimental studies. We will show by analysing a 2D visualisation of a dataset obtained by *Multidimensional Scaling* (MDS) that in practice most of the datasets are seriously disturbed and that examples from the minority class can be of different nature. In our opinion, one can distinguish the following *types* of these examples: *safe*, *borderline*, *outliers* and *rare examples*.

The other aim of our study is to introduce a new method for identification of these types of examples in the data which is based on analysing a local neighbourhood of learning examples. In the experiments carried out with the same 19 datasets we plan to evaluate the amount of each type of examples. Depending on the main type of identified examples, we will categorize the datasets representing different characteristics of the minority class.

Finally, within each category of datasets, we compare the classification abilities of the classifiers – J48, PART, JRip, kNN, RBF and SVM. We want to verify whether they reveal different behaviour in face of different data types and how much they are sensitive to them.

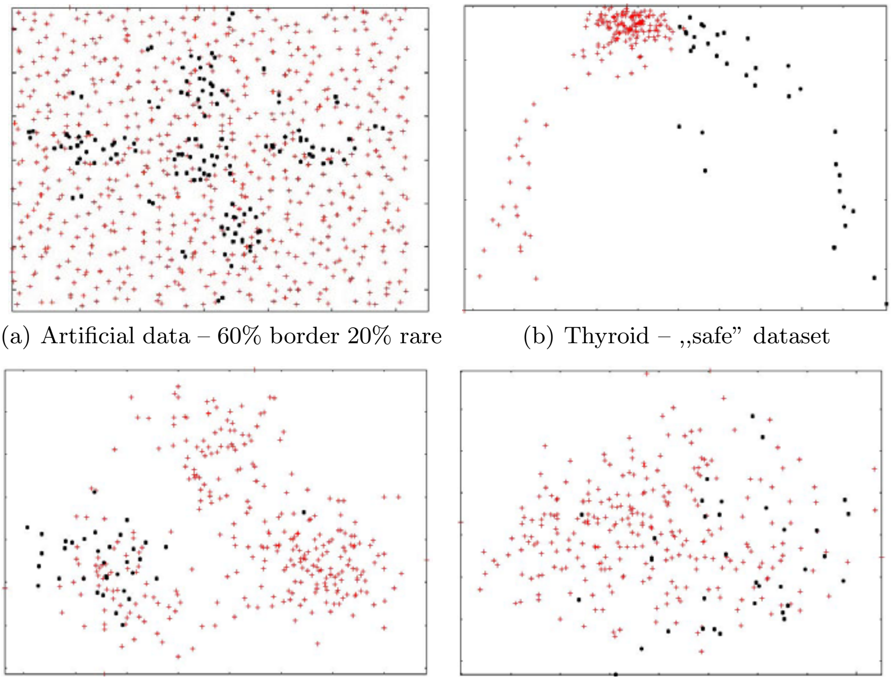
2 Distribution of Examples in the Minority Class

It is often claimed that learning from data with clearly separated classes is not difficult for most classifiers. It also concerns imbalanced data, as showed in the experimental studies, e.g. in [9]. Recognizing the minority class becomes more difficult when the distribution of examples from different classes is heavily mixed. Some researchers have also claimed that mutual position of examples has a crucial impact on learning from imbalanced data [3,6].

Several types of examples can be distinguished. The most common is the distinction between safe and unsafe examples [6]. *Safe* examples are located in the homogenous regions populated by the examples from one class only, otherwise they are treated as *unsafe* ones. Unsafe examples are often further discriminated between *borderline* and *noisy* examples as e.g. in [6]. *Borderline* examples are placed in the boundary regions between classes, where the examples from both classes overlap. Singular examples located deeper in the regions where the opposite class prevails, are usually treated as noisy examples. However, we share a different point of view. We claim that the minority class is often underrepresented in the dataset, so even the singular observations may represent a meaningful concept. What is more, as we will show in our experimental study, such

[1] `http://www.ics.uci.edu/~mlearn/MLRepository.html`

examples often represent a considerable number of minority examples (even as much as half of the class). Therefore, we would like to pay special attention to these examples. If they are single examples surrounded by many examples from majority classes, we treat them as *outliers*. Although some of them may indeed be noisy observations, in general they are too precious to be automatically discarded. The observations distant from the core of the minority class may also form small groups of two-three examples. In such a situation they are even less likely to be noisy. We call them *rare examples*.


