Journal of Artificial Intelligence Research 61 (2018) 863-905 Submitted 06/17; published 04/18

SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary

Alberto Fernández ALBERTO@DECSAI.UGR.ES
Salvador García SALVAGL@DECSAI.UGR.ES
Francisco Herrera HERRERA@DECSAI.UGR.ES
Department of Computer Science and Artificial Intelligence
University of Granada, Spain

Nitesh V. Chawla NCHAWLA@ND.EDU
Department of Computer Science and Engineering
and Interdisciplinary Center
for Network Science & Applications
University of Notre Dame, IN, USA

Abstract

The Synthetic Minority Oversampling Technique (SMOTE) preprocessing algorithm is considered "de facto" standard in the framework of learning from imbalanced data. This is due to its simplicity in the design of the procedure, as well as its robustness when applied to different type of problems. Since its publication in 2002, SMOTE has proven successful in a variety of applications from several different domains. SMOTE has also inspired several approaches to counter the issue of class imbalance, and has also significantly contributed to new supervised learning paradigms, including multilabel classification, incremental learning, semi-supervised learning, multi-instance learning, among others. It is standard benchmark for learning from imbalanced data. It is also featured in a number of different software packages — from open source to commercial. In this paper, marking the fifteen year anniversary of SMOTE, we reflect on the SMOTE journey, discuss the current state of affairs with SMOTE, its applications, and also identify the next set of challenges to extend SMOTE for Big Data problems.

1. Introduction

In the 1990s as more data and applications of machine learning and data mining started to become prevalent, an important challenge emerged: how to achieve desired classification accuracy when dealing with data that had significantly skewed class distributions (Sun et al., 2009; He & Garcia, 2009; López et al., 2013; Branco et al., 2016; Cieslak et al., 2012; Hoens et al., 2012b; Hoens & Chawla, 2013; Lemaitre et al., 2017; Khan et al., 2018). Authors from several disciplines observed an unexpected behavior for standard classification algorithms over datasets with uneven class distributions (Anand, Mehrotra, Mohan, & Ranka, 1993; Bruzzone & Serpico, 1997; Kubat, Holte, & Matwin, 1998). In many cases, the specificity or local accuracy on the majority class examples overwhelmed the one achieved on the minority ones. This led to the beginning of an active area of research in machine learning, now termed as "learning from imbalanced data". It was in the beginning of the 2000's when the foundations of the topic were established during the first workshop on class imbalanced learning during the American Association for Artificial Intelligence Conference (Japkowicz & Holte, 2000). The second milestone was set in 2003 during the ICML-KDD Workshop on learning from imbalanced datasets, leading to a special issue on the topic (Chawla, Japkowicz, & Kolcz, 2004).

The significance of this area of research continues to grow largely driven by the challenging problem statements from different application areas (such as face recognition, software engineering, social media, social networks, and medical diagnosis), providing a novel and contemporaneous set of challenges to the machine learning and data science researchers (Krawczyk, 2016; Haixiang et al., 2017; Maua & Galinac Grbac, 2017; Zhang et al., 2017; Zuo et al., 2016; Lichtenwalter et al., 2010; Krawczyk et al., 2016; Bach et al., 2017; Cao et al., 2017a). The overarching question that researchers have been trying to solve is: how to push the boundaries of prediction on the underrepresented or minority classes while managing the trade-off with with false positives? The solution space has ranged from sampling approaches to new learning algorithms designed specifically for imbalanced data. The sampling approaches are broadly divided into two broad categories — undersampling or oversampling.

Undersampling techniques are known to provide a compact balanced training set that also reduces the cost of the learning stage. However, it also leads to some derived problems. First, it increases the variance of the classifier and ii) it produces warped posterior probabilities (Dal Pozzolo, Caelen, & Bontempi, 2015). It may also might discard some useful examples for the modeling of the classifier. Particularly when the ratio of imbalance is high, then more examples need to be removed leading to the problem of lack of data (Wasikowski & Chen, 2010). This may affect the generalization ability of the classifier. As a result, researchers developed oversampling methods that may not lead to reduction of majority class examples, and tackle the class imbalance issue by replicating the minority class examples.However, applying random oversampling only implies a higher weight or cost for the minority instances. Therefore, the correct modeling of those clusters of minority data by the classification algorithm might still be hard in the case of overlapping (García, Mollineda, & Sánchez, 2008; Cieslak & Chawla, 2008) or small disjuncts (Jo & Japkowicz, 2004).

