A Survey of Predictive Modelling under Imbalanced Distributions

Paula Branco[1,2], Luís Torgo[1,2], and Rita P. Ribeiro[1,2]

[1]LIAAD - INESC TEC
[2] DCC - Faculdade de Ciências - Universidade do Porto
paobranco@gmail.com, ltorgo@dcc.fc.up.pt, rpribeiro@dcc.fc.up.pt

May 14, 2015

Abstract

Many real world data mining applications involve obtaining predictive models using data sets with strongly imbalanced distributions of the target variable. Frequently, the least common values of this target variable are associated with events that are highly relevant for end users (e.g. fraud detection, unusual returns on stock markets, anticipation of catastrophes, etc.). Moreover, the events may have different costs and benefits, which when associated with the rarity of some of them on the available training data creates serious problems to predictive modelling techniques. This paper presents a survey of existing techniques for handling these important applications of predictive analytics. Although most of the existing work addresses classification tasks (nominal target variables), we also describe methods designed to handle similar problems within regression tasks (numeric target variables). In this survey we discuss the main challenges raised by imbalanced distributions, describe the main approaches to these problems, propose a taxonomy of these methods and refer to some related problems within predictive modelling.

1 Introduction

Predictive modelling is a data analysis task whose goal is to build a model of an unknown function $Y = f(X_1, X_2, \cdots, X_p)$, based on a training sample $\{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n}$ with examples of this function. Depending on the type of the variable Y, we face either a classification task (nominal Y) or a regression task (numeric Y). Models are obtained through an optimisation process that tries to find the "optimal" model parameters according to some criterion. The most frequent criteria are the error rate for classification and the mean squared error for regression. For some real world applications it is of key

importance that the obtained models are particularly accurate at some subrange of the domain of the target variable. Examples include diagnostic of rare diseases, forecasting rare extreme returns on financial markets, among many others. Frequently, these specific sub-ranges of the target variable are poorly represented on the available training sample. In these cases we face what is usually known as a problem of imbalanced data distributions, or imbalanced data sets. In other words, in these domains the cases that are more important for the user are rare and few exist on the available training set. The conjugation of the specific preferences of the user with the poor representation of these situations creates problems to modelling approaches at several levels. Namely, we typically need (i) special purpose evaluation metrics that are biased towards the performance of the models on these rare cases, and moreover, we need means for (ii) making the learning algorithms focus on these rare events. Without addressing these two questions, models will tend to be biased to the most frequent (and uninteresting for the user) cases, and the results of the "standard" evaluation metrics will not capture the competence of the models on these rare cases.

In this paper we provide a general definition for the problem of imbalanced domains that is suitable for both classification and regression tasks. We present an extensive survey of existing performance assessment measures and approaches to the problem of imbalanced data distributions. Existing surveys address only the problem of imbalanced domains for classification tasks (e.g. Kotsiantis et al. (2006); He and Garcia (2009); Sun et al. (2009)). Therefore, the coverage of performance assessment measures and approaches to tackle both classification and regression tasks is an innovative aspect of our paper. Another key feature of our work is the proposal of a broader taxonomy of methods for handling imbalanced domains. Our proposal extends previous taxonomies by including post-processing strategies.

The main contributions of this work are: i) provide a general definition of the problem of imbalanced domains suitable for classification and regression tasks; ii) review the main performance assessment measures for classification and regression tasks under imbalanced domains; iii) provide a taxonomy of existing approaches to tackle the problem of imbalanced domains both for classification and regression tasks; and iv) describe the most important techniques to address this problem.

The paper is organised as follows. Section 2 defines the problem of imbalanced data distributions and the type of existing approaches to address this problem. Section 3 describes several evaluation metrics that are biased towards performance assessment on the relevant cases in these domains. Section 4 provides a taxonomy of the modelling approaches to imbalanced domains, describing some of the most important techniques in each category. Finally, Section 5 explores some problems related with imbalanced domains and Section 6 concludes the paper.

2 Problem Definition

As we have mentioned before the problem of imbalanced data distributions occurs in the context of predictive tasks where the goal is to obtain a good approximation of the unknown function $Y = f(X_1, X_2, \cdots, X_p)$ that maps the values of a set of p predictor variables into the values of a target variable. These approximations to the function are obtained using a training data set $D = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^n$. At the center of the problem of imbalanced distribution is the fact that the user assigns more importance to the performance of the obtained approximation on a subset of the range of values of the target variable Y. Let us express this user preference bias by an importance or relevance function $\phi()$ that maps the values of the target variable into a range of importance, where 1 is maximal importance and 0 minimum relevance,

$$\phi(Y) : \mathcal{Y} \to [0, 1] \tag{1}$$

where $\mathcal{Y}$ is the domain of the target variable Y.

Suppose the user defines a relevance threshold t_R which sets the boundary above which the target variable values are relevant for the user. Let $D_R \in D$ be the subset of the training samples for which the relevance of the target value is high (or above t_R), i.e. $D_R = \{\langle \mathbf{x}_i, y_i \rangle \in D : \phi(y_i) > t_R\}$, and $D_N \in D$ be the subset of the training sample with the normal (or less important) cases, i.e $D_N = \{\langle \mathbf{x}_i, y_i \rangle \in D : \phi(y_i) \leq t_R\} = D \setminus D_R$.

The problem of imbalanced data sets can be described by the following assertions:

- $\phi(Y)$ is not uniform across the domain of Y

- The cardinality of the set of examples D_R is much smaller than the cardinality of D_N

- Standard evaluation criteria for both learning the models and evaluating their performance assume an uniform $\phi(Y)$, i.e. they are insensitive to $\phi(Y)$.

In this context, we potentially have a situation where the obtained models are sub-optimal with respect to the user-preference biases, and moreover, the metrics used to evaluate them are not in accordance with these biases and thus may be misleading.

Regarding the evaluation issue, traditional metrics are not adequate as they do not take into account the user preferences. Several solutions have been proposed to address this problem and overcome existing difficulties, mainly for classification tasks.

With respect to the inadequacy of the obtained models a large number of solutions has also appeared in the literature. We propose a categorisation of these approaches that considers three types of strategies: (i) modifications

on the learning algorithms, (ii) changes on the data before the the learning process takes place and finally (iii) transformations applied to the predictions of the learned models.

3   Performance Metrics for Imbalanced Domains

Obtaining a model from data can be seen as a search problem guided by an evaluation criterion that establishes a preference ordering among different alternatives. The main problem of imbalanced data sets lies on the fact that they are often associated with an user preference bias towards the performance on cases that are poorly represented in the available data sample. Standard evaluation criteria tend to focus the evaluation of the models on the most frequent cases, which is against the user preferences on these tasks. In fact, the use of common metrics in imbalanced domains can lead to sub-optimal classification models (He and Garcia, 2009; Weiss, 2004; Kubat and Matwin, 1997) and might produce misleading conclusions since these measures are insensitive to skewed domains (Ranawana and Palade, 2006; Daskalaki et al., 2006). As such, selecting proper evaluation metrics plays a key role in the task of correctly handling data imbalance. Adequate metrics should not only provide means to compare the models according to the user preferences, but can also be used to drive the learning of these models.

As the problem of imbalanced domains has been addressed mainly in classification problems, there are far more solutions for this type of tasks. We start by addressing the problem of evaluation metrics in classification and then move to regression.

Table 1 summarises the main references concerning performance assessment proposals for imbalanced domains in classification and regression.

Task type (Section)	Main References
Classification (3.1)	Estabrooks and Japkowicz (2001); Kubat et al. (1998); Bradley (1997) Provost et al. (1998); Davis and Goadrich (2006) García et al. (2008, 2009, 2010); Ranawana and Palade (2006) Batuwita and Palade (2009, 2012); Hand (2009); Thai-Nghe et al. (2011)
Regression (3.2)	Zellner (1986); Cain and Janssen (1995); Christoffersen and Diebold (1997) Crone et al. (2005); Lee (2008); Hernández-Orallo (2013) Bi and Bennett (2003); Torgo (2005); Torgo and Ribeiro (2007, 2009) Ribeiro (2011)

Table 1: Metrics for classification and regression, corresponding sections and main bibliographic references

3.1   Metrics for Classification Tasks

The confusion matrix for a two-class problem presents the results obtained by a given classifier (cf. Table 2). This table provides for each class the in-

		Predicted	
		Positive	Negative
True	Positive	TP	FN
	Negative	FP	TN

Table 2: Confusion matrix for a two-class problem.

stances that were correctly classified, i.e. the number of True Positives (TP) and True Negatives (TN), and the instances that were wrongly classified, i.e. the number of False Positives (FP) and False Negatives (FN).

Accuracy (cf. Equation 2) and its complement *error rate* are the most frequently used metrics for estimating the performance of learning systems in classification problems. For two-class problems, *accuracy* can be defined as follows,

$$accuracy = \frac{TP+TN}{TP+FN+TN+FP} \tag{2}$$

Considering a user preference bias towards the minority (positive) class examples, *accuracy* is not suitable because the impact of the least represented, but more important examples, is reduced when compared to that of the majority class. For instance, if we consider a problem where only 1% of the examples belong to the minority class, an high *accuracy* of 99% is achievable by predicting the majority class for all examples. Yet, all minority class examples, the rare and more interesting cases for the user, are misclassified. This is worthless when the goal is the identification of the rare cases.

The metrics used in imbalanced domains must consider the user preferences and, thus, should take into account the data distribution. To fulfill this goal several performance measures were proposed. From Table 2 the following measures (cf. Equations 3-8) can be obtained,

$$true\ positive\ rate\ (recall\ or\ sensitivity): TP_{rate} = \frac{TP}{TP+FN} \tag{3}$$

$$true\ negative\ rate\ (specificity\ ): TN_{rate} = \frac{TN}{TN+FP} \tag{4}$$

$$false\ positive\ rate: FP_{rate} = \frac{FP}{TN+FP} \tag{5}$$

$$false\ negative\ rate: FN_{rate} = \frac{FN}{TP+FN} \tag{6}$$

$$positive\ predictive\ value\ (precision\ ): PP_{value} = \frac{TP}{TP+FP} \tag{7}$$

$$negative\ predictive\ value: \ NP_{value} = \frac{TN}{TN+FN} \qquad (8)$$

However, as some of these measures exhibit a trade-off and it is impractical to simultaneously monitor several measures, new metrics have been developed, such as the *F-measure* (Rijsbergen, 1979),the *geometric mean* (Kubat et al., 1998) or the *receiver operating characteristic* (ROC) curve (Egan, 1975).

The *F-Measure* (F_{β}), a combination of both *precision* and *recall*, is defined as follows:

$$F_{\beta} = \frac{(1+\beta)^2 \cdot recall \cdot precision}{\beta^2 \cdot recall + precision} \qquad (9)$$

where β is a coefficient to adjust the relative importance of *recall* with respect to *precision* (if $\beta = 1$ *precision* and *recall* have the same weight, large values of β will increase the weight of *recall* whilst values less than 1 will give more importance to *precision*).

F_{β} is commonly used and is more informative about the effectiveness of a classifier on predicting correctly the cases that matter to the user (e.g. Estabrooks and Japkowicz (2001)). This metric value is high when both *recall* (a measure of completeness) and *precision* (a measure of exactness) are high.

An also frequently used metric when dealing with imbalanced data sets is the *geometric mean* (*G-Mean*) which is defined as:

$$G - Mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} = \sqrt{sensitivity \times specificity}$$
$$(10)$$

G-Mean is an interesting measure because it computes the *geometric mean* of the accuracies of the two classes, attempting to maximise them while obtaining good balance.

Two popular tools used in imbalanced domains are the *receiver operating characteristics* (ROC) curve (cf. Figure 1) and the corresponding area under the ROC curve (AUC) (Metz, 1978) . Provost et al. (1998) proposed ROC and AUC as alternatives to *accuracy*. The ROC curve allows the visualisation of the relative trade-off between benefits (TP_{rate}) and costs (FP_{rate}). The performance of a classifier for a certain distribution is represented by a single point in the ROC space. A ROC curve consists of several points each one corresponding to a different value of a decision/threshold parameter used for classifying an example as belonging to the positive class.