(a) Artificial data – 60% border 20% rare

(b) Thyroid – ,,safe" dataset

(c) Ecoli – ,,borderline" dataset

(d) Cleveland – ,,rare" / ,,outlier" dataset

Fig. 1. MDS visualisation of selected imbalanced datasets

To illustrate this categorization, in Fig. 1(a) we present an artificial dataset (coming from [8]) described with two numerical attributes, where the minority class (black circles) consists of different types of examples. It is divided into five sub-concepts (clusters). In each of these concepts, only the examples lying near the center of the cluster can be considered as *safe*. Many more examples belong to the border between the classes, in which the majority examples overlap with the minority ones. Finally, there are some examples more distant from the clusters, which could represent *outliers* or *rare examples*.

To confirm these observations in the real-world datasets, we visualise three commonly used imbalanced datasets from the UCI repository: thyroid, ecoli and cleveland (Fig 1(b)-1(d)). As these datasets are described with more than two attributes, we use a *Multidimensional Scaling* technique (MDS) to reduce the dimensionality of the datasets. MDS performs a nonlinear mapping of dimensions with the aim of preserving the pairwise distances between data points in the original high dimensional data space into the projected low dimensional subspace [1]. As the datasets have both numeric and nominal attributes, we calculate the distances between the points using the HVDM metric [10].

Let us remark that using 2 dimensions in MDS requires keeping the data variance at a sufficient level. For instance, we could not use this technique to visualize the hepatitis dataset, as MDS with two dimensions preserved only 25% of variance in the dataset. For the other datasets, including three datasets visualised in Fig 1(b)-1(d), the percentage of preserved variance was higher than 60%, which in our opinion is enough to analyse the data.

Looking at (Fig 1(b)-1(d)) one can notice that the three data sets are of different nature. In thyroid dataset (Fig. 1(b)), the classes are clearly separated (even linearly), so most of the minority examples represent safe examples. In ecoli dataset (Fig. 1(c)) however, the classes seriously overlap. The consistent region belonging solely to the minority class (on the very left) is rather small – most examples lie in a mixed region between the classes. Finally, the cleveland dataset (Fig. 1(d)) is even more difficult to learn, as the minority class is very scattered – most examples form very small groups of few examples and some of the other are singular observations, surrounded by the opposite class. This dataset consists mostly of *rare examples* and *outliers.*

3 Assessing Types of Examples

Following the hypothesis about different types of examples in the minority class, we need an automatic procedure for their identification. We propose to assess the type of an example by analysing its local neighbourhood in the original attribute space[2]. For each minority example, we analyse the class assignment of its k-nearest neighbours. We use $k = 5$, because $k = 3$ may poorly distinguish the nature of examples, and 5 is often used in the preprocessing methods for class imbalance. With such k, the proportion of neighbours from the same class against neighbours from the opposite class can range from 5:0 (all neighbours are from the same class as the analysed example) to 0:5 (all neighbours belong to the opposite class). Depending on this proportion, we propose to assign the labels to the examples in the following way:

- 5:0 or 4:1 – an example is labelled as a safe example (further denoted as S).
- 3:2 or 2:3 – a borderline example (denoted as B). The examples with the proportion 3:2 are correctly classified by its neighbours, so they might still be safe. However, we prefer to be more pessimistic, and assume that they could be located too close to the decision boundary between the classes.