In 2002, Chawla, Bowyer, Hall, and Kegelmeyer (2002) proposed a novel approach as an alternative to the standard random oversampling. The idea was to overcome the overfitting rendered by simply oversampling by replication, and assist the classifier to improve its generalization on the testing data. Instead of "weighting" data points, the basis of this new data preprocessing technique was to create new minority instances. This technique was titled Synthetic Minority Oversampling Technique, now widely known as SMOTE (Chawla et al., 2002). The basis of the SMOTE procedure was to carry out an interpolation among neighboring minority class instances. As such, it is able to increase the number of minority class instances by introducing new minority class examples in the neighborhood, thereby assisting the classifiers to improve its generalization capacity.

SMOTE preprocessing technique became a pioneer for the research community in imbalanced classification. Since its release, many extensions and alternatives have been proposed to improve its performance under different scenarios. Due to its popularity and influence, SMOTE is considered as one of the most influential data preprocessing/sampling algorithms in machine learning and data mining (García, Luengo, & Herrera, 2016). Some approaches combine SMOTE with data cleaning techniques (Batista, Prati, & Monard, 2004). Other authors focus on the inner procedure by modifying some of its components, such as the selection of the instances for new data generation (Han, Wang, & Mao, 2005), or the type of interpolation (Bunkhumpornpat, Sinapiromsaran, & Lursinsap, 2012), among others.

In this paper, we present a summary of SMOTE and its impact in the last 15 years, celebrate its contributions to machine learning and data mining, and present the next state of challenges to keep pushing the frontier on learning from imbalanced data. While we don't include a discussion on the over 5,370 citations of SMOTE (as of Feb 1st, 2018), we specifically focus this paper on enumerating various SMOTE extensions as well as discussing the road ahead. For example, we discuss the extensions of SMOTE to other learning paradigms, such as streaming data (Krawczyk et al., 2017; Brzezinski & Stefanowski, 2017; Hoens et al., 2011), incremental learning (Ditzler, Polikar, & Chawla, 2010), concept drift (Hoens & Chawla, 2012; Hoens, Polikar, & Chawla, 2012a), or multi-label/multi-instance classification tasks (Herrera et al., 2016a, 2016b), among others. We also present an analysis about potential scenarios within imbalanced data that require a deeper dive into application of SMOTE, such as the data intrinsic characteristics (López et al., 2013), including small disjuncts, overlapping classes, and so on. Finally, we posit challenges of imbalanced classification in Big Data problems (Fernandez, del Rio, Chawla, & Herrera, 2017). Our hope is that this paper provides a summative overview of SMOTE, its extensions, and challenges that remain to be addressed in the community.

This paper is organized as follows. Section 2 introduces the SMOTE algorithm. Then, Section 3 enumerates those extensions to the standard SMOTE that have been proposed along these years. Section 4 presents the use of SMOTE under different learning paradigms. The challenges and topics for future work on SMOTE based preprocessing algorithms are given in Section 5. Finally, Section 6 summarizes and concludes the paper.

2. Synthetic Minority Oversampling Technique

In this section, we will first point out the origins of the SMOTE algorithm, setting the context under which it was designed (Section 2.1). Then, we will describe its properties in detail in order to present the working procedure of this preprocessing approach (Section 2.2).

2.1 Why to Propose SMOTE

Chawla reminisces the origins of SMOTE to a classification problem that he was tackling as a graduate student in 2000. He was working on developing a classification algorithm to learn and predict about cancerous pixels — the mammography data discussed in the original paper. A basic decision tree classifier provided him with an accuracy of around 97%. His first reaction was celebratory, as he achieved over 97% accuracy on a problem that was presented as a challenge to him. However, that celebration was short-lived. He quickly realized that merely by guessing majority class, he would have achieved an accuracy of 97.68% (which was the majority class distribution in the original data). So he actually did worse than a majority class guess classifier. Moreover, the decision tree classifier performed poorly in the important task of predicting calcifications correctly. This, thus presented the challenge of: *how to improve the performance of the classifier on minority class instances?*

An accompanying challenge was a low tolerance of false positives, i.e. examples of the majority class identified as minority ones. That is, one had to achieve an appropriate trade-off between the true positives and false positives, and not just be overly aggressive in predicting minority class (cancerous pixels) to compensate for the 2.32% distribution. This was because there were costs associated with errors — every false negative bore the burden of misclassifying a cancer as non-cancer, and every false positive bore the cost of additional tests by misclassifying a non-cancer as a cancer. The errors clearly were not of equal types.