However, comparing several models through ROC curves is not an easy task unless one of the curves dominates all the others (Provost and Fawcett, 1997). Moreover, ROC curves do not provide a single-value performance score which motivates the use of AUC. The AUC (cf. Equation 11) allows
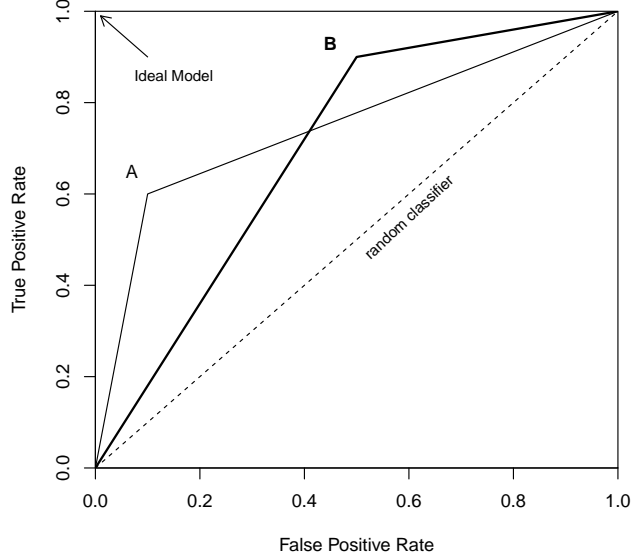
Figure 1: *ROC* curve of three classifiers: A, B and random.

the evaluation of the best model on average. Still, it is not biased towards the minority class.

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} = \frac{TP_{rate} + TN_{rate}}{2} \qquad (11)$$

Precision-recall curves (*PR curves*) are recommended for highly skewed domains where *ROC* curves may provide an excessively optimistic view of the performance (Davis and Goadrich, 2006). *PR curves* have the *recall* and *precision* rates represented on the axes. A strong relation between *PR* and *ROC* curves was found by Davis and Goadrich (2006).

Several other measures were proposed for dealing with some particular disadvantages of the previously mentioned metrics. For instance, a metric called *dominance* (García et al., 2008) (cf. Equation 12) was proposed to deal with the inability of *AUC* and *G-Mean* to explain how each class contributes to the overall performance.

$$dominance = TP_{rate} - TN_{rate} \qquad (12)$$

This measure ranges from -1 to $+1$.A value of $+1$ represents situations where perfect *accuracy* is achieved on the minority (positive) class, but all cases of the majority class are missed. A value of -1 corresponds to the opposite situation.

7

Another example is the *index of balanced accuracy* (*IBA*) (García et al., 2009, 2010) (cf. Equation 13) which quantifies a trade-off between an index of how balanced both class accuracies are and a chosen unbiased measure of overall *accuracy*.

$$IBA_\alpha(M) = (1 + \alpha \cdot dominance)M \tag{13}$$

where $(1 + \alpha \cdot dominance)$ is the weighting factor and M represents any performance metric.

Several other metrics exist such as *optimized precision* (Ranawana and Palade, 2006), *adjusted geometric mean* (Batuwita and Palade, 2009, 2012), *H-measure* (Hand, 2009) or *B42* (Thai-Nghe et al., 2011). All of them try to overcome some specific disadvantage detected in another metric when addressingthe challenge of assessing the performance in imbalanced domains.

3.2 Metrics for Regression Tasks

Very few efforts have been made regarding evaluation metrics for regression tasks in imbalanced domains. Performance measures commonly used in regression, such as *Mean Squared Error* (MSE) and *Mean Absolute Deviation* (MAD) (cf. Equations 14 and 15) are not adequate to these specific problems. These measures assume an uniform relevance of the target variable domain and evaluate only the magnitude of the error.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{14}$$

$$MAD = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{15}$$

Although the magnitude of the numeric error is important, for tasks with imbalanced distribution of the target variable, the metric must also be sensitive to the errors location within the target variable domain, because as in classification tasks, users of these domains are frequently biased to the performance on poorly represented values of the target. A simple solution, such as the introduction of weights, would not fulfil this goal because it would neglect the errors of predicting a rare value when it is a normal one (Ribeiro, 2011).

Within finance several attempts have been made for considering differentiated prediction costs through the proposal of asymmetric loss functions (Zellner, 1986; Cain and Janssen, 1995; Christoffersen and Diebold, 1996, 1997; Crone et al., 2005; Granger, 1999; Lee, 2008). However, the proposed solutions, such as *LIN-LIN* or *QUAD-EXP* error metrics, all suffer from the same problem: they can only distinguish between over- and underpredictions. Therefore, they are still unsuitable for addressing the problem

of imbalanced domains with a user preference bias towards some specific ranges of values.

Following the efforts made within classification, some attempts were made to adapt the existing notion of ROC curves to regression tasks. One of these attempts is the ROC *space for regression* ($RROC$ space) (Hernández-Orallo, 2013) which is motivated by the asymmetric loss often present on regression applications where both over-estimations and under-estimations entail different costs. $RROC$ space is defined by plotting the total over-estimation and under-estimation on the x-axis and y-axis, respectively (cf. Figure 2). $RROC$ curves are obtained when the notion of shift is used, which allows to adjust the model to an asymmetric operating condition by adding or subtracting a constant to the predictions. The notion of dominance can also be assessed by plotting the curves of different regression models, similarly to ROC curves in classification problems. Other evaluation metrics
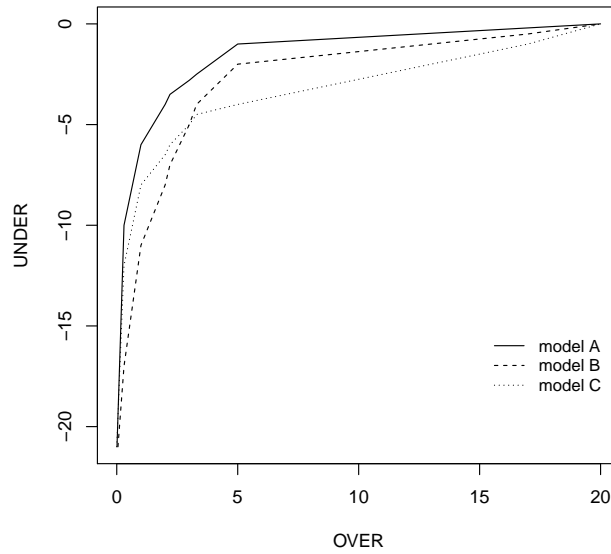

Figure 2: $RROC$ curve of three models: A, B and C.

were explored, such as the *Area Over the RROC curve* (AOC) which was shown to be equivalent to the error variance. In spite of the importance of this approach, it still only distinguishes over from under predictions.

Another relevant effort towards the adaptation of the concept of ROC curves to regression tasks was made by Bi and Bennett (2003) with the proposal of *Regression Error Characteristic* (REC) curves that provide a graphical representation of the cumulative distribution function (cdf) of the

9

error of a model. These curves plot the error tolerance and the accuracy of a regression function which is defined as the percentage of points predicted within a given tolerance ϵ. *REC* curves illustrate the predictive performance of a model across the range of possible errors (cf. Figure 3). The *Area Over the Curve* (*AOC*) can also be evaluated and is a biased estimate of the expected error of a model (Bi and Bennett, 2003). *REC* curves, although interesting, are still not sensitive to the error location across the target variable domain.
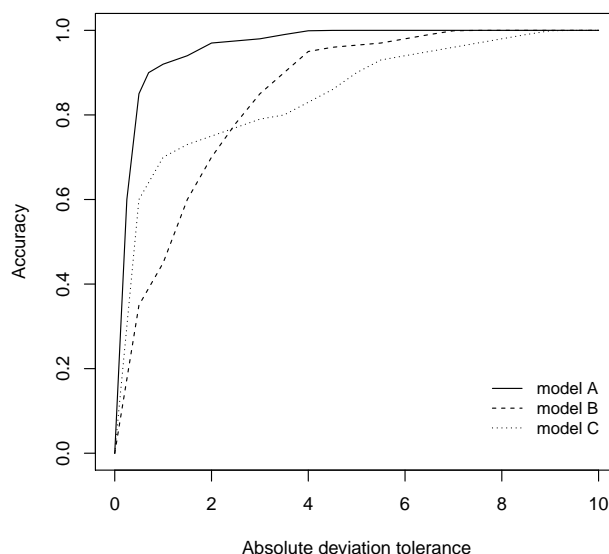

Figure 3: *REC* curve of three models: A, B and C.

To address this problem *Regression Error Characteristic Surfaces* (*RECS*) (Torgo, 2005) were proposed. These surfaces incorporate an additional dimension into *REC* curves representing the cumulative distribution of the target variable. *RECS* show how the errors corresponding to a certain point of the REC curve are distributed across the range of the target variable (cf. Figure 4). This tool allows the study of the behaviour of alternative models for certain specific values of the target variable. By zooming on specific regions of REC surfaces we can carry out two types of analysis that are highly relevant for some application domains. The first involves checking how certain values of prediction error are distributed across the domain of the target variable, which tells us where this type of errors are more frequent. The second type of analysis involves inspecting the type of errors a model has on a certain range of the target variable that is of particular
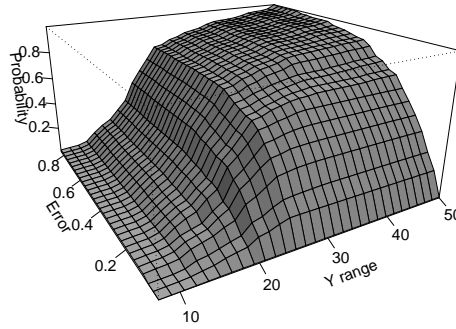
10

interest to us.

Figure 4: An example of the *REC* surface.

Another existing approach is the precision/recall evaluation framework, based on the concept of utility-based regression (Ribeiro, 2011; Torgo and Ribeiro, 2007). Utility-based regression establishes the notion of relevance of the target variable values and the existence of a non uniform relevance across the domain of this variable. In this context, the usefulness of a prediction dependes on both the numeric error of the prediction (which is provided by a certain loss function $L(\hat{y}, y)$) and the relevance (importance) of the predicted $\hat{y}$ and true y values. The relevance function, $\phi()$, is a continuous function as defined in Equation 1 which expresses the importance of the target variable values. Considering the goal of being accurate at rare extreme values, Ribeiro (2011) describes some methods for automatically obtaining these functions. The methods are based on the simple observation that, in these cases, the notion of relevance is inversely proportional to the target variable probability. Figure 5 shows an example of the relevance function ϕ in a data set where the high extreme values of the target variable are the most important, and Figure 6 shows the corresponding utility surface .

Using this utility-based framework, the notions of precision and recall were adapted to regression problems with non-uniform relevance of the target values by Torgo and Ribeiro (2009) and Ribeiro (2011). Ribeiro (2011) defines the notion of event using the concept of utility. In this context, the ratios of the two metrics are also defined as functions of utility, finally lead-
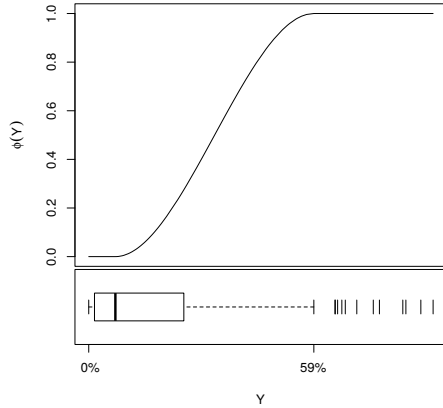
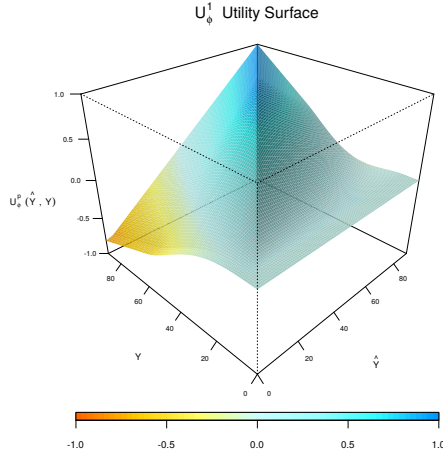Figure 5: Relevance function $\phi()$ automatically generated

Figure 6: Utility surface obtained with relevance function $\phi()$ shown in Figure 5

ing to definitions of *precision* and *recall* for regression[1]. The notion of utility led to the proposal of other measures, such as the *Mean Utility* and *Normalized Mean Utility* (Ribeiro, 2011). These metrics are derived from the utility and enable the comparison of different regression models according to the user preference bias.