[2] The MDS projection to the reduced attribute space is applied for visualization aims only.

- 1:4 – labelled as a rare example (denoted as R), only if its neighbour from the same class has the proportion of neighbours either 0:5 or 1:4, but pointing to the analysed example. Otherwise there are some other examples from the same class in the proximity (although not in the immediate surrounding of $k = 5$), which suggests that it is rather a borderline example B.
- 0:5 – an example is labelled as an outlier and denoted as O.

To calculate the distance between examples we use the HVDM distance metric. It aggregates normalized Euclidean distances for numeric attributes with Stanfil and Valtz value difference metric for nominal attributes [10].

As our method is based on a simple analysis of a fixed number of neighbours, we want to check whether the assigned labels can precisely reflect the known distribution of examples. Inspired by good experience with artificial data in [8], we generated a number of such datasets (with 800 examples described by 2 numerical attributes) with varying imbalance ratios and number of the minority class sub-concepts, in which we changed the percentage of safe, borderline, rare and outlying examples. Table 1 presents the description of several analysed datasets and the labelling results.

Table 1. Labelling of artificial datasets

Dataset Description					Identified Labels			
Imbalance Ratio	Sub-concepts	Border [%]	Rare [%]	Outlier [%]	Safe [%]	Border [%]	Rare [%]	Outlier [%]
1:5	1	60	20	0	17.04	60.74	21.48	0.74
1:5	3	60	20	0	18.52	57.78	23.70	0.00
1:5	5	60	20	0	17.78	64.44	17.78	0.00
1:5	5	0	0	10	64.44	25.93	0.00	9.63
1:7	5	0	0	10	54.00	36.00	0.00	10.00
1:9	5	0	0	10	52.00	36.00	2.00	10.00

The first three datasets are disturbed in the same way (60% of borderline examples and 20% of rare examples), but differ in the number of sub-concepts. One of them (with 5 sub-concepts) is plotted in Fig. 1(a). Proportions of the identified labels show that our labelling method can correctly reconstruct the percentage of safe, borderline and rare examples, regardless of the number of sub-concepts. The other three datasets contain 10% of outliers and differ according to the imbalance ratio. Here, the labels also correctly reflect the percentage of outliers. However, although the classes in these datasets are not overlapped, a considerable number of examples is labelled as borderline. This is to some extent understandable, as the examples close to the border between the classes can contain in their neighbourhood some examples from the opposite class. Moreover, while labelling examples as borderline, we pessimistically assume that safe examples (3:2) also belong to this category.

4 Analysing Real-World Datasets

4.1 Datasets

We will conduct our analysis on 19 real-world datasets representing different domains, sizes and imbalance ratios. Their main characteristics are presented in the left-hand part of Table 2. 15 datasets come from the UCI repository and are often used in other works on class imbalance. Four datasets are retrospective medical datasets, which we used in our earlier works concerning imbalanced data[3]. If some datasets contain more than one majority class, we aggregate them into one class. The data are not modified, e.g. missing attribute values are handled directly by our methods and classifiers.

4.2 Labelling Results and Categorization of Datasets

The results of labelling the minority class examples in all the datasets are presented in the right-hand part of Table 2. The first observation is that most of the datasets contain the examples of all four types. Moreover, a majority of datasets contains rather a small number of safe examples. There are even such datasets as cleveland, glass, hsv or solar-flare, which do not contain any safe examples. Most of the data is characterized by a large number of difficult examples. Let us try to categorize considered datasets depending on the dominating type of examples from the minority class.

Only in abdominal-pain, acl, new-thyroid and vehicle datasets, safe minority examples prevail. Therefore, we can treat these 4 datasets as representatives of *safe* datasets (category S).