Chawla tried the standard tools in the research arsenal at that time — oversampling by replication and undersampling. Both of the approaches, while improving the performance, did not provide satisfactorily results. On further investigation, he noticed the challenge arising from overfitting the minority class instances because of oversampling. This observation led to the question of: *how to improve the generalization capacity of the underling classifier?* And thus SMOTE was created to synthetically generate new instances to provide new information to the learning algorithm to improve its predictability about the minority class instances. SMOTE provided statistically significantly superior performance on the mammography data, as well as several others, thus laying the foundation for learning from imbalanced datasets. Of course, SMOTE like other sampling approaches, faces the challenge of the sampling amount, which Chawla and his colleagues also tried to mitigate by developing a wrapper framework, akin to feature selection (Chawla, Cieslak, Hall, & Joshi, 2008).

2.2 SMOTE Description

The SMOTE algorithm carries out an oversampling approach to rebalance the original training set. Instead of applying a simple replication of the minority class instances, the key idea of SMOTE is to introduce synthetic examples. This new data is created by interpolation between several minority class instances that are within a defined neighborhood. For this reason, the procedure is said to be focused on the "feature space" rather than on the "data space", in other words, the algorithm is based on the values of the features and their relationship, instead of considering the data points as a whole. This has also led to studying the theoretical relationship between original and synthetic instances must be analyzed in depth, including the data dimensionality. Some properties such as variance and correlation in the data and feature space, as well as the relationship between training and test examples distribution must be considered (Blagus & Lusa, 2013). We will discuss these issues hereinafter in Section 5 A simple example of SMOTE is illustrated in Figure 1. An x_i minority class instance is selected as basis to create new synthetic data points. Based on a distance metric, several nearest neighbors of the same class (points x_{i1} to x_{i4}) are chosen from the training set. Finally, a randomized interpolation is carried out in order to obtain new instances r_1 to r_4.

The formal procedure works as follows. First, the total amount of oversampling N (an integer value) is set up, which can either be set-up to obtain an approximate 1:1 class distribution or discovered via a wrapper process (Chawla et al., 2008). Then, an iterative process is carried out, composed of several steps. First, a minority class instance is selected at random from the training set. Next, its K nearest neighbors (5 by default) are obtained. Finally, N of these K instances are randomly chosen to compute the new instances by

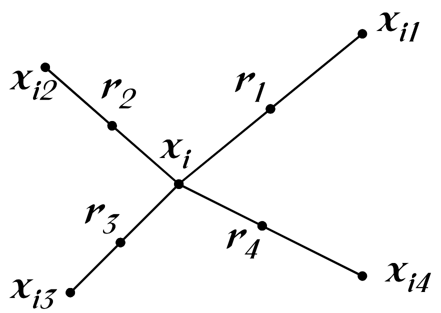


Figure 1: An illustration of how to create the synthetic data points in the SMOTE algorithm

interpolation. To do so, the difference between the feature vector (sample) under consideration and each of the selected neighbors is taken. This difference is multiplied by a random number drawn between 0 and 1, and then it is added to the previous feature vector. This causes the selection of a random point along the "line segment" between the features. In case of nominal attributes, one of the two values is selected at random. The whole process is summarized in Algorithm 1.

Algorithm 1 SMOTE algorithm

function SMOTE(T, N, k)
  Input: T; N; k ▷ #minority class examples, Amount of oversampling, #nearest neighbors
  Output: $(N/100)$ * T synthetic minority class samples
  Variables: $Sample$[][]: array for original minority class samples; $newindex$: keeps a count of number of synthetic samples generated, initialized to 0; $Synthetic$[][]: array for synthetic samples
  if $N < 100$ **then**
    Randomize the T minority class samples
    $T = (N/100)*T$
    $N = 100$
  end if
  $N = (int)N/100$ ▷ The amount of SMOTE is assumed to be in integral multiples of 100.
  for $i = 1$ to T **do**
    Compute k nearest neighbors for i, and save the indices in the $nnarray$
    POPULATE($N, i, nnarray$)
  end for
end function

Figure 2 shows a simple example of the SMOTE application in order to understand how synthetic instances are computed.