4   Modelling Strategies for Handling Imbalanced Domains

Imbalanced domains raise significant challenges when building predictive models. The scarce representation of the most important cases leads to models that tend to be more focused on the normal examples, neglecting the rare events. Several strategies have been developed to address this problem, mainly in a classification setting. We propose that the existing approaches to learn under imbalanced data distributions can be grouped into the following four main categories:

- Data Pre-processing;

- Special-purpose Learning Methods;

- Prediction Post-processing;

- Hybrid Methods.

---

[1]Full details can be obtained in Chapter 4 of Ribeiro (2011).

Data Pre-processing approaches include solutions that pre-process the given imbalanced data set, changing the data distribution to make standard algorithms focus on the cases that are more relevant for the user. These methods have the following advantages: (i) can be applied to any existing learning tool; and (ii) the chosen models are biased to the goals of the user (because the data distribution was previously changed to match these goals), and thus it is expected that the models are more interpretable in terms of these goals. The main inconvenient of this strategy is that it may be difficult to relate the modifications in the data distribution with the target loss function.This means that mapping the given data distribution into an optimal new distribution according to the user goals is not easy.

Special-purpose learning methods comprise solutions that change the existing algorithms to be able to learn from imbalanced data. The following are important advantages: (i) the user goals are incorporated directly into the models; and (ii) it is expected that the models obtained this way are more comprehensible to the user. The main disadvantages of these approaches are: (i) the user is restricted in his choice to the learning algorithms that have been modified to be able to optimise his goals, or has to develop new algorithms for the task; (ii) if the target loss function changes, the model must be relearned, and moreover, it may be necessary to introduce further modifications in the algorithm which may not be straightforward; and (iii) it requires a deep knowledge of the learning algorithms implementations.

Prediction Post-processing approaches use the original data set and a standard learning algorithm, only manipulating the predictions of the models according to the user preferences and the imbalance of the data. As advantages, we can enumerate that: (i) it is not necessary to be aware of the user preference biases at learning time; (ii) the obtained model can, in the future, be applied to different deployment scenarios (i.e. different loss functions), without the need of re-learning the models or even keeping the training data available; and (iii) any standard learning tool can be used. However, these methods also have some drawbacks: (i) the models do not reflect the user preferences; (ii) the models interpretability is meaningless as they were obtained optimising a loss function that is not in accordance with the user preference bias.

Approaches following these three types of strategies will be reviewed in Sections 4.1, 4.2 and 4.3, and will include solutions for both classification and regression tasks. In Section 4.4 hybrid solutions will be addressed. Hybrid methods combine approaches of different types trying to take advantage of their best characteristics. Figure 7 synthesizes the different existing approaches within each of the categories.
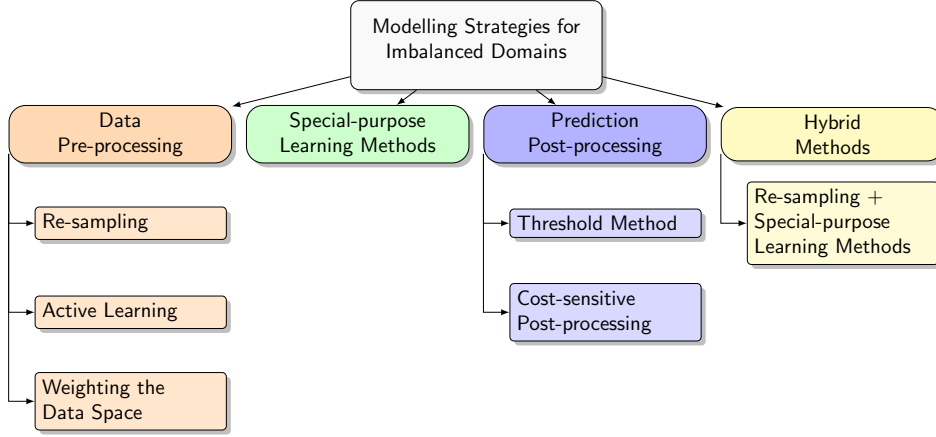
Figure 7: Main modelling strategies for imbalanced domains.

4.1 Data Pre-processing

Pre-processing strategies consist of methods of using the available data set in a way that is more in accordance with the user preference biases. This means that instead of applying a learning algorithm directly to the provided training data, we will first somehow pre-process this data according to the goals of the user. Any standard learning algorithm can be applied to the pre-processed data set.

Existing data pre-processing approaches can be grouped into three main types:

- **re-sampling:** change the data distribution of the data set forcing the learner to focus on the least represented examples;

- **active learning:** actively selecting the best (more valuable) samples to learn, leaving the ones with less information to improve the learner performance;

- **weighting the data space:** modify the training set distribution using information concerning misclassification costs, such that the learned model avoids costly errors.

Table 3 summarizes the main bibliographic references for data pre-processing strategies.

4.1.1 Re-sampling

Applying re-sampling strategies to obtain a more balanced data distribution is an effective solution to the imbalance problem (Estabrooks et al., 2004; Batuwita and Palade, 2010a; Fernández et al., 2008, 2010).

Strategy type (Section)		Main References
Re-sampling (4.1.1)	Random Under/Over-sampling	Chawla et al. (2002); Drummond and Holte (2003) Estabrooks et al. (2004); Seiffert et al. (2010); Chen et al. (2004); Wang and Yao (2009); Chang et al. (2003); Tao et al. (2006); Torgo et al. (2013)
	Distance Based	Chyi (2003); Mani and Zhang (2003)
	Data Cleaning Based	Kubat and Matwin (1997); Laurikkala (2001); Batista et al. (2004); Naganjaneyulu and Kuppa (2013)
	Recognition Based	Chawla et al. (2004); Zhuang and Dai (2006b); Raskutti and Kowalczyk (2004); Japkowicz (2000); Bellinger et al. (2012); Lee and Cho (2006); Zhuang and Dai (2006a)
	Cluster Based	Jo and Japkowicz (2004); Yen and Lee (2006, 2009); Cohen et al. (2006)
	Synthesising New Data	Lee (1999, 2000); Chawla et al. (2002); Liu et al. (2007); Menardi and Torelli (2010); Chawla et al. (2003); Martínez-García et al. (2012); Wang and Yao (2009); Torgo et al. (2013)
	Adaptive Synthetic Sampling	Batista et al. (2004); Verbiest et al. (2012); Hu et al. (2009); Zhang et al. (2011); Barua et al. (2012); Ramentol et al. (2012b,a); Bunkhumpornpat et al. (2012); Nakamura et al. (2013); Bunkhumpornpat et al. (2009); Han et al. (2005); He et al. (2008); Maciejewski and Stefanowski (2011)
	Evolutionary Sampling	García et al. (2006a); Doucette and Heywood (2008); García and Herrera (2009); Drown et al. (2009); Del Castillo and Serrano (2004); Yong (2012); Maheshwari et al. (2011); García et al. (2012); Galar et al. (2013)
	Re-sampling Combinations	Stefanowski and Wilk (2008); Napierała et al. (2010); Songwattanasiri and Sinapiromsaran (2010); Yang and Gao (2012); Li et al. (2008); Vasu and Ravi (2011); Bunkhumpornpat et al. (2011); Jeatrakul et al. (2010); Liu et al. (2006); Mease et al. (2007); Chen et al. (2010)
Active Learning (4.1.2)		Ertekin et al. (2007b,a); Zhu and Hovy (2007) Ertekin (2013); Mi (2013)
Weighting the Data Space (4.1.3)		Zadrozny et al. (2003); Wang and Japkowicz (2010)

Table 3: Pre-processing strategy types, corresponding sections and main bibliographic references

However, changing the data distribution may not be as easy as expected. Decide what is the optimal distribution is not straightforward as it is a domain dependent decision. Moreover, it was proved for classification tasks that a perfectly balanced distribution does not always provide optimal results (Weiss and Provost, 2003). In this context, some solutions were proposed to find the right amount of re-sampling for a data set (Weiss and Provost, 2003; Chawla et al., 2005, 2008).

For classification problems, changing the class distribution of the training data improves classifiers performance on an imbalanced context because it imposes non-uniform misclassification costs. This equivalence between the two concepts of altering the data distribution and the misclassification cost ratio is well-known and was first pointed out by Breiman et al. (1984).

The existing re-sampling strategies are based on a diverse set of techniques such as: random under/over-sampling, distance methods, data cleaning approaches, clustering algorithms, synthesising new data or evolutionary algorithms. We now briefly describe the most significant re-sampling strategies.

Two of the most simple re-sampling approaches that can be applied are under- and over-sampling. The first one removes data from the original data set reducing the sample size, while the second one adds data increasing the sample size. In random under-sampling, a random set of majority class examples are discarded. This may eliminate useful examples leading to a worse performance. Oppositely, in random over-sampling, a random set of copies of minority class examples is added to the data. This may increase the likelihood of overfitting, specially for higher over-sampling rates (Chawla et al., 2002; Drummond and Holte, 2003). Moreover, it may decrease the classifier performance and increase the computational effort.

Random under-sampling was also used in the context of ensembles. Namely, it was combined with boosting (Seiffert et al., 2010), bagging (Wang and Yao, 2009; Chang et al., 2003; Tao et al., 2006) and was applied to both classes in random forests in a method named Balanced Random Forest (BRF) (Chen et al., 2004).

For regression tasks, Torgo et al. (2013) perform random under-sampling of the common values as a strategy for addressing the imbalance problem. This method uses a relevance function and an user defined threshold to determine which are the common and uninteresting values that should be under-sampled.

Despite the potential of randomly selecting examples, under- and over-sampling strategies can also be carried out by other, more informed, methods. For instance, under-sampling can be accomplished resorting to distance evaluations (Chyi, 2003; Mani and Zhang, 2003). These approaches perform under-sampling based on a certain distance criteria that determines which are the examples from the majority class to include in the training set. These strategies are very time consuming which is a major disadvantage, specially

when dealing with large data sets.

Under-sampling can also be achieved through data cleaning methods. The main goal of these methods is to identify possibly noisy examples or overlapping regions and then decide on the removal of examples. One of those methods uses Tomek links (Tomek, 1976) which consist of points that are each other's closest neighbours, but do not share the same class label. This method allows for two options: only remove Tomek links examples belonging to the majority class or eliminate Tomek links examples of both classes (Batista et al., 2004). The notion of Condensed Nearest Neighbour Rule (CNN) (Hart, 1968) was also applied to perform under-sampling (Kubat and Matwin, 1997). CNN is used to find a subset of examples consistent with the training set, i.e., a subset that correctly classifies the training examples using a 1-nearest neighbour classifier. CNN and Tomek links methods were combined in this order by Kubat and Matwin (1997) in a strategy called One-Sided-Selection (OSS), and in the reverse order in a proposal of Batista et al. (2004).

Recognition-based methods as one-class learning or autoencoders offer the possibility to perform the most extreme type of under-sampling where all the examples from the majority class are removed. In this type of approach, and contrary to discrimination-based inductive learning, the model is learned using only examples of the target class, and no counter examples are included. This lack of examples from the other class(es) is the key distinguishing feature between recognition-based and discrimination-based learning.

One-class learning tries to set up boundaries which surround the target concept. This method starts by measuring the similarity between the target class and an object. Classification is then performed using a threshold on the obtained similarity score. One-class learning methods have the disadvantage of requiring the tuning of the threshold imposed on the similarity. In fact, this is a sensitive issue because if we choose a too narrow threshold the minority class examples are disregarded. However, too wide thresholds may lead to including examples from the majority class. Therefore, establishing an efficient threshold is vital with this method. Also, some learners actually need examples from more than one class and are unable to adapt to this method. Despite all these possible disadvantages, recognition-based learning algorithms have been proved to provide good prediction performance in most domains. Developments made in this context include one-class SVMs (e.g. Schölkopf et al. (2001); Manevitz and Yousef (2002); Raskutti and Kowalczyk (2004); Zhuang and Dai (2006b,a); Lee and Cho (2006)) and the use of an autoencoder (or autoassociator) (e.g. Japkowicz et al. (1995); Japkowicz (2000)).