In the next category the borderline examples dominate in the distribution of the minority class. As could be observed in Table 1, even in datasets with clean borders a considerable amount of examples (up to 36%) can be labelled as borderline ones. So, the percentage of borderline examples must be even higher to represent some overlapping between classes. We treat a dataset as a *borderline* dataset if it contains more than 50% of B examples – these are credit-g, ecoli, haberman, hepatitis. Two additional datasets – car and scrotal-pain – are located somewhere between S and B categories. As the amount of safe examples is too low, we decide to assign them to the B category.

Then, several datasets contain many rare examples. Although they are not that numerous as B or S examples, they constitute even 20-30% of the minority class. The R category includes haberman (also assigned to B category), cmc, breast-cancer, cleveland, glass, hsv and abalone datasets, which have at least 20% of rare examples. Other datasets contain less than 10% of these examples.

Finally, some datasets contain a relatively high number of outlier examples – sometimes more than a half of the whole minority class. We assign the dataset to O category if more than 20% of examples are labelled as outliers. In Table 2

[3] We are grateful to prof. W.Michalowski and the MET Research Group from the University of Ottawa for abdominal-pain and scrotal-pain datasets; and to prof. K. Slowinski from Poznan University of Medical Science for hsv and acl datasets.

Table 2. Labelling of real-world datasets

Dataset Description				Identified Labels				
Dataset	Abbrev.	Size	Imbalance Ratio [%]	Safe [%]	Border [%]	Rare [%]	Outlier [%]	Type
abdominal-pain	AP	723	27.94	59.90	22.28	8.90	7.92	S
acl	AC	140	28.57	67.50	30.00	0.00	2.50	S
new-thyroid	NT	215	16.28	68.57	31.43	0.00	0.00	S
vehicle	VE	846	23.52	74.37	24.62	0.00	1.01	S
car	CA	1728	3.99	47.83	39.13	8.70	4.35	B
scrotal-pain	SP	201	29.35	38.98	45.76	10.17	5.08	B
credit-g	CG	1000	30	9.33	63.67	10.33	16.67	B
ecoli	EC	336	10.42	28.57	54.29	2.86	14.29	B
hepatitis	HE	155	20.65	15.63	62.50	6.25	15.63	B
haberman	HA	306	26.47	4.94	61.73	18.52	14.81	B, R
cmc	CM	1473	22.61	17.72	44.44	18.32	19.52	R
breast-cancer	BC	286	29.72	24.71	25.88	32.94	16.47	R
cleveland	CL	303	11.55	0.00	31.43	17.14	51.43	R, O
glass	GL	214	7.94	0.00	35.29	35.29	29.41	R, O
hsv	HS	122	11.48	0.00	0.00	28.57	71.43	R, O
abalone	AB	4177	8.02	8.36	20.60	20.60	50.45	R, O
solar-flare	SF	1066	4.03	0.00	48.84	11.63	39.53	O
transfusion	TR	748	23.8	18.54	47.19	11.24	23.03	O
yeast	YE	1484	3.44	5.88	47.06	7.84	39.22	O

these datasets are listed from cleveland to yeast. For many datasets, R and O categories appear together.

This categorization can be partly backed up by the MDS visualisation. The three datasets visualised in Fig. 1(b)-1(d) also show that new-thyroid is a safe dataset, ecoli can be assigned to a B category, while cleveland represents R and O categories.

5 Impact of Different Data Categories on Classifiers

The analysis of Table 2 showed that most datasets are seriously disturbed with a large number of B, R and O examples (or a mixture of them) which should cause difficulties in recognizing the minority class. Thus, in the next experiment we study the influence of these examples on the performance of popular classifiers.