To conclude this section, we aim at introducing some of the first real applications that made a successful use of the SMOTE preprocessing algorithm, both of which are based on the area Bioinformatics. Specifically, we stress a multi-class problem of molecular functions

Algorithm 2 Function to generate synthetic samples

function Populate($N, i, nnarray$)
Input: N; i; $nnarray$ ▷ #instances to create, original sample index, array of nearest neighbors
Output: N new synthetic samples in $Synthetic$ array
    while $N \neq 0$ **do**
        nn = random(1,k)
        for attr = 1 to numattrs **do** ▷ $numattrs$ = Number of attributes
            Compute: $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$
            Compute: $gap = random(0,1)$
            $Synthetic[newindex][attr] = Sample[i][attr] + gap \cdot dif$
        end for
        $newindex++$
        $N--$
    end while
end function

```
Consider a sample (6,4) and let (4,3) be its nearest neighbor.
(6,4) is the sample for which k-nearest neighbors are being
identified (4,3) is one of its k-nearest neighbors.
Let: f1_1 = 6 f2_1 = 4,  f2_1 - f1_1 = -2
f1_2 = 4  f2_2 = 3,  f2_2 - f1_2 = -1
The new samples will be generated as
f1',f2' = (6,4) + rand(0-1) * (-2,-1)
rand(0-1) generates a vector of two random numbers between 0 and 1.
```

Figure 2: Example of the SMOTE application.

of yeast proteins (Hwang, Fotouhi, Finley Jr., & Grosky, 2003). The original problem was divided into imbalanced binary subsets, so that new synthetic instances were needed prior to the learning stage of a modular neural network to avoid the bias towards the majority classes.

3. Extensions to SMOTE

In the following, we present the most significant SMOTE-based approaches proposed in the last 15 years and a set of common properties shared by them. We consider SMOTE as a foundation for over-sampling with artificial generation of minority class instances. For this reason, we understand that any preprocessing method in the area of imbalanced classification that is based on the synthetic creation of examples by any type of interpolation or other process has some degree of relationship with the original SMOTE algorithm. First, in Section 3.1, the essential characteristics will be outlined. Next, in Section 3.2, we will enumerate all the extensions based on SMOTE proposed in the scientific literature until now. Then, each method will be categorized according to the studied properties to provide a comprehensive taxonomy. Next, in Section 3.3 we will present a list of SMOTE-based multiclassifiers proposed together with their categorization. Finally, Section 3.4 will outline the most influential experimental studies presented in the literature involving SMOTE as key point.

3.1 Properties for Categorizing the SMOTE-Based Extensions

This section provides a framework for the organization of the SMOTE-based extensions that will be presented in Sections 3.2 and 3.3. The aspects discussed here consist of (1) initial selection of instances to be oversampled, (2) integration with Undersampling as step in the technique, (3) type of interpolation, (4) operation with dimensionality changes, (5) adaptive generation of synthetic examples, (6) possibility of relabeling and (7) filtering of noisy generated instances. These mentioned facets are involved in the definition of the categorization, because they determine the way of operation of each technique. Next, we describe in detail each property.

- **Initial selection of instances to be oversampled:** It is usual to determine the best candidates to be oversampled in the data before the process of synthetic example generation starts. This strategy is intended to reduce the overlapping and noise in the final dataset. Many techniques opt to choose the instances near to the boundary classes (Han et al., 2005) or to not generate a synthetic example depending on the number of minority class examples belonging to the neighborhood (Bunkhumpornpat, Sinapiromsaran, & Lursinsap, 2009). Although many alternatives of initial selection have been proposed in the literature, almost all follow any of the two mentioned strategies. Two exceptions are the generation of synthetic examples after a LVQ optimization process (Nakamura, Kajiwara, Otsuka, & Kimura, 2013) and the selection of initial points from the support vectors obtained by a SVM (Cervantes, García-Lamont, Rodríguez-Mazahua, Chau, Ruiz-Castilla, & Trueba, 2017).

- **Integration with Undersampling:** The examples belonging to the majority class are also removed by either using a random or an informed technique of undersampling. The undersampling step can be either done at the beginning of the oversampling or as an internal operation together with the generation of synthetic examples. Generally, oversampling follows undersampling.

- **Type of interpolation:** This property offers varied mechanisms about generation of artificial or synthetic examples and is frequently associated with the main originality of the novel development. It defines the way new artificial examples are created and many alternatives can be found. The interpolation mechanisms can be *range restricted* (Han et al., 2005; Bunkhumpornpat et al., 2009; Maciejewski & Stefanowski, 2011), for example by looking not only for nearest neighbours from the minority class but also from majority class; creating new examples closer to the selected instance than its neighbor or by using feature weighting (Hukerikar, Tumma, Nikam, & Attar, 2011); *multiple* interpolations (de la Calleja & Fuentes, 2007; Gazzah & Amara, 2008) involving more than two examples or following topologies based on geometric shapes, such as ellipses (Abdi & Hashemi, 2016) and voronoi diagrams (Young, Nykl, Weckman, &