Bellinger et al. (2012) investigated the performance variations of binary and one-class classifiers for different levels of imbalance. The results on both artificial and real world data sets showed that as the level of imbal-

ance increased, the performance of binary classifiers decreased, whereas the performance of one-class classifiers stayed relatively stable.

Imbalanced domains can influence the performance and the efficiency of clustering algorithms (Xuan et al., 2013). However, due to their flexibility, several approaches appeared for dealing with imbalanced data sets using clustering methods . For instance, the cluster-based oversampling (CBO) algorithm proposed by Jo and Japkowicz (2004) addresses both the imbalance problem and the problem of small disjuncts. Small disjuncts are subclusters of a certain class which have a low coverage, i.e., classify only few examples (Holte et al., 1989). CBO consists of clustering the training data of each class separately with the k-means technique and then performing random over-sampling in each cluster. All majority class clusters are over-sampled until they reach the cardinality of the largest cluster of this class. Then the minority class clusters are over-sampled until both classes are balanced maintaining all minority class subclusters with the same number of examples. Several other proposals based on clustering techniques exist (e.g. Yen and Lee (2006, 2009); Cohen et al. (2006)).

Another important approach for dealing with the imbalance problem as a pre-processing step, is the generation of new synthetic data. Several methods exist for building new synthetic examples. Most of the proposals are focused on classification tasks. Synthesising new data has several known advantages (Chawla et al., 2002; Menardi and Torelli, 2010), namely: (i) reduces the risk of overfitting which is introduced when replicas of the examples are inserted in the training set; (ii) improves the ability of generalisation which was compromised by the over-sampling methods. The methods for synthesising new data can be organized in two groups: (i) one that uses interpolation of existing examples, and (ii) another that introduces perturbations.

A famous method that uses interpolation is the synthetic minority over-sampling technique - SMOTE (Chawla et al., 2002). SMOTE algorithm over-samples the minority class by generating new synthetic data. This technique is then combined with a certain percentage of random under-sampling of the majority class that depends on a user defined parameter. Artificial data is created using an interpolation strategy that introduces a new example along the line segment joining a seed example and one of its k minority class nearest neighbours. The number of minority class neighbours (k) is another user defined parameter. For each minority class example a certain number of examples is generated according to a predefined over-sampling percentage.

SMOTE algorithm has been applied with several different classifiers and was also integrated with boosting (Chawla et al., 2003) and bagging (Wang and Yao, 2009).

SMOTE generates synthetic examples with the positive class label disregarding the negative class examples which may lead to overgeneraliza-

tion (Yen and Lee, 2006; Maciejewski and Stefanowski, 2011; Yen and Lee, 2009). This strategy may be specially problematic in the case of highly skewed class distributions where the minority class examples are very sparse, thus resulting in a greater chance of class mixture.

The group of techniques that introduces perturbations for generating new data does not suffer from this problem. Lee (1999) proposed an over-sampling method that produces noisy replicates of the rare cases while keeping the majority class unchanged. The synthetic examples are generated by adding normally distributed noise to the minority class examples. This simple strategy was tested with success, and a new version was developed by Lee (2000). This new approach generates, for a given data set, multiple versions of training sets with added noise. Then, an average of multiple model estimates is obtained.

Another framework, named ROSE (Random Over Sampling Examples), for dealing with the problem of imbalanced classification was presented by Menardi and Torelli (2010) based on a smoothed bootstrap re-sampling technique. ROSE generates a more balanced and completely new data set from the given training set combining over- and under-sampling. One observation is draw from the training set by giving the same probability to both existing classes. A new example is generated in the neighbourhood of this observation, using a width for the neighbourhood determined by a chosen smoothing matrix.

Several other proposals exist for classification tasks (e.g. Liu et al. (2007); Martínez-García et al. (2012)). However, for regression problems only one method for generating new synthetic data was proposed. Torgo et al. (2013) have adapted the SMOTE algorithm to regression tasks. Three key components of the SMOTE algorithm required adaptation for regression: (i) how to define which are the relevant observations and the "normal" cases; (ii) how to generate the new synthetic examples (i.e. over-sampling); and (iii) how to determine the value of the target variable in the synthetic examples. Regarding the first issue, a relevance function and a user-specified threshold were used to define D_R and D_N sets. The observations in D_R are over-sampled, while cases in D_N are under-sampled. For the generation of new synthetic examples the same interpolation method used in SMOTE for classification was applied. Finally, the target value of each synthetic example was calculated as an weighted average of the target variable values of the two seed examples. The weights were calculated as an inverse function of the distance of the generated case to each of the two seed examples.

Some drawbacks identified in the SMOTE algorithm motivated the appearance of several variants of this method (Barua et al., 2012; Han et al., 2005; Bunkhumpornpat et al., 2009; Chawla et al., 2003; He et al., 2008; Maciejewski and Stefanowski, 2011; Ramentol et al., 2012b; Verbiest et al., 2012; Stefanowski and Wilk, 2007).

We can identify three main types of SMOTE variants: (i) application

19

of some pre- or post- processing before or after the use of SMOTE; (ii) apply SMOTE only in some selected regions of the input space; or (iii) introducing small modifications to the SMOTE algorithm. Most of the first type of SMOTE variants start by applying the SMOTE algorithm and, afterwards, use a post-processing mechanism for removing some data. Examples of this type of approaches include: SMOTE+Tomek (Batista et al., 2004), SMOTE+ENN (Batista et al., 2004), SMOTE+FRST (Ramentol et al., 2012b) or SMOTE+RSB (Ramentol et al., 2012a). An exception is the Fuzzy Rough Imbalanced Prototype Selection (FRIPS) (Verbiest et al., 2012) method that pre-processes the data set before applying the SMOTE algorithm. The second type of SMOTE variants only generates synthetic examples in specific regions that are considered useful for the learning algorithms. As the notion of what is a good region is not straightforward, several strategies were developed. Some of these variants focus the synthesising effort on the borders between classes while others try to find which are the harder to learn instances and concentrate on these ones. Examples of these approaches are: Borderline-SMOTE (Han et al., 2005), ADASYN (He et al., 2008), Modified Synthetic Minority Oversampling Technique (MSMOTE) (Hu et al., 2009), MWMOTE (Barua et al., 2012), FSMOTE (Zhang et al., 2011), among others. Regarding the last type of SMOTE variants, some modifications are introduced in the way SMOTE generates the synthetic examples. For instance, the synthetic examples may be generated closer or further apart from a seed depending on some measure. The following proposals are examples within this group: Safe-Level-SMOTE (Bunkhumpornpat et al., 2009), Safe Level Graph (Bunkhumpornpat and Subpaiboonkit, 2013), LN-SMOTE (Maciejewski and Stefanowski, 2011) and DBSMOTE (Bunkhumpornpat et al., 2012).

Another approach to re-sampling concerns the use of Evolutionary Algorithms (EA). These algorithms started to be applied to imbalanced domains as a strategy to perform under-sampling through a prototype selection (PS) procedure (e.g. García et al. (2006a); García and Herrera (2009)).

García et al. (2006a) made one of the first contributions with a new evolutionary method proposed for balancing the data set. The method presented uses a new fitness function designed to perform a prototype selection process. Some proposals have also emerged in the area of heuristics and metrics for improving several genetic programming classifiers performance in imbalanced domains (Doucette and Heywood, 2008).

However, EA have been used for more than under-sampling. More recently, Genetic Algorithms (GA) and clustering techniques were combined to perform both under and over-sampling (Maheshwari et al., 2011; Yong, 2012). Evolutionary under-sampling has also been combined with boosting (Galar et al., 2013).

Finally, several other interesting methods have appeared which combine some of the previous techniques (Stefanowski and Wilk, 2008; Bunkhumporn-

pat et al., 2011; Songwattanasiri and Sinapiromsaran, 2010; Yang and Gao, 2012). For instance, Jeatrakul et al. (2010) presents a method that uses Complementary Neural Networks (CMTNN) to perform under-sampling and combines it with SMOTE. The combination of strategies was also applied to ensembles (e.g. Liu et al. (2006); Mease et al. (2007); Chen et al. (2010)).

Some attention has also been given to SVMs, leading to proposals such as the one of Kang and Cho (2006) where an ensemble of under-sampled SVMs is presented. Multiple different training sets are built by sampling examplesfrom the majority class and combining them with the minority class examples. Each training set is used for training an individual SVM classifier. The ensemble is produced by aggregating the outputs of all individual classifiers. Another similar approach is the EnSVM (Liu et al., 2006) which adopts a rebalance strategy combining the over-sampling strategy of SMOTE algorithm and under-sampling to form a number of new training sets while using all the positive examples.Then, an ensemble of SVMs is built.

Several ensembles have been adapted and combined with re-sampling approaches to better tackle the problem of imbalanced domains. Essentially, for every type of ensembles, some attempt has been made. For a more complete review on ensembles for the class imbalance problem see Galar et al. (2012).

4.1.2 Active Learning

Active learning is a semi-supervised strategy in which the learning algorithm is able to interactively obtain information from the user. Although this method is traditionally used with unlabelled data, it can also be applied when all class labels are known. In this case, the active learning strategy provides the ability of actively selecting the best, i.e. the most informative, examples to learn from.

Several approaches for imbalanced domains based on active learning have been proposed (Ertekin et al., 2007b,a; Zhu and Hovy, 2007; Ertekin, 2013). These approaches are concentrated on SVM learning systems and are based on the fact that, for this type of learners, the most informative examples are the ones closest to the hyperplane.

This property is used to guide under-sampling by selecting the most informative examples , i.e., choosing the examples closer to the hyperplane.

More recent developments try to combine active learning with other techniques to further improve learners performance. Ertekin (2013) presents a novel adaptive over-sampling algorithm named Virtual Instances Resampling Technique Using Active Learning (VIRTUAL), that combines the benefits of over-sampling and active learning. Contrary to traditional re-sampling methods, which are applied before the training stage, VIRTUAL generates synthetic examples for the minority class during the training pro-

cess. Therefore, the need for a separate pre-processing step is discarded. In the context of learning with SVMs, VIRTUAL outperforms competitive over-sampling techniques both in terms of generalisation performance and computational complexity. Mi (2013) developed a method that combines SMOTE and active learning with SVMs.

Some efforts have also been made for integrating active learning with other classifiers. Hu (2012) proposed an active learning method for imbalance data using the Localized Generalization Error Model (L-GEM) of radial basis function neural networks (RBFNN).

4.1.3 Weighting the Data Space

The strategy of weighting the data space is a way of implementing cost-sensitive learning. In fact, misclassification costs are applied to the given data set with the goal of selecting the best training distribution. Essentially, this method is based on the fact that changing the original sampling distribution by multiplying each case by a factor that is proportional to its importance (relative cost), allows any standard learner to accomplish expected cost minimisation on the original distribution. Although it is a simple technique and easy to apply, it also has some drawbacks. There is a risk of model overfitting and it is also possible that the real cost values are unavailable which can introduce the extra difficulty of exploring effective cost setups.

This approach has a strong theoretical foundation, building on the *Translation Theorem* derived by Zadrozny et al. (2003). Namely, to obtain a modified distribution biased towards the costly classes, the training set distribution is modified with regards to misclassification costs. Zadrozny et al. (2003) presented two different ways of accomplishing this conversion: in a transparent box or in a black box way. In the first, the weights are provided to the classifier while for the second a careful subsampling is performed according to the same weights. The first approach cannot be applied to an arbitrary learner, while the second one results in severe overfitting if re-sampling with replacement is used. Thus, to overcome the drawbacks of the later approach, the authors have presented a method called *cost-proportionate rejection sampling* which accepts each example in the input sample with probability proportional to its associated weight.

Wang and Japkowicz (2010) proposes an ensemble of SVMs with asymmetric misclassification costs. The proposed system works by modifying the base classifier (SVM) using costs and uses boosting as the combination scheme.