We have decided to choose classifiers which are often used in related experimental studies and are based on different principles[4]. These are: decision tree learner J48 (a WEKA implementation of C4.5 classifier), two rule learners PART and Ripper (JRip), k-nearest neighbour (kNN), Naive Bayes, neural network (RBF) and SVM (SMO version). We parametrize them in the following way. J48 and PART are used without pruning. For JRip we do not change standard

[4] All implementation comes from WEKA platform.

options. kNN is used with $k = 1, 3, 5$ as we want to study whether increasing k influences the classifier. Naive Bayes is used with a supervised discretization of numeric attributes option from the WEKA's implementation. Standard values of parameters for RBF and SVM have failed to recognize the minority class. For RBF we have scanned several configurations trying to get the best sensitivity measures on all datasets. As a result, we changed a number of clusters to 5 and minimum standard deviation to 0.1. The similar optimization has been done for the SVM classifiers. We have used RBF kernel function, and selected two best combinations of complexity C and gamma G parameters – $(C = 50, G = 1.0)$, further referred to as SVM1 and $(C = 30, G = 0.1)$, denoted as SVM2.

Performance of the classifiers is evaluated with *Sensitivity* (true positive rate or an accuracy of the minority class), *Specificity* (accuracy of the majority class) and their aggregation by the geometric mean (*G-mean*) [4]. Their values are estimated by means of a 10-fold stratified cross-validation repeated 5 times to reduce possible variance. Table 3 presents the sensitivity and Table 4 – G-mean, with respect to 4 categories of datasets, which we will discuss below.

Table 3. Sensitivity of real-world datasets [%]

	DS	**PART**	**J48**	**JRip**	**NB**	**1NN**	**3NN**	**5NN**	**RBF**	**SVM1**	**SVM2**
S	**AP**	72.6	69.8	72.5	81.9	76.4	78.5	77.5	75.0	63.8	71.8
	AC	80.0	85.5	84.5	82.0	72.0	78.5	73.0	84.0	79.5	82.5
	NT	93.3	92.2	86.7	89.3	96.3	90.2	86.7	99.5	96.8	89.8
	VE	88.3	87.0	89.0	95.9	89.1	87.9	86.5	88.0	97.2	95.2
B	**CA**	90.0	77.7	47.0	0.0	3.1	3.1	3.1	49.6	27.0	88.2
	CG	47.7	46.5	37.6	50.5	50.3	39.9	37.1	43.6	2.5	52.2
	EC	42.0	58.0	59.7	81.0	52.2	50.8	57.8	54.7	64.0	58.5
	HA	33.4	41.0	34.0	25.0	30.1	26.9	18.1	18.3	14.7	1.3
	HE	45.7	43.2	31.2	75.5	44.0	37.0	47.5	60.7	39.3	51.5
	SP	63.4	55.3	53.4	56.5	58.4	58.7	49.2	62.5	32.0	65.9
R	**AB**	18.8	30.4	29.7	33.1	20.5	16.5	13.7	12.3	9.1	0.2
	BC	41.1	38.7	32.4	43.4	40.4	27.6	26.1	40.8	7.1	45.3
	CL	25.2	23.7	6.3	45.5	20.3	12.5	4.2	9.5	12.5	9.0
	CM	37.7	39.2	30.0	44.6	37.6	33.8	30.8	12.1	24.9	5.2
	GL	34.0	30.0	7.0	0.0	30.0	16.0	1.0	25.0	0.0	0.0
	HA	33.4	41.0	34.0	25.0	30.1	26.9	18.1	18.3	14.7	1.3
	HS	2.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
O	**AB**	18.8	30.4	29.7	33.1	20.5	16.5	13.7	12.3	9.1	0.2
	CL	25.2	23.7	6.3	45.5	20.3	12.5	4.2	9.5	12.5	9.0
	GL	34.0	30.0	7.0	0.0	30.0	16.0	1.0	25.0	0.0	0.0
	HS	2.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
	SF	18.7	20.9	3.7	60.6	9.1	8.2	0.0	10.2	0.0	15.7
	TR	42.9	41.3	39.7	51.5	31.9	34.3	31.9	32.9	8.5	2.2
	YE	26.7	30.9	36.7	42.9	38.1	26.2	19.4	15.1	7.9	0.0

Category S
All classifiers can learn the minority class quite well – they recognize 70-90% of the minority examples, regardless of the used parameters. The G-mean values are also very high. It is difficult to appoint the best classifier. This confirms the results of the earlier experiments with artificial datasets, where two classes have been clearly separated, see e.g. [9].