Chelberg, 2015), and graphs (Bunkhumpornpat et al., 2012); *clustering*-based interpolation (Barua, Islam, & Murase, 2011), in which the new examples can be either the centroids of the cluster or can be created involving examples that belong to the same cluster; interpolations that use different random distributions, such as the *gaussian* (Sandhan & Choi, 2014), estimations of the *probability distribution function* of the data (Gao, Hong, Chen, Harris, & Khalaf, 2014b), probability *smoothing* (Wang, Li, Chao, & Cao, 2012), *preservation of covariances* (Cateni, Colla, & Vannucci, 2011) among data and more complex interpolations, such as Markov chains (Das, Krishnan, & Cook, 2015) or Q-unions (Rong, Gong, & Ng, 2014). It is even possible to have no interpolation, such as when the new data is generated using only a single point, through jittering (Mease, Wyner, & Buja, 2007), gaussians disturbances (de la Calleja, Fuentes, & González, 2008), just simple copies with changes of label (Stefanowski & Wilk, 2008) or even by combining oversampling with pushing the majority samples out of a sphere (Koziarski, Krawczyk, & Wozniak, 2017).

- **Operation with dimensionality changes:** This occurs when the technique incorporates either a reduction or augmentation of dimensionality before or during the generation of artificial or synthetic examples. The most common approach is to change the dimensionality of the data at the beginning and then to work in the new dimensional space; either by reducing it through Principal Component Analysis (PCA) (Abdi & Hashemi, 2016) or related techniques (Gu, Cai, & Zhu, 2009; Xie, Jiang, Ye, & Li, 2015), feature selection (Koto, 2014), Bagging (Wang, Yun, li Huang, & ao Liu, 2013a), manifold techniques (Bellinger, Drummond, & Japkowicz, 2016) and auto-encoders (Bellinger, Japkowicz, & Drummond, 2015), and by using kernel functions (Mathew, Luo, Pang, & Chan, 2015; Tang & He, 2015; Pérez-Ortiz, Gutiérrez, Tiño, & Hervás-Martínez, 2016). Also, an estimation of the principal components of the data may be used to lead the interpolation (Tang & Chen, 2008).

- **Adaptive generation of synthetic examples:** The hypothesis of adaptive generation, ADASYN (He, Bai, Garcia, & Li, 2008), was to use a weighted distribution depending on each minority class example according to their degree of difficulty when learning. This way, more synthetic data will be generated for some minority class instances that are more complicated to learn compared to other. Inspired by ADASYN, lots of techniques incorporate similar mechanisms to control the quantity of new artificial examples to be generated associated with each minority example or subgroups of minority examples (Alejo, García, & Pacheco-Sánchez, 2015; Rivera, 2017).

- **Relabeling:** The technique offers the choice to relabel the examples belonging to the majority class during the synthetic generation of examples (Dang, Tran, Hirose, & Satou, 2015) or replacing the interpolation mechanism (Blaszczynski, Deckert, Stefanowski, & Wilk, 2012).

- **Filtering of noisy generated instances:** The first extensions of SMOTE motivated by its well known drawback of generating overlapped and noisy examples was the addition of a noise filtering step just after SMOTE process ends. Two typical techniques are SMOTE-TomekLinks and SMOTE+ENN (Batista et al., 2004). Filtering of artificial examples is a frequent operation that supports the success of SMOTE on real data.

Many kind of filters have been proposed for enhancing SMOTE, such as greedy filtering strategies (Puntumapon & Waiyamai, 2012), rough-sets based filtering (Ramentol, Caballero, Bello, & Herrera, 2012; Hu & Li, 2013; Ramentol, Gondres, Lajes, Bello, Caballero, Cornelis, & Herrera, 2016), ensembles-based filtering (Sáez, Luengo, Stefanowski, & Herrera, 2015) and bioinspired optimization procedures (López, Triguero, Carmona, García, & Herrera, 2014; Zieba, Tomczak, & Gonczarek, 2015; Jiang, Lu, & Xia, 2016; Cervantes et al., 2017).

3.2 SMOTE-Based Extensions for Oversampling

Till date, more than 85 extensions of SMOTE have been proposed in the specialized literature. This section is devoted to enumerate and categorize them according to the properties studied before. Table 1 presents an enumeration of the methods reviewed in this paper. In this field, it is usual that the authors provide a name for their proposal, with a few exceptions.