Strategy type (Section)	Main References
Special-purpose Learning Methods (4.2)	Maloof (2003); Akbani et al. (2004); Tang et al. (2009); Weiguo et al. (2012); Zhou and Liu (2006); Oh (2011); Castro and de Pádua Braga (2013); Sun et al. (2007); Song et al. (2009); Chen et al. (2004); Joshi et al. (2001); Hwang et al. (2011); Alejo et al. (2007); Cao et al. (2013); Wu and Chang (2003); Imam et al. (2006); Tang and Zhang (2006); Batuwita and Palade (2010b) Li et al. (2009); Barandela et al. (2003); Huang et al. (2004); Liu et al. (2010); Tan et al. (2003); Cieslak et al. (2012); Rodríguez et al. (2012); Wu and Chang (2005); Xiao et al. (2012); Cieslak and Chawla (2008); Ribeiro (2011); Torgo and Ribeiro (2003); Ribeiro and Torgo (2003)

Table 4: Special-purpose Learning Methods, corresponding section and main bibliographic references

4.2   Special-purpose Learning Methods

The approaches at this level consist of solutions that modify existing algorithms to provide a better fit to the imbalanced training data. The task of developing a solution based on algorithm modifications is not an easy one. It requires a deep knowledge of both the learning algorithm and the target domain. In order to perform a modification on a selected algorithm, it is essential to understand why it fails when the distribution is skewed. Also, some of the adaptations assume that a cost/cost-benefit matrix is known for different error types, which is frequently not the case. On the other hand, these methods have the advantage of being very effective in the contexts for which they were designed.

Existing solutions for dealing with imbalanced domains at the learning level are focused on the introduction of modifications in the algorithm preference criteria.

Table 4 summarizes the main bibliographic references for strategies involving modifications of algorithms.

The incorporation of benefits and/or costs (negative benefits) in existing algorithms, as a way to express the utility of different predictions, is one of the known approaches to cope with imbalanced domains. This includes the well known cost-sensitive algorithms for classification tasks which directly incorporate costs in the learning process. In this case, the goal of the prediction task is to minimise the total cost, knowing that misclassified examples may have different costs. In an imbalanced context, the cost of misclassifying a minority class example is superior to the cost of misclassifying a majority class example and, usually, there is no cost associated with making a correct prediction.

The research literature includes several works describing the adaptation

of different classifiers in order to make them cost-sensitive. For decision trees, the impact of the incorporation of costs under imbalanced domains was addressed by Maloof (2003)Regarding support vector machines several ways of integrating costs have been considered such as assigning different penalties to false negatives and positives (Akbani et al., 2004) or including a weighted attribute strategy (Yuanhong et al., 2009) among others (Weiguo et al., 2012). Regarding neural networks, the possibility of making them cost-sensitive has also been considered (e.g. Zhou and Liu (2006); Alejo et al. (2007); Oh (2011)). A Cost-Sensitive Multilayer Perceptron (CSMLP) algorithm was proposed by Castro and de Pádua Braga (2013) for asymmetrical learning of MLPs via a modified (backpropagation) weight update rule. Cao et al. (2013) present a framework for improving the performance of cost-sensitive neural networks that uses Particle Swarm Optimization (PSO) for optimizing misclassification cost, feature subset and intrinsic structure parameters. Alejo et al. (2007) propose two strategies for dealing with imbalanced domains using RBF neural networks which include a cost function in the training phase.

Ensembles have also been considered in the cost-sensitive framework to handle imbalanced domains. Several ensemble methods have been successfully adapted to include costs during the learning phase. However, boosting was the most extensively explored. AdaBoost is the most representative algorithm of the boosting family. When the class distribution is imbalanced, AdaBoost biases the learning (through the weights) towards the majority class, as it contributes more to the overall accuracy. Several proposals appeared which modify AdaBoost weight update process by incorporating cost items so that examples from different classes are treated unequally. Important proposals in the context of imbalanced distributions are: RareBoost (Joshi et al., 2001), AdaC1, AdaC2 and AdaC3 (Sun et al., 2007), and BABoost (Song et al., 2009). All of them modify the AdaBoost algorithm by introducing costs in the used weight updating formula. These proposals differ in how they modify the update rule. Random Forests have also been adapted to better cope with imbalanced domains undergoing a cost-sensitive transformation. Chen et al. (2004) proposes a method called Weighted Random Forest (WRF) for dealing with highly-skewed class distributions based on the Random Forest algorithm. WRF strategy operates by assigning a higher misclassification cost to the minority class. For an extensive review on ensembles for handling class imbalance see Galar et al. (2012).

Several other solutions exist that also modify the preference criteria of the algorithms while not relying directly on the definition of a cost/cost-benefit matrix. Regarding SVMs, several proposals try to bias the algorithm so that the hyperplane is further away from the positive class because the skew associated with imbalanced data sets pushes the hyperplane closer to the positive class. Wu and Chang (2003) accomplish this with an algorithm

that changes the kernel function. Fuzzy Support Vector Machines for Class Imbalance Learning (FSVM-CIL) was a method proposed by Batuwita and Palade (2010b). This algorithm is based on an SVM variant for handling the problem of outliers and noise called FSVM (Lin and Wang, 2002)and improves it for also dealing with imbalanced data sets. Potential Support Vector Machine (P-SVM) differs from standard SVM learners by defining a new objective function and constraints. An improved P-SVM algorithm (Li et al., 2009) was proposed to better cope with imbalanced data sets.

k-NN learners were also adapted to cope with the imbalance problem. Barandela et al. (2003) present a weighted distance function to be used in the classification phase of k-NN without changing the class distribution. This method assigns different weights to the respective classes and not to the individual prototypes. Since more weight is given to the majority class, the distance to minority class examples becomes much lower than the distance to examples from the majority class. This biases the learner to find their nearest neighbour among examples of the minority class.

A new decision tree algorithm - Class Confidence Proportion Decision Tree (CCPDT) - was proposed by Liu et al. (2010). CCPDT is robust and insensitive to class distribution and generates rules that are statistically significant. The algorithm adopts a new proposed measure,called Class Confidence Proportion (CCP), which forms the basis of CCPDT. CCP measure is embedded in the information gain and used as the splitting criteria. In this algorithm, a new approach , using Fisher exact test, to prune branches of the tree that are not statistically significant is presented.

Hellinger distance was introduced as a decision tree splitting criterion to build Hellinger Distance Decision Trees (HDDT) (Cieslak and Chawla, 2008). This proposal was shown to be insensitive towards class distribution skewness. More recently, Cieslak et al. (2012) recommended the use of bagged HDDTs as the preferred method for dealing with imbalanced data sets when using decision trees.

For regression tasks, some works have addressed the problem of imbalanced domains by changing the splitting criteria of regression trees (e.g. Torgo and Ribeiro (2003); Ribeiro and Torgo (2003)).

In Wu and Chang (2005) the Kernel Boundary Alignment algorithm (KBA) is proposed. This method adjusts the boundary towards the majority class by modifying the kernel matrix generated by a kernel function according to the imbalanced data distribution.

An ensemble method for learning over multi-class imbalanced data sets, named ensemble Knowledge for Imbalance Sample Sets (eKISS), was proposed by Tan et al. (2003). This algorithm was specifically designed to increase classifiers sensitivity without losing the corresponding specificity. The eKISS approach combines the rules of the base classifiers to generate new classifiers for final decision making.

Recently, more sophisticated approaches were proposed as the Dynamic

Classifier Ensemble method for Imbalanced Data (DCEID) presented by Xiao et al. (2012). DCEID combines dynamic ensemble learning with cost-sensitive learning and is able to adaptively select the more appropriate ensemble approach.

For regression problems one work exists that is able to tackle the problem of imbalanced domains through an utility-based algorithm. The utility-based Rules (ubaRules) approach was proposed by Ribeiro (2011). ubaRules is an utility-based regression rule ensemble system designed for obtaining models biased according to a specific utility-based metric. The system main goal is to obtain accurate and interpretable predictions in the context of regression problems with non-uniform utility. It consists in two main steps: generation of different regression trees, which are converted to rule ensembles, and selection of the best rules to include in the final ensemble. An utility function is used as criterion at several stages of the algorithm.

All these algorithm modification strategies are specifically designed to address the problem of imbalanced domains and have great potential. However, some disadvantages exist, such as: i) an often unavailable cost/cost-benefit matrix; ii) the need of a deep knowledge of the selected learner to accomplish a good modification of the preference criteria and iii) the difficulty of using an already existing method with a different learning system which contrasts with pre-processing approaches.

4.3 Prediction Post-processing

For dealing with imbalanced domains at the post-processing level, we will consider two main types of solutions:

- **threshold method:** uses the ranking provided by a score, that expresses the degree to which an example is a member of a class, to produce several learners by varying the threshold for class membership;

- **cost-sensitive post-processing:** associates costs to prediction errors and minimizes the expected cost.

Table 5 summarizes the main bibliographic references of post-processing strategies.

4.3.1 Threshold Method

Some classifiers are named soft classifiers because they provide a score which expresses the degree to which an example is a member of a class. This score can, in fact, be used as a threshold to generate other classifiers. This task can be accomplished by varying the threshold for an example belonging to a class Weiss (2004). A study of this method (Maloof, 2003) concluded that the operations of moving the decision threshold, applying a sampling

Strategy type (Section)	Main References
Threshold Method (4.3.1)	Maloof (2003); Weiss (2004)
Cost-sensitive Post-processing (4.3.2)	Hernández-Orallo (2012, 2014)

Table 5: Post-processing strategy types, corresponding sections and main bibliographic references

strategy, and adjusting the cost matrix produce classifiers with the same performance.

4.3.2 Cost-sensitive Post-processing

Several methods exist for making models cost-sensitive in a post hoc manner. This type of strategy was mainly explored for classification tasks and aims at changing only the model predictions for making it cost-sensitive (e.g. Domingos (1999); Sinha and May (2004)). This means that these approaches could potentially be applicable to imbalanced data distributions. However, to the best of our knowledge, these methods have never been applied or evaluated on these tasks.

In regression , introducing costs at a post-processing level has only recently been proposed (Bansal et al., 2008; Zhao et al., 2011). It is an issue still under-explored with few limited solutions. Similarly to what happens in classification, no progress was yet made for evaluating these solutions in imbalanced domains. However, one interesting proposal called reframing (Hernández-Orallo, 2012, 2014) was recently presented. Although not developed specifically for imbalanced domains, this framework aims at adjusting the predictions of a previously built model to different data distributions. Therefore, it is also potentially suitable for being applied to the problem of imbalanced domains. The notion of reframing was established as the process of applying a previously built model to a new operating context by the proper transformation of inputs, outputs and patterns. The reframing framework acts at a post-processing level, changing the obtained predictions by adapting them to a different distribution.

The reframing method essentially consists of two steps:

- the conversion of any traditional crisp regression model with one parameter into a soft regression model with two parameters, seen as a normal conditional density estimator (NCDE), by the use of enrichment methods;

- the reframing of an enriched soft regression model to new contexts by an instance-dependent optimisation of the expected loss derived from the conditional normal distribution.

Strategy type (Section)	Main References
Re-sampling and Special-purpose Learning Methods (4.4.1)	Phua et al. (2004); Kotsiantis and Pintelas (2003); Estabrooks and Japkowicz (2001); Estabrooks et al. (2004); Yoon and Kwek (2005); Liu et al. (2009)

Table 6: Hybrid strategies, corresponding sections and main bibliographic references

4.4 Hybrid Methods

In recent years, several methods involving the combination of some of the basic approaches described in the previous sections, have appeared in the research literature. Due to their characteristics these methods can be seen as hybrid methods to handle imbalanced distributions. They try to capitalise on some of the main advantages of the different approaches we have described previously.

Existing hybrid approaches combine the use of re-sampling strategies with special-purpose learning algorithms. Table 6 summarizes the main bibliographic references concerning these strategies.

4.4.1 Re-sampling and Special-purpose Learning Methods

One of the first hybrid strategies was presented by Estabrooks and Japkowicz (2001) and Estabrooks et al. (2004). The motivation for this proposal is related to the fact that a perfectly balanced data may not be optimal and that the right amount of over/under-sample to apply is difficult to determine. To overcome these difficulties, a mixture-of-experts framework was proposed (Estabrooks and Japkowicz, 2001; Estabrooks et al., 2004) in an architecture with three levels: a classifier level, an expert level and an output level. The system has two experts in the expert level: an under-sampling expert and an over-sampling expert. The architecture incorporates 10 classifiers on the over-sampling expert and another 10 classifiers on the under-sampling expert. All these classifiers are trained in data sets re-sampled at different rates of over and under-sampling, respectively. At the classifier level an elimination strategy is applied for removing the learners that are considered unreliable according to a predefined test. Then a combination scheme is applied both at the expert and output levels. These combination schemes use the following simple heuristic: if one of the classifiers decides that the example is positive so does the expert, and if one of the two experts decides that the example is positive so does the output level. This strategy is clearly heavily biased towards the minority (positive) class.