Category B
Overlapping of examples causes more difficulty for the classifiers – they can usually recognize 30-50% of the minority class. However, the borderline examples influence classifiers in a different degree. Naive Bayes seems to work well on these datasets – it usually gives the highest sensitivity (except for car and haberman datasets). J48 and PART, which have quite similar learning strategies, also give very good results – often not as high as Naive Bayes, but more stable (they do not decrease on some data). The same refers to the RBF network. JRip also performs rather stable, but is usually worse than J48 and PART. For kNN, different settings of k result in a difference of about 10%, and it is difficult to say which values of k is the best. Generally, kNN works rather well on borderline datasets (with the exception of car dataset). SVM can achieve good results, but its performance depends on the used parameters. It seems that SVM2 is better for borderline datasets (apart from haberman dataset). By analysing Specificity and G-mean, we observed that overlapping affected less the majority class (Specificity values ranged between 80-90%); this is consistent with the conclusions from [3].

Category R
Datasets with many rare examples seem to be more difficult than borderline datasets – the average recognition of the minority class ranges between 0% and 40%. Again, Naive Bayes can give the best results, but it sometimes fails (e.g. on glass). One can also notice a more visible decrease of G-mean. Symbolic classifiers (PART, J48, JRip) perform relatively well and, what is important, their classification abilities are more stable (although JRip performs worse on glass and cleveland). KNN's performance is heavily sensitive to k. 1NN definitely dominates other configurations. We noticed that rare examples form groups of two/three examples, which can be correctly classified using one neighbour, while using $k = 5$ increases the probability of finding a majority neighbour, which often results in a negative prediction. At the same time, our analysis of Specificity measure showed that 1NN tends to degrade the performance in the majority class more than other classifiers, which is consistent with the observations from the experiments with artificial datasets conducted in [3]. However, as this degradation is not that serious (few percents), it does not impact the G–mean measure. Finally, SVM classifier is not suited for this kind of data. Although SVM1 seems better than SVM2 (contrary to B datasets), it is still worse than other classifiers. RBF gives unstable results, but it works better than SVM.

Category O
The results show that O type of data is definitely the most difficult for all classifiers. They usually cannot recognize more than 30% of the minority examples and they often cannot recognize any examples from this class (e.g. hsv

dataset). Nevertheless, it is still reasonable to use Naive Bayes, J4.8, PART or 1NN. 3NN, 5NN, RBF and both SVMs usually cannot learn the minority class. As for a majority class, all the classifiers can recognize it in a similar degree, reaching 95–100% on Specificity. So, the tendencies observed for Sensitivity are also demonstrated by values of the G-mean measure. This also indicates that for difficult data distributions, the classifiers are especially strongly biased toward the majority classes.

Table 4. G–mean of real-world datasets [%]