As we can see in Table 1, the most frequent properties exploited by the techniques are the initial selection and adaptive generation of synthetic examples. Filtering is becoming more common in recent years, as well as the use of kernel functions. Regarding the interpolation procedure, it is also usual to replace the original method with other more complex ones, such as clustering-based or derived from a probabilistic function. It is worth mentioning that there is no technique that applies the four mechanisms pertinent to the calibration of the generation of artificial examples, selection and removal of harmful examples either synthetic or belonging to the majority class; namely initial selection, integration with undersampling, adaptive generation and filtering all-together. Due to space limitations, it is not possible to describe all the reviewed techniques. Nevertheless, we will provide brief explanations for the most well-known techniques from Table 1:

- **Borderline-SMOTE** (Han et al., 2005): This algorithm draws from the premise of that the examples far from the borderline may contribute little to the classification success. Thus, the technique indentifies those examples which belong to the borderline by using the ratio between the majority and minority examples within the neighborhood of each instance to be oversampled. Noisy examples, those that have all the neighbours from the majority class, are not considered. The so-called dangerous examples, with a suitable ratio, are oversampled.

- **AHC** (Cohen et al., 2006): It was the first attempt to use clustering to generate new synthetic examples to balance the data. The K-means algorithm was used to undersample the majority examples and agglomerative hierarchical clustering was used to oversample the minority examples. Here, the clusters are gathered from all levels of the resulting dendograms and their centroids are interpolated with the original minority class examples.

- **ADASYN** (He et al., 2008): Its main idea proceeds from the assumption of utilizing a weighted distribution depending on the type of minority examples according to their complexity for learning. The quantity of synthetic data for each one is associated with the level of difficulty of each minority example. This difficulty estimation is based on