A different idea involving re-sampling and the combination of different learners was proposed by Kotsiantis and Pintelas (2003). The proposed approach uses a facilitator agent and three learning agents each one with its own learning system. The facilitator starts by filtering the features of

the data set. The filtered data is then passed to the three learning agents. Each learning agent re-samples the data set, learns using the respective system (Naive Bayes, C4.5 and 5NN) and returns the predictions for each instance back to the facilitator agent. Finally, the facilitator makes the final prediction according to majority voting.

In the proposal of Phua et al. (2004) re-sampling is performed and then stacking and boosting are used together. The applied re-sampling strategy partitions the data set into eleven new data sets which include all the minority class examples and a portion of the majority class examples. The proposed system uses three different learners (Naive Bayes, C4.5 and back-propagation classifier) each one processing the eleven partitions of the data. Bagging is used to combine the classifiers trained by the same algorithm. Then stacking is used to combine the multiple classifiers generated by the different algorithms identifying the best mix of classifiers.

Other approaches combine pre-processing techniques with bagging and boosting, simultaneously, composing an ensemble of ensembles. EasyEnsemble and BalanceCascade algorithms (Liu et al., 2009) are examples of this type of approach. Both algorithms use bagging as the main ensemble method and use Adaboost for training each bag. As for the pre-processing technique, both construct balanced bags by randomly under-sampling examples from the majority class. In EasyEnsemble algorithm all Adaboost iterations can be performed simultaneously because each Adaboost ensemble uses a previously determined subset of the data. All the generated classifiers are combined for a final solution. On the other hand, in the BalanceCascade algorithm, after the Adaboost learning, the majority examples correctly classified with higher confidence are discarded from further iterations.

Wang (2008) presents an approach that combines the SMOTE algorithm with Biased-SVM (Veropoulos et al., 1999). The proposed approach applies the Biased-SVM in the imbalanced data and stores the obtained support vectors from both classes. Then SMOTE is used to over-sample the support vectors with two alternatives: only use the obtained support vectors or use the entire minority class. A final classification is obtained with the new data using the biased-SVM.

Finally, a strategy using a clustering method based on class purity maximization is proposed by Yoon and Kwek (2005). This method generates clusters of pure majority class examples and non-pure clusters based on the improvement of the clusters class purity. When the clusters are formed, all minority class examples are added to the non-pure clusters and a decision tree is built for each cluster. An unlabelled example is clustered according to the same algorithm. If it falls on a non-pure cluster, the decision tree committee votes the prediction, but if it falls on a pure majority class cluster the final prediction is the majority class.If the committee votes for a majority class prediction, then that will be the final prediction. Onn the other hand, if it is a minority class prediction, then the example will be submitted

to a final classifier which is constructed using a neural network.

5 Related Problems

In this section we describe some problems that frequently coexist with imbalanced data distributions and further contribute to degrade the performance of predictive models. These related problems have been addressed mainly within a classification setting. Problems such as small disjuncts, class overlap and small sample size, usually coexist with imbalanced classification domains and are also identified as possible causes of classifiers performance degradation (Weiss, 2004; He and Garcia, 2009; Sun et al., 2009). We will briefly describe the major developments made for the following related problems: class overlapping or class separability, small sample size and lack of density in the training set, high dimensionality of the data set, noisy data and small disjuncts.

The overlap problem occurs when a given region of the data space contains an identical number of training cases for each class. In this situation, a learner will have an increased difficulty in distinguishing between the classes present on the overlapping region. The problems of imbalanced data sets and overlapping regions were mostly treated separately. However, in the last decade, some attention was given to the relationship between these two problems (Prati et al., 2004a; García et al., 2006b). The combination of imbalanced domains with overlapping regions causes an important deterioration of the learner performance and both problems acting together produce much more difficulties than expected when considering their effects individually (Denil and Trappenberg, 2010). Recent works (Alejo Eleuterio et al., 2011; Alejo et al., 2013) presented combinations of solutions for handling, simultaneously, both the class imbalance and the class overlap problem and apply a blend of techniques for addressing these issues.

The small training set, or small sample problem, is also naturally related with imbalanced domains. In effect, having too few examples from the minority class will prevent the learner from capturing their characteristics and will hinder the generalisation capability of the algorithm. The relation between imbalanced domains and small sample problems was addressed by Japkowicz and Stephen (2002) and Jo and Japkowicz (2004), where it was highlighted that class imbalance degrades classification performance in small data sets although this loss of performance tends to gradually reduce as the training set size increases. As expected, the subconcepts defined by the minority class examples can be better learned if their number can be increased.

The small sample problem may trigger problems such as rare cases (Weiss, 2005), which bring an additional difficulty to the learning system. Rare examples are extremely scarce cases that are difficult to detect and use

for generalisation. The small training set problem may also be accompanied by a variable class distribution that may not match the target distribution. Forman and Cohen (2004) showed that, for imbalanced domains, obtaining a balanced training set is not the most favourable setting and classifiers performance can be greatly improved by non-random sampling that favours the minority class.

In some domains, such as text classification, the imbalance problem co-exists with high dimensional data sets, i.e., domains with a high number of predictors. The main challenge here is to adequately select features that contain the key information of the problem. Feature selection is recommended (Wasikowski and Chen, 2010) and is also pointed as the solution for addressing the class imbalance problem (Mladenic and Grobelnik, 1999; Zheng et al., 2004; Chen and Wasikowski, 2008; Van Der Putten and Van Someren, 2004; Forman, 2003). Several proposals exist for handling the imbalance problem in conjunction with the high dimensionality problem, all using a feature selection strategy (Zheng et al., 2004; Del Castillo and Serrano, 2004; Forman and Cohen, 2004; Chu et al., 2010).

Noise is a known factor that usually affects models performance. In imbalanced domains, noisy data has a greater impact on the least represented examples (Weiss, 2004). A recent study (Seiffert et al., 2011) on the effect of noise in a data set intrinsically characterised by the presence of both class imbalance and class noise concluded that, generally, class noise has a more significant impact on learners than imbalance. It was also noticed that the interaction between the level of imbalance and the level of noise within a data set is a significant factor, and that studying these two cal effects individually may not be sufficient.

One of the most studied related problems is the problem of small disjuncts which is associated to the imbalance in the subclusters of each class in the data set (Japkowicz, 2001; Jo and Japkowicz, 2004). When a subcluster has a low *coverage*, i.e., it classifies few examples, it is called small (Holte et al., 1989). Small disjuncts are a problem because the learners are typically biased towards classifying large disjuncts and therefore they will tend to overfit and misclassify the cases in the small disjuncts. This problem is often present along with the problem of class imbalance in real world data sets and the connection between the two problems is not yet well understood Jo and Japkowicz (2004). Due to the importance of these two problems, several works address the relation between the problem of small disjuncts and the class imbalance problem (Japkowicz, 2003; Weiss and Provost, 2003; Jo and Japkowicz, 2004; Pearson et al., 2003; Japkowicz, 2001; Prati et al., 2004b). A new metric called *error concentration* (Weiss and Hirsh, 2000) was proposed for evaluating the error concentration towards the smaller disjuncts. The work in Weiss (2010) analyses the impact of several factors on small disjuncts and in the error distribution across disjuncts. Among the studied factors are pruning, training-set size, noise and class imbalance. Regarding

pruning, it was not considered an effective strategy for dealing with small disjuncts in the presence of class imbalance (Prati et al., 2004b; Weiss, 2010). Weiss (2010) also concluded that even with a balanced data set, errors tend to be concentrated towards the smaller disjuncts. However, when there is class imbalance, the error concentration increases. Moreover, the increase in the class imbalance also increases the error concentration. Thus, class imbalance is partly responsible for the problem with small disjuncts, and artificially modifying the class distribution of the training data to be more balanced, causes a decrease in the error concentration.

All the considered problems coexist and are related with the imbalance problem. The conjunction of these problems with imbalanced domains tends to further degrade the classifiers performance and therefore this relationship should not be ignored.

6 Conclusions

Imbalanced domains pose important challenges to existing approaches to predictive modelling. In this paper we propose a formulation of the problem of modelling using imbalanced data sets that includes both classification and regression tasks. We present a survey of the state of the art solutions for obtaining and evaluating predictive models for both classification and regression tasks. We propose a new taxonomy for the existing approaches grouping them into: (i) data pre-processing, (ii) special-purpose learning methods and (iii) prediction post-processing.

Most existing solutions to modelling under imbalanced distributions are focused on classification tasks. This fact is also present on previous surveys of this important research area. In this paper, we propose the first survey that also addresses existing approaches to imbalanced data sets within regression tasks.

Finally, we describe some problems that are strongly related with imbalanced data distributions, highlighting works that explore the relationship of these other problems with imbalance data sets.

References

Akbani, R., Kwek, S., and Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. In *Machine Learning: ECML 2004*, pages 39–50. Springer.

Alejo, R., García, V., Sotoca, J. M., Mollineda, R. A., and Sánchez, J. S. (2007). Improving the performance of the rbf neural networks trained with imbalanced samples. In *Computational and Ambient Intelligence*, pages 162–169. Springer.

Alejo, R., Valdovinos, R. M., García, V., and Pacheco-Sanchez, J. (2013). A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. *Pattern Recognition Letters*, 34(4):380–388.

Alejo Eleuterio, R., Martínez Sotoca, J., García Jiménez, V., and Valdovinos Rosas, R. M. (2011). Back propagation with balanced mse cost function and nearest neighbor editing for handling class overlap and class imbalance.

Bansal, G., Sinha, A. P., and Zhao, H. (2008). Tuning data mining methods for cost-sensitive regression: a study in loan charge-off forecasting. *Journal of Management Information Systems*, 25(3):315–336.

Barandela, R., Sánchez, J. S., Garcıa, V., and Rangel, E. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3):849–851.

Barua, S., Islam, M., Yao, X., and Murase, K. (2012). Mwmote-majority weighted minority oversampling technique for imbalanced data set learning.

Batista, G. E., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29.

Batuwita, R. and Palade, V. (2009). A new performance measure for class imbalance learning. application to bioinformatics problems. In *Machine Learning and Applications, 2009. ICMLA'09. International Conference on*, pages 545–550. IEEE.

Batuwita, R. and Palade, V. (2010a). Efficient resampling methods for training support vector machines with imbalanced datasets. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–8. IEEE.

Batuwita, R. and Palade, V. (2010b). Fsvm-cil: fuzzy support vector machines for class imbalance learning. *Fuzzy Systems, IEEE Transactions on*, 18(3):558–571.

Batuwita, R. and Palade, V. (2012). Adjusted geometric-mean: a novel performance measure for imbalanced bioinformatics datasets learning. *Journal of Bioinformatics and Computational Biology*, 10(04).

Bellinger, C., Sharma, S., and Japkowicz, N. (2012). One-class versus binary classification: Which and when? In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 2, pages 102–106. IEEE.

Bi, J. and Bennett, K. P. (2003). Regression error characteristic curves. In *Proc. of the 20th Int. Conf. on Machine Learning*, pages 43–50.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees. wadsworth & brooks. *Monterey, CA*.

Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap, C. (2009). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Advances in Knowledge Discovery and Data Mining*, pages 475–482. Springer.

Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap, C. (2011). Mute: Majority under-sampling technique. In *Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on*, pages 1–4. IEEE.

Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap, C. (2012). Db-smote: Density-based synthetic minority over-sampling technique. *Applied Intelligence*, 36(3):664–684.

Bunkhumpornpat, C. and Subpaiboonkit, S. (2013). Safe level graph for synthetic minority over-sampling techniques. In *Communications and Information Technologies (ISCIT), 2013 13th International Symposium on*, pages 570–575. IEEE.

Cain, M. and Janssen, C. (1995). Real estate price prediction under asymmetric loss. *Annals of the Institute of Statistical Mathematics*, 47(3):401–414.

Cao, P., Zhao, D., and Zaïane, O. R. (2013). A pso-based cost-sensitive neural network for imbalanced data classification. In *Trends and Applications in Knowledge Discovery and Data Mining*, pages 452–463. Springer.

Castro, C. L. and de Pádua Braga, A. (2013). Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. *IEEE Trans. Neural Netw. Learning Syst.*, 24(6):888–899.