	DS	**PART**	**J48**	**JRip**	**NB**	**1NN**	**3NN**	**5NN**	**RBF**	**SVM1**	**SVM2**
S	**AP**	78.6	78.1	80.0	85.7	79.8	82.6	82.8	82.6	76.8	79.9
	AC	84.8	89.1	88.4	87.5	81.2	86.6	83.7	88.8	85.0	87.8
	NT	95.3	94.3	91.6	93.4	97.3	93.9	92.1	99.1	97.6	94.3
	VE	91.9	91.3	92.2	80.6	92.1	91.9	91.4	89.7	98.0	96.4
B	**CA**	94.3	86.8	65.7	0.0	7.9	7.9	7.9	67.9	47.5	93.3
	CG	60.2	59.1	56.7	65.7	63.7	58.1	56.9	61.0	11.5	65.2
	EC	55.4	69.2	70.9	81.7	66.8	66.3	70.1	65.7	74.8	71.1
	HA	46.8	53.8	47.4	33.9	44.6	43.9	33.4	34.4	31.0	3.1
	HE	54.9	53.9	43.8	79.6	56.1	51.5	61.5	71.9	52.7	64.7
	SP	70.7	67.2	63.0	70.5	68.7	72.3	66.1	74.0	51.2	74.7
R	**AB**	41.9	53.9	53.2	55.3	43.2	38.8	35.8	32.2	28.2	1.4
	BC	52.9	53.1	50.6	58.9	56.1	47.3	47.5	56.7	17.8	59.0
	CL	38.2	34.3	10.6	60.2	30.7	22.2	8.1	16.0	18.6	14.1
	CM	54.3	56.9	51.7	59.4	53.8	53.0	51.7	32.2	46.0	20.0
	GL	40.7	36.2	8.9	0.0	36.2	20.0	1.4	29.8	0.0	0.0
	HA	46.8	53.8	47.4	33.9	44.6	43.9	33.4	34.4	31.0	3.1
	HS	2.8	0.0	0.0	0.0	0.0	0.0	0.0	1.3	0.0	0.0
O	**AB**	41.9	53.9	53.2	55.3	43.2	38.8	35.8	32.2	28.2	1.4
	CL	38.2	34.3	10.6	60.2	30.7	22.2	8.1	16.0	18.6	14.1
	GL	40.7	36.2	8.9	0.0	36.2	20.0	1.4	29.8	0.0	0.0
	HS	2.8	0.0	0.0	0.0	0.0	0.0	0.0	1.3	0.0	0.0
	SF	31.9	37.6	6.4	73.8	17.8	16.6	0.0	18.8	0.0	26.8
	TR	60.2	59.9	58.8	64.6	50.2	53.9	52.9	54.4	25.7	8.6
	YE	42.0	49.7	56.9	59.7	58.3	43.8	34.1	27.1	17.7	0.0

The observed differences between classifiers could be analysed more precisely on a single dataset, by focusing attention on the classification errors made on the particular classified example. In other words, we want to analyse the distribution of error rates over types/labels of examples. To get their valid estimations, the dataset has to be big enough, to assure that there is a sufficient number of examples for all four labels. We choose one of the biggest datasets, abalone. In Table 5 we present the error rates for all labels. Let us observe that for each classifier, the error rate rises from the left to the right, confirming that most of the errors occur for the difficult types of examples. SVM, RBF, 3NN and 5NN cannot predict any of the rare and outlier examples. However, Naive Bayes and

Table 5. Error rates on labeled testing examples for abalone dataset [%]

Classifier	Safe	Border	Rare	Outlier
J48	9.29	48.16	75.51	82.72
PART	25.71	72.24	87.64	89.59
JRip	0.71	50.61	75.96	84.50
NB	0.00	38.37	70.79	84.14
1NN	18.57	54.69	77.53	97.87
3NN	6.43	56.33	95.06	98.11
5NN	13.57	62.04	97.75	99.29
RBF	17.86	64.08	99.33	100.00
SVM1	30.71	80.00	98.65	99.88
SVM2	100.00	100.00	99.10	100.00

a decision tree can recognize some of these examples. Most classifiers classify rather well the safe examples (but for SVM a choice of parameters is crucial), while in the borderline region all classifiers can recognize some of the examples.

6 Conclusions

Distribution of examples in the minority class and its influence on learning classifiers is the main topic of our study. We distinguish four types of examples – besides safe examples, we focus our attention on borderline, rare and outlier examples. The method for identification of these examples in the data is proposed, which is based on the analysis of the local neighbourhood of learning examples.

Our experiments with real-world datasets show that most datasets contain many unsafe examples. The minority class is usually decomposed or scattered, with only a small number of safe regions. This observation is confirmed by analysing a 2D visualisation of datasets obtained by Multidimensional Scaling.