Table 1: Enumeration and categorization of SMOTE algorithm extensions

Ref.	Algorithm name	Initial selection	Integration US	Type of Interpolation	Dimensionality change	Adaptive generation	Relabeling	Filtering
(Batista et al., 2004)	SMOTE+TomekLinks							✓
(Batista et al., 2004)	SMOTE+ENN							✓
(Han et al., 2005)	Borderline-SMOTE	✓		Range restricted				
(Cohen, Hilario, Sax, Hugonnet, & Geissbühler, 2006)	AHC		✓	Clustering				
(Wang, Xu, Wang, & Zhang, 2006)	LLE-SMOTE				LLE			
(de la Calleja & Fuentes, 2007)	Distance-SMOTE			Multiple				
(de la Calleja et al., 2008)	SMMO	✓		Without-Gaussian				
(Gazzah & Amara, 2008)	Polynom-Fit-OS			Topologies				
(He et al., 2008)	ADASYN					✓		
(Stefanowski & Wilk, 2008)	< *no name* >	✓		Without-Copy			✓	
(Tang & Chen, 2008)	ADOMS			With PCA	PCA			
(Bunkhumpornpat et al., 2009)	Safe-Level-SMOTE	✓		Range restricted		✓		
(Gu et al., 2009)	Isomap-Hybrid		✓		MDS			
(Liang, Hu, Ma, & He, 2009)	MSMOTE	✓						
(Chen, Cai, Chen, & Gu, 2010a)	DE-Oversampling		✓	DE operators				
(Chen, Guo, & Chen, 2010c)	CE-SMOTE	✓						
(Kang & Won, 2010)	Edge-Det-SMOTE	✓						
(Barua et al., 2011)	CBSO			Clustering		✓		
(Cao & Wang, 2011)	SMOBD	✓				✓		
(Cateni et al., 2011)	SUNDO	✓	✓	Gaussian+Cov.				
(Deepa & Punithavalli, 2011)	E-SMOTE				FS with GA			
(Dong & Wang, 2011)	Random-SMOTE			Multiple				
(Fan, Tang, & Weise, 2011)	MSYN					✓		✓
(Fernández-Navarro, Hervás-Martínez, & Gutiérrez, 2011)	DSRBF					✓		
(Maciejewski & Stefanowski, 2011)	LN-SMOTE	✓		Range restricted				
(Zhang & Wang, 2011b)	Distribution-SMOTE	✓						
(Zhang & Wang, 2011a)	NDO-Sampling			Without-Gaussian				
(Bunkhumpornpat et al., 2012)	DBSMOTE	✓		Graph based				
(Farquad & Bose, 2012)	SVM-Balance						✓	
(Puntumapon & Waiyamai, 2012)	TRIM-SMOTE					✓		✓
(Ramentol et al., 2012)	SMOTE-RSB*							✓
(Wang et al., 2012)	ASMOBD	✓		Smoothing		✓		
(Barua, Islam, & Murase, 2013)	ProWSyn			Clustering		✓		
(Bunkhumpornpat & Subpaiboonkit, 2013)	SL-Graph-SMOTE	✓		Range restricted		✓		
(Hu & Li, 2013)	NRSBoundary-SMOTE							✓
(Li, Zou, Wang, & Xia, 2013b)	ISMOTE		✓					
(Nakamura et al., 2013)	LVQ-SMOTE	✓(LVQ)			FS			
(Pérez-Ortiz, Gutiérrez, & Hervás-Martínez, 2013)	BKS	✓		Range restricted	Kernels			
(Sánchez, Morales, & Gonzalez, 2013)	SOI-CJ	✓		Clustering+Jittering				
(Wang et al., 2013a)	TSMOTE+AB			Range restricted	Bagging	✓		
(Wang, Yao, Zhou, Leng, & Chen, 2013b)	MST-SMOTE			Graph based				
(Zhou, Yang, Guo, & Hu, 2013)	Assembled-SMOTE	✓						
(Menardi & Torelli, 2014)	ROSE	✓	✓	Without-Smoothing	Kernels			
(Barua, Islam, Yao, & Murase, 2014)	MWMOTE	✓		Clustering				
(Gao et al., 2014b)	PDFOS			PDF+Gaussian				
(Koto, 2014)	SMOTE-Out			Range restricted				
(Koto, 2014)	SMOTE-Cosine	✓						
(Koto, 2014)	Selected-SMOTE				FS			
(Li, Zhang, Lu, & Fang, 2014)	SDSMOTE	✓				✓		
(López et al., 2014)	IPADE-ID		✓			✓		✓
(Mahmoudi, Moradi, Ahklaghian, & Moradi, 2014)	DSMOTE	✓						
(Rong et al., 2014)	SSO			Gaussian+Q-union				
(Sandhan & Choi, 2014)	G-SMOTE			Gaussian+Non-linear				
(Xu, Le, & Tian, 2014)	NT-SMOTE			Multiple				
(Zhang & Li, 2014)	RWO-Sampling			Without-Gaussian				
(Lee, Kim, & Lee, 2015)	< *no name* >	✓						
(Almogahed & Kakadiaris, 2015)	NEATER							✓
(Alejo et al., 2015)	MSEBPOS					✓		
(Bellinger et al., 2015)	DEAGO			Without	Auto-Encoder			
(Dang et al., 2015)	SPY						✓	
(Das et al., 2015)	wRACOG	✓		Without-Markov				
(Gazzah, Hechkel, & Amara, 2015)	< *no name* >		✓	Topologies	PCA			
(Jiang, Qiu, & Li, 2015)	MCT			Without-Copy				
(Li, Fong, & Zhuang, 2015)	SMOTE-PSO/BAT							
(Mao, Wang, & Wang, 2015)	MinorityDegree-SMOTE		✓					✓
(Mathew et al., 2015)	K-SMOTE				Kernels			
(Pourhabib, Mallick, & Ding, 2015)	ADG	✓		Without-Gaussian	Kernels			
(Sáez et al., 2015)	SMOTE-IPF							✓
(Tang & He, 2015)	KernelADASYN				Kernels	✓		
(Xie et al., 2015)	MOT2LD	✓		Clustering	t-SNE			
(Young et al., 2015)	V-synth	✓		Voronoi				
(Zieba et al., 2015)	RBM-SMOTE					✓		✓
(Abdi & Hashemi, 2016)	MDO	✓		Ellipse	PCA	✓		
(Bellinger et al., 2016)	DAE			Without	PCA+Auto-Encoder			
(Borowska & Stepaniuk, 2016)	VIS-RST	✓					✓	✓
(Gong & Gu, 2016)	DGSMOTE		✓	Clustering				✓
(Jiang et al., 2016)	GASMOTE					✓		
(Nekooeimehr & Lai-Yuen, 2016)	A-SUWO	✓		Clustering				
(Peng, Zhang, Yang, Chen, & Zhou, 2016)	SMOTE-DGC					✓		✓
(Pérez-Ortiz et al., 2016)	OEFS				Kernels			
(Ramentol et al., 2016)	SMOTE-FRST-2T							✓
(Rivera & Xanthopoulos, 2016)	OUPS	✓						
(Torres, Carrasco-Ochoa, & Martínez Trinidad, 2016)	SMOTE-D			Range restricted		✓		
(Yun, Ha, & Lee, 2016)	AND-SMOTE					✓		
(Cervantes et al., 2017)	SMOTE-PSO	✓(SVs)				✓		✓
(Ma & Fan, 2017)	CURE-SMOTE	✓		Clustering				
(Rivera, 2017)	NRAS					✓		✓
(Cao, Liu, Zhang, Zhao, Huang, & Zaïane, 2017b)	MKOS				FS + Kernels			
(Douzas & Bacao, 2017)	SOMO			Clustering	SOM	✓		
(Li, Fong, Wong, & Chu, 201)	AMSCO	✓	✓			✓		✓

the ratio of examples belonging to the majority class in the neighborhood. Then a density distribution is computed using all the ratios of the minority instances, which will be used to compute the number of synthetic examples required to be generated for each minority example.