Chang, E. Y., Li, B., Wu, G., and Goh, K. (2003). Statistical learning for effective visual information retrieval. In *ICIP (3)*, pages 609–612.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *JAIR*, 16:321–357.

Chawla, N. V., Cieslak, D. A., Hall, L. O., and Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery*, 17(2):225–252.

Chawla, N. V., Hall, L. O., and Joshi, A. (2005). Wrapper-based computation and evaluation of sampling methods for imbalanced datasets. In *Proceedings of the 1st international workshop on Utility-based data mining*, pages 24–33. ACM.

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6.

Chawla, N. V., Lazarevic, A., Hall, L. O., and Bowyer, K. W. (2003). Smoteboost: Improving prediction of the minority class in boosting. In *Knowledge Discovery in Databases: PKDD 2003*, pages 107–119. Springer.

Chen, C., Liaw, A., and Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*.

Chen, S., He, H., and Garcia, E. A. (2010). Ramoboost: Ranked minority oversampling in boosting. *Neural Networks, IEEE Transactions on*, 21(10):1624–1642.

Chen, X.-w. and Wasikowski, M. (2008). Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 124–132. ACM.

Christoffersen, P. F. and Diebold, F. X. (1996). Further results on forecasting and model selection under asymmetric loss. *Journal of applied econometrics*, 11(5):561–571.

Christoffersen, P. F. and Diebold, F. X. (1997). Optimal prediction under asymmetric loss. *Econometric theory*, 13(06):808–817.

Chu, L., Gao, H., and Chang, W. (2010). A new feature weighting method based on probability distribution in imbalanced text classification. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*, volume 5, pages 2335–2339. IEEE.

Chyi, Y.-M. (2003). Classification analysis techniques for skewed class distribution problems. *Master Thesis, Department of Information Management, National Sun Yat-Sen University*.

Cieslak, D. A. and Chawla, N. V. (2008). Learning decision trees for unbalanced data. In *Machine Learning and Knowledge Discovery in Databases*, pages 241–256. Springer.

Cieslak, D. A., Hoens, T. R., Chawla, N. V., and Kegelmeyer, W. P. (2012). Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, 24(1):136–158.

Cohen, G., Hilario, M., Sax, H., Hugonnet, S., and Geissbuhler, A. (2006). Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence in Medicine*, 37(1):7–18.

Crone, S. F., Lessmann, S., and Stahlbock, R. (2005). Utility based data mining for time series analysis: cost-sensitive learning for neural network predictors. In *Proceedings of the 1st international workshop on Utility-based data mining*, pages 59–68. ACM.

Daskalaki, S., Kopanas, I., and Avouris, N. M. (2006). Evaluation of classifiers for an uneven class distribution problem. *Applied Artificial Intelligence*, 20(5):381–417.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *ICML'06: Proc. of the 23rd Int. Conf. on Machine Learning*, ACM ICPS, pages 233–240. ACM.

Del Castillo, M. D. and Serrano, J. I. (2004). A multistrategy approach for digital text categorization from imbalanced documents. *ACM SIGKDD Explorations Newsletter*, 6(1):70–79.

Denil, M. and Trappenberg, T. (2010). Overlap versus imbalance. In *Advances in Artificial Intelligence*, pages 220–231. Springer.

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *KDD'99: Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*, pages 155–164. ACM Press.

Doucette, J. and Heywood, M. I. (2008). Gp classification under imbalanced data sets: Active sub-sampling and auc approximation. In *Genetic Programming*, pages 266–277. Springer.

Drown, D. J., Khoshgoftaar, T. M., and Seliya, N. (2009). Evolutionary sampling and software quality modeling of high-assurance systems. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 39(5):1097–1107.

Drummond, C. and Holte, R. C. (2003). C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on Learning from Imbalanced Datasets II*, volume 11. Citeseer.

Egan, J. P. (1975). Signal detection theory and {ROC} analysis.

Ertekin, Ş. (2013). Adaptive oversampling for imbalanced data classification. In *Information Sciences and Systems 2013*, pages 261–269. Springer.

Ertekin, Ş., Huang, J., Bottou, L., and Giles, L. (2007a). Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 127–136. ACM.

Ertekin, Ş., Huang, J., and Giles, C. L. (2007b). Active learning for class imbalance problem. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 823–824. ACM.

Estabrooks, A. and Japkowicz, N. (2001). A mixture-of-experts framework for learning from imbalanced data sets. In *Advances in Intelligent Data Analysis*, pages 34–43. Springer.

Estabrooks, A., Jo, T., and Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36.

Fernández, A., del Jesus, M. J., and Herrera, F. (2010). On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets. *Information Sciences*, 180(8):1268–1291.

Fernández, A., García, S., del Jesus, M. J., and Herrera, F. (2008). A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets and Systems*, 159(18):2378–2398.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research*, 3:1289–1305.

Forman, G. and Cohen, I. (2004). Learning from little: Comparison of classifiers given little training. In *Knowledge Discovery in Databases: PKDD 2004*, pages 161–172. Springer.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., and Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(4):463–484.

Galar, M., Fernández, A., Barrenechea, E., and Herrera, F. (2013). Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition*.

García, Salvador Derrac, J., Triguero, I., Carmona, C. J., and Herrera, F. (2012). Evolutionary-based selection of generalized instances for imbalanced classification. *Knowledge-Based Systems*, 25(1):3–12.

García, S., Cano, J. R., Fernández, A., and Herrera, F. (2006a). A proposal of evolutionary prototype selection for class imbalance problems. In *Intelligent Data Engineering and Automated Learning–IDEAL 2006*, pages 1415–1423. Springer.

García, S. and Herrera, F. (2009). Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary Computation*, 17(3):275–306.

García, V., Alejo, R., Sánchez, J. S., Sotoca, J. M., and Mollineda, R. A. (2006b). Combined effects of class imbalance and class overlap on instance-based classification. In *Intelligent Data Engineering and Automated Learning–IDEAL 2006*, pages 371–378. Springer.

García, V., Mollineda, R. A., and Sánchez, J. S. (2008). A new performance evaluation method for two-class imbalanced problems. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 917–925. Springer.

García, V., Mollineda, R. A., and Sánchez, J. S. (2009). Index of balanced accuracy: A performance measure for skewed class distributions. In *Pattern Recognition and Image Analysis*, pages 441–448. Springer.

García, V., Mollineda, R. A., and Sánchez, J. S. (2010). Theoretical analysis of a performance measure for imbalanced data. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 617–620. IEEE.

Granger, C. W. (1999). Outline of forecast theory using generalized cost functions. *Spanish Economic Review*, 1(2):161–173.

Han, H., Wang, W.-Y., and Mao, B.-H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *Advances in intelligent computing*, pages 878–887. Springer.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine learning*, 77(1):103–123.

Hart, P. E. (1968). The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:515–516.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 1322–1328. IEEE.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284.

Hernández-Orallo, J. (2012). Soft (gaussian cde) regression models and loss functions. *arXiv preprint arXiv:1211.1043*.

Hernández-Orallo, J. (2013). {ROC} curves for regression. *Pattern Recognition*, 46(12):3395 – 3411.

Hernández-Orallo, J. (2014). Probabilistic reframing for cost-sensitive regression. *ACM Trans. Knowl. Discov. Data*, 8(4):17:1–17:55.

Holte, R. C., Acker, L. E., and Porter, B. W. (1989). Concept learning and the problem of small disjuncts. In *IJCAI*, volume 89, pages 813–818. Citeseer.

Hu, J. (2012). Active learning for imbalance problem using l-gem of rbfnn. In *ICMLC*, pages 490–495.

Hu, S., Liang, Y., Ma, L., and He, Y. (2009). Msmote: improving classification performance when training data is imbalanced. In *Computer Science and Engineering, 2009. WCSE'09. Second International Workshop on*, volume 2, pages 13–17. IEEE.

Huang, K., Yang, H., King, I., and Lyu, M. R. (2004). Learning classifiers from imbalanced data based on biased minimax probability machine. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–558. IEEE.

Hwang, J. P., Park, S., and Kim, E. (2011). A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function. *Expert Systems with Applications*, 38(7):8580–8585.

Imam, T., Ting, K. M., and Kamruzzaman, J. (2006). z-svm: An svm for improved classification of imbalanced data. In *AI 2006: Advances in Artificial Intelligence*, pages 264–273. Springer.

Japkowicz, N. (2000). Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, volume 68. Menlo Park, CA.

Japkowicz, N. (2001). Concept-learning in the presence of between-class and within-class imbalances. In *Advances in Artificial Intelligence*, pages 67–77. Springer.

Japkowicz, N. (2003). Class imbalances: are we focusing on the right issue. In *Workshop on Learning from Imbalanced Data Sets II*, volume 1723, page 63.

Japkowicz, N., Myers, C., and Gluck, M. (1995). A novelty detection approach to classification. In *IJCAI*, pages 518–523.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Jeatrakul, P., Wong, K. W., and Fung, C. C. (2010). Classification of imbalanced data by combining the complementary neural network and smote algorithm. In *Neural Information Processing. Models and Applications*, pages 152–159. Springer.

Jo, T. and Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, 6(1):40–49.

Joshi, M. V., Kumar, V., and Agarwal, R. C. (2001). Evaluating boosting algorithms to classify rare classes: Comparison and improvements. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 257–264. IEEE.

Kang, P. and Cho, S. (2006). Eus svms: Ensemble of under-sampled svms for data imbalance problems. In *Neural Information Processing*, pages 837–846. Springer.

Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36.

Kotsiantis, S. and Pintelas, P. (2003). Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics, Computing & Teleinformatics*, 1(1):46–55.

Kubat, M., Holte, R. C., and Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine learning*, 30(2-3):195–215.

Kubat, M. and Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. In *Proc. of the 14th Int. Conf. on Machine Learning*, pages 179–186. Morgan Kaufmann.

Laurikkala, J. (2001). *Improving identification of difficult small classes by balancing class distribution*. Springer.

Lee, H.-j. and Cho, S. (2006). The novelty detection approach for different degrees of class imbalance. In *Neural Information Processing*, pages 21–30. Springer.

Lee, S. S. (1999). Regularization in skewed binary classification. *Computational Statistics*, 14(2):277.

Lee, S. S. (2000). Noisy replication in skewed binary classification. *Computational statistics & data analysis*, 34(2):165–191.

Lee, T.-H. (2008). Loss functions in time series forecasting. *International encyclopedia of the social sciences*.

Li, C., Jing, C., and Xin-tao, G. (2009). An improved p-svm method used to deal with imbalanced data sets. In *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*, volume 1, pages 118–122. IEEE.

Li, P., Qiao, P.-L., and Liu, Y.-C. (2008). A hybrid re-sampling method for svm learning from imbalanced data sets. In *Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on*, volume 2, pages 65–69. IEEE.

Lin, C.-F. and Wang, S.-D. (2002). Fuzzy support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):464–471.

Liu, A., Ghosh, J., and Martin, C. E. (2007). Generative oversampling for mining imbalanced datasets. In *DMIN*, pages 66–72.

Liu, W., Chawla, S., Cieslak, D. A., and Chawla, N. V. (2010). A robust decision tree algorithm for imbalanced data sets. In *SDM*, volume 10, pages 766–777. SIAM.

Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(2):539–550.

Liu, Y., An, A., and Huang, X. (2006). Boosting prediction accuracy on imbalanced datasets with svm ensembles. In *Advances in Knowledge Discovery and Data Mining*, pages 107–118. Springer.

Maciejewski, T. and Stefanowski, J. (2011). Local neighbourhood extension of smote for mining imbalanced data. In *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on*, pages 104–111. IEEE.

Maheshwari, S., Agrawal, J., and Sharma, S. (2011). A new approach for classification of highly imbalanced datasets using evolutionary algorithms. *Intl. J. Sci. Eng. Res*, 2:1–5.

Maloof, M. A. (2003). Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML-2003 workshop on learning from imbalanced data sets II*, volume 2, pages 2–1.

Manevitz, L. and Yousef, M. (2002). One-class svms for document classification. *the Journal of machine Learning research*, 2:139–154.

Mani, I. and Zhang, J. (2003). knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of Workshop on Learning from Imbalanced Datasets*.