Moreover, the distribution of the minority class can be of different nature – it may consist of borderline, rare or outlier examples. We categorize the datasets depending on the dominating type of examples and study the performance of different classifiers. Our experiments show that safe datasets are generally quite easy for all considered classifiers. Borderline and, even more, rare or outlier datasets, are a real source of difficulties and they influence classifiers in a different degree. We could also observe that the imbalance ratio and the size of the data are not as influential as the above distribution types. Comparison of the abilities of different classifiers shows that Naive Bayes and J4.8 trees or PART rules are the most robust to unsafe types of the minority class examples – also for more difficult types. Performance of kNN depends on the type of examples (works better for borderline and rare examples) and k=1 is usually a better value, but it can adveresly affect the majority class. Then, RBF networks and SVM are quite sensitive to tuning parameters and fail to recognize rare or outliers examples.

Our observations are partly consistent with some earlier works with artificial datasets. In [3,5,8] it has also been shown that imbalance ratio is not the main source of difficulty. In conclusions from [3], it has been suggested that when there is a large overlapping between the classes, SVM is significantly worse than any other algorithm when the minority class recognition is concerned, while 1NN tends to degrade the majority class more than other classifiers.

However, these earlier works do not attempt to analyse real-world datasets. They also do not generalize the observations on classifier's performance. We think that it is worth looking for methods able to evaluate the nature of real-world datasets and their degree of difficulty. Such analysis can help to foresee the behaviour of classifiers and their possible sensitivity to the type of examples which prevail in an analysed dataset. Besides our proposals of using labelling and MDS visualisation, other approaches could be developed – see e.g. some new techniques of data visualisation recently studied in [7].

Acknowledgments. The research has been supported by the Ministry of Science and Higher Education, grant no. N N519 441939.

References

1. Cox, T., Cox, M.: Multidimensional Scaling. Chapman and Hall (1994)
2. Fernández, A., García, S., Herrera, F.: Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution. In: Corchado, E., Kurzyński, M., Woźniak, M. (eds.) HAIS 2011, Part I. LNCS (LNAI), vol. 6678, pp. 1–10. Springer, Heidelberg (2011)
3. García, V., Sánchez, J., Mollineda, R.A.: An Empirical Study of the Behaviour of Classifiers on Imbalanced and Overlapped Data Sets. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 397–406. Springer, Heidelberg (2007)
4. He, H., Garcia, E.: Learning from imbalanced data. IEEE Transactions on Data and Knowledge Engineering 21(9), 1263–1284 (2009)
5. Jo, T., Japkowicz, N.: Class Imbalances versus small disjuncts. ACM SIGKDD Explorations Newsletter 6(1), 40–49 (2004)
6. Kubat, M., Matwin, S.: Addresing the curse of imbalanced training sets: one-side selection. In: Proc. of Int. Conf. on Machine Learning ICML 1997, pp. 179–186 (1997)
7. Moreno-Torres, J.G., Herrera, F.: A Preliminary Study on Overlapping and Data Fracture in Imbalanced Domains by means of Genetic Programming-based Feature Extraction. In: Proc. of 10th Int. Conf. ISDA, pp. 501–506 (2010)
8. Napierała, K., Stefanowski, J., Wilk, S.: Learning from Imbalanced Data in Presence of Noisy and Borderline Examples. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS (LNAI), vol. 6086, pp. 158–167. Springer, Heidelberg (2010)
9. Prati, R., Batista, G., Monard, M.: Class imbalance versus class overlapping: an analysis of a learning system behavior. In: Proc. 3rd Mexican Int. Conf. on Artificial Intelligence, pp. 312–321 (2004)
10. Wilson, D.R., Martinez, T.R.: Improved heterogeneous distance functions. J. Artif. Intell. Res. 6, 1–34 (1997)