- **Safe-Level-SMOTE** (Bunkhumpornpat et al., 2009): It assigns each minority example a safe level before generating synthetic instances. Each synthetic instance will be positiones closer to the largest safe level, thus generating all synthetic instances only in safe regions. The safe level is the ratio between the number of minority examples within the neighborhood and the safe level ratio depends on the safe level of each instance and that of the examples in its neighborhood. The interpolation is controlled by a gap which depends on the safe level ratio of each minority instance.

- **DBSMOTE** (Bunkhumpornpat et al., 2012): This algorithm relies on a density-based approach of clustering called DBSCAN and performs oversampling by generating synthetic samples along a shortest path from each minority instance to a pseudocentroid of a minority-class cluster. DBSMOTE was inspired by Borderline-SMOTE in the sense it operates in an overlapping region, but unlike Borderline-SMOTE, it also tries to maintain both the minority and majority class accuracies.

- **ROSE** (Menardi & Torelli, 2014): ROSE is an oversampling technique proposed within a complete framework to obtain classification rules in imbalanced data. It is established from the generation of new artificial data from the classes, according to a smoothed bootstrap form and the idea behind it is supported by the theoretical well-known properties of the kernel methods. The algorithm samples a new instance using the probability distribution centered at a randomly selected example and depending on a smoothing matrix of scale parameters.

- **MWMOTE** (Barua et al., 2014): Based on the assumption of that existing oversampling methods may generate wrong synthetic minority samples, MWMOTE analyzes the most difficult minority examples and assigns each them a weight according to their distance from the nearest majority examples. The synthetic examples are then generated from the weighted informative minority class instance using a clustering approach, ensuring that they must lie inside a minority class cluster.

- **MDO** (Abdi & Hashemi, 2016): It is one of the recent multi-class approaches inspired by Mahalanobis distance. MDO builds synthetic examples having the same Mahalanobis distance from each examined class mean as the other minority examples. Thus, the region of minority instances can be better learned by preserving the covariance during the generation of synthetic examples along the probability contours. Also, the risk of overlapping between different class regions is reduced.

3.3 SMOTE-Based Extensions for Ensembles

Ensembles of classifiers has emerged as a popular learning framework to address imbalanced classification problems. SMOTE has also involved in and / or extended to many ensemble based methods. Table 2 shows a list of ensemble based techniques that incorporate SMOTE itself or a derivative of SMOTE as a major step to achieve the diversity of the set of classifiers learned to form the ensemble. Note that this table only contains the methods concerned with

Table 2: Enumeration and categorization of SMOTE-based ensemble methods

Ref.	Algorithm name	Type of multi-classifier	Initial selection	Integration US	Type of Interpolation	Adaptive generation	Relabeling	Multi-Class
(Chawla, Lazarevic, Hall, & Bowyer, 2003)	SMOTEBoost	Boosting						
(Guo & Viktor, 2004)	DataBoost-IM	Boosting			Without			
(Frank & Pfahringer, 2006)	inputSmearing	Bagging			Without-Gaussian			
(Mease et al., 2007)	JOUS-Boost	Boosting		✓	Jittering			
(Wang & Yao, 2009)	SMOTEBagging	Bagging						
(Chen, He, & Garcia, 2010b)	RAMOBoost	Boosting				✓		
(Peng & Yao, 2010)	AdaOUBoost	Boosting	✓	✓				
(Hukerikar et al., 2011)	SkewBoost	Boosting			Feature-weighted			
(Blaszczynski et al., 2012)	IIvotes+SPIDER	Bagging	✓	✓	Without-Copy		✓	
(Jeatrakul & Wong, 2012)	OAA-DB	OVA	✓					✓
(Thanathamathee & Lursinsap, 2013)	< *no name* >	Boosting	✓		Boostrap-Resampling			
(Yongqing, Min, Danling, Gang, & Daichuan, 2013)	I-SMOTEBagging	Bagging						
(Abdi & Hashemi, 2016)	MDOBoost	Boosting	✓		MDO(Abdi & Hashemi, 2016)	✓		✓
(Bhagat & Patil