Martínez-García, J. M., Suárez-Araujo, C. P., and Báez, P. G. (2012). Sneom: a sanger network based extended over-sampling method. application to imbalanced biomedical datasets. In *Neural Information Processing*, pages 584–592. Springer.

Mease, D., Wyner, A., and Buja, A. (2007). Cost-weighted boosting with jittering and over/under-sampling: Jous-boost. *J. Machine Learning Research*, 8:409–439.

Menardi, G. and Torelli, N. (2010). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, pages 1–31.

Metz, C. E. (1978). Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier.

Mi, Y. (2013). Imbalanced classification based on active learning smote. *Research Journal of Applied Sciences*, 5.

Mladenic, D. and Grobelnik, M. (1999). Feature selection for unbalanced class distribution and naive bayes. In *ICML*, volume 99, pages 258–267.

Naganjaneyulu, S. and Kuppa, M. R. (2013). A novel framework for class imbalance learning using intelligent under-sampling. *Progress in Artificial Intelligence*, 2(1):73–84.

Nakamura, M., Kajiwara, Y., Otsuka, A., and Kimura, H. (2013). Lvq-smote–learning vector quantization based synthetic minority over–sampling technique for biomedical data. *BioData mining*, 6(1):16.

Napierała, K., Stefanowski, J., and Wilk, S. (2010). Learning from imbalanced data in presence of noisy and borderline examples. In *Rough Sets and Current Trends in Computing*, pages 158–167. Springer.

Oh, S.-H. (2011). Error back-propagation algorithm for classification of imbalanced data. *Neurocomputing*, 74(6):1058–1061.

Pearson, R., Goney, G., and Shwaber, J. (2003). Imbalanced clustering for microarray time-series. In *Proceedings of the ICML*, volume 3.

Phua, C., Alahakoon, D., and Lee, V. (2004). Minority report in fraud detection: classification of skewed data. *ACM SIGKDD Explorations Newsletter*, 6(1):50–59.

Prati, R. C., Batista, G. E., and Monard, M. C. (2004a). Class imbalances versus class overlapping: an analysis of a learning system behavior. In *MICAI 2004: Advances in Artificial Intelligence*, pages 312–321. Springer.

Prati, R. C., Batista, G. E., and Monard, M. C. (2004b). Learning with class skews and small disjuncts. In *Advances in Artificial Intelligence–SBIA 2004*, pages 296–306. Springer.

Provost, F. J. and Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *KDD*, volume 97, pages 43–48.

Provost, F. J., Fawcett, T., and Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *ICML'98: Proc. of the 15th Int. Conf. on Machine Learning*, pages 445–453. Morgan Kaufmann Publishers.

Ramentol, E., Caballero, Y., Bello, R., and Herrera, F. (2012a). Smote-rsb*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory. *Knowledge and Information Systems*, 33(2):245–265.

Ramentol, E., Verbiest, N., Bello, R., Caballero, Y., Cornelis, C., and Herrera, F. (2012b). Smote-frst: a new resampling method using fuzzy rough set theory. In *10th International FLINS conference on uncertainty modelling in knowledge engineering and decision making (to appear)*.

Ranawana, R. and Palade, V. (2006). Optimized precision-a new measure for classifier performance evaluation. In *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*, pages 2254–2261. IEEE.

Raskutti, B. and Kowalczyk, A. (2004). Extreme re-balancing for svms: a case study. *ACM Sigkdd Explorations Newsletter*, 6(1):60–69.

Ribeiro, R. P. (2011). *Utility-based Regression*. PhD thesis, Dep. Computer Science, Faculty of Sciences - University of Porto.

Ribeiro, R. P. and Torgo, L. (2003). Predicting harmful algae blooms. In *Progress in Artificial Intelligence*, pages 308–312. Springer.

Rijsbergen, C. V. (1979). Information retrieval. dept. of computer science, university of glasgow, 2nd edition.

Rodríguez, J. J., Díez-Pastor, J.-F., Maudes, J., and García-Osorio, C. (2012). Disturbing neighbors ensembles of trees for imbalanced data. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 2, pages 83–88. IEEE.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., and Folleco, A. (2011). An empirical study of the classification performance of learners on imbalanced and noisy software quality data. *Information Sciences*.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., and Napolitano, A. (2010). Rusboost: A hybrid approach to alleviating class imbalance. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 40(1):185–197.

Sinha, A. P. and May, J. H. (2004). Evaluating and tuning predictive data mining models using receiver operating characteristic curves. *Journal of Management Information Systems*, 21(3):249–280.

Song, J., Lu, X., and Wu, X. (2009). An improved adaboost algorithm for unbalanced classification data. In *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on*, volume 1, pages 109–113. IEEE.

Songwattanasiri, P. and Sinapiromsaran, K. (2010). Smoute: Synthetics minority over-sampling and under-sampling techniques for class imbalanced problem. In *Proceedings of the Annual International Conference on Computer Science Education: Innovation and Technology, Special Track: Knowledge Discovery*, pages 78–83.

Stefanowski, J. and Wilk, S. (2007). Improving rule based classifiers induced by modlem by selective pre-processing of imbalanced data. In *Proc. of the RSKD Workshop at ECML/PKDD, Warsaw*, pages 54–65.

Stefanowski, J. and Wilk, S. (2008). Selective pre-processing of imbalanced data for improving classification performance. In *Data Warehousing and Knowledge Discovery*, pages 283–292. Springer.

Sun, Y., Kamel, M. S., Wong, A. K., and Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378.

Sun, Y., Wong, A. K., and Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719.

Tan, A., Gilbert, D., and Deville, Y. (2003). Multi-class protein fold classification using a new ensemble machine learning approach.

Tang, Y. and Zhang, Y.-Q. (2006). Granular svm with repetitive under-sampling for highly imbalanced protein homology prediction. In *Granular Computing, 2006 IEEE International Conference on*, pages 457–460. IEEE.

Tang, Y., Zhang, Y.-Q., Chawla, N. V., and Krasser, S. (2009). Svms modeling for highly imbalanced classification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(1):281–288.

Tao, D., Tang, X., Li, X., and Wu, X. (2006). Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(7):1088–1099.

Thai-Nghe, N., Gantner, Z., and Schmidt-Thieme, L. (2011). A new evaluation measure for learning from imbalanced data. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 537–542. IEEE.

Tomek, I. (1976). Two modifications of cnn. *IEEE Trans. Syst. Man Cybern.*, (11):769–772.

Torgo, L. (2005). Regression error characteristic surfaces. In *KDD'05: Proc. of the 11th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 697–702. ACM Press.

Torgo, L. and Ribeiro, R. P. (2003). Predicting outliers. In *Knowledge Discovery in Databases: PKDD 2003*, pages 447–458. Springer.

Torgo, L. and Ribeiro, R. P. (2007). Utility-based regression. In *PKDD'07: Proc. of 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases*, pages 597–604. Springer.

Torgo, L. and Ribeiro, R. P. (2009). Precision and recall in regression. In *DS'09: 12th Int. Conf. on Discovery Science*, pages 332–346. Springer.

Torgo, L., Ribeiro, R. P., Pfahringer, B., and Branco, P. (2013). Smote for regression. In *Progress in Artificial Intelligence*, pages 378–389. Springer.

Van Der Putten, P. and Van Someren, M. (2004). A bias-variance analysis of a real world learning problem: The coil challenge 2000. *Machine Learning*, 57(1-2):177–195.

Vasu, M. and Ravi, V. (2011). A hybrid under-sampling approach for mining unbalanced datasets: applications to banking and insurance. *International Journal of Data Mining, Modelling and Management*, 3(1):75–105.

Verbiest, N., Ramentol, E., Cornelis, C., and Herrera, F. (2012). Improving smote with fuzzy rough prototype selection to detect noise in imbalanced classification data. In *Advances in Artificial Intelligence–IBERAMIA 2012*, pages 169–178. Springer.

Veropoulos, K., Campbell, C., and Cristianini, N. (1999). Controlling the sensitivity of support vector machines. In *Proceedings of the international joint conference on artificial intelligence*, volume 1999, pages 55–60. Citeseer.

Wang, B. X. and Japkowicz, N. (2010). Boosting support vector machines for imbalanced data sets. *Knowledge and information systems*, 25(1):1–20.

Wang, H.-Y. (2008). Combination approach of smote and biased-svm for imbalanced datasets. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 228–231. IEEE.

Wang, S. and Yao, X. (2009). Diversity analysis on imbalanced data sets by using ensemble models. In *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*, pages 324–331. IEEE.

Wasikowski, M. and Chen, X.-w. (2010). Combating the small sample class imbalance problem using feature selection. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1388–1400.

Weiguo, D., Li, W., Yiyang, W., and Zhong, Q. (2012). An improved svm-km model for imbalanced datasets. In *Industrial Control and Electronics Engineering (ICICEE), 2012 International Conference on*, pages 100–103. IEEE.

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *SIGKDD Explorations Newsletter*, 6(1):7–19.

Weiss, G. M. (2005). Mining with rare cases. In *Data Mining and Knowledge Discovery Handbook*, pages 765–776. Springer.

Weiss, G. M. (2010). The impact of small disjuncts on classifier learning. In *Data Mining*, pages 193–226. Springer.

Weiss, G. M. and Hirsh, H. (2000). A quantitative study of small disjuncts. In *AAAI/IAAI*, pages 665–670.

Weiss, G. M. and Provost, F. J. (2003). Learning when training data are costly: the effect of class distribution on tree induction. *J. Artif. Intell. Res.(JAIR)*, 19:315–354.

Wu, G. and Chang, E. Y. (2003). Class-boundary alignment for imbalanced dataset learning. In *ICML 2003 workshop on learning from imbalanced data sets II, Washington, DC*, pages 49–56.

Wu, G. and Chang, E. Y. (2005). Kba: Kernel boundary alignment considering imbalanced data distribution. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):786–795.

Xiao, J., Xie, L., He, C., and Jiang, X. (2012). Dynamic classifier ensemble model for customer classification with imbalanced class distribution. *Expert Systems with Applications*, 39(3):3668–3675.

Xuan, L., Zhigang, C., and Fan, Y. (2013). Exploring of clustering algorithm on class-imbalanced data. In *Computer Science & Education (ICCSE), 2013 8th International Conference on*, pages 89–93. IEEE.

Yang, Z. and Gao, D. (2012). An active under-sampling approach for imbalanced data classification. In *Computational Intelligence and Design (ISCID), 2012 Fifth International Symposium on*, volume 2, pages 270–273. IEEE.

Yen, S.-J. and Lee, Y.-S. (2006). Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In *Intelligent Control and Automation*, pages 731–740. Springer.

Yen, S.-J. and Lee, Y.-S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):5718–5727.

Yong, Y. (2012). The research of imbalanced data set of sample sampling method based on k-means cluster and genetic algorithm. *Energy Procedia*, 17:164–170.

Yoon, K. and Kwek, S. (2005). An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics. In *Hybrid Intelligent Systems, 2005. HIS'05. Fifth International Conference on*, pages 6–pp. IEEE.

Yuanhong, D., Hongchang, C., and Tao, P. (2009). Cost-sensitive support vector machine based on weighted attribute. In *Information Technology and Applications, 2009. IFITA'09. International Forum on*, volume 1, pages 690–692. IEEE.

Zadrozny, B., Langford, J., and Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 435–442. IEEE.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81(394):446–451.

Zhang, D., Liu, W., Gong, X., and Jin, H. (2011). A novel improved smote resampling algorithm based on fractal. *Journal of Computational Information Systems*, 7(6):2204–2211.

Zhao, H., Sinha, A. P., and Bansal, G. (2011). An extended tuning method for cost-sensitive regression and forecasting. *Decision Support Systems*, 51(3):372–383.

Zheng, Z., Wu, X., and Srihari, R. (2004). Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter*, 6(1):80–89.

Zhou, Z.-H. and Liu, X.-Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *Knowledge and Data Engineering, IEEE Transactions on*, 18(1):63–77.

Zhu, J. and Hovy, E. H. (2007). Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *EMNLP-CoNLL*, volume 7, pages 783–790.

Zhuang, L. and Dai, H. (2006a). Parameter estimation of one-class svm on imbalance text classification. In *Advances in Artificial Intelligence*, pages 538–549. Springer.

Zhuang, L. and Dai, H. (2006b). Parameter optimization of kernel-based one-class classifier on imbalance learning. *Journal of Computers*, 1(7):32–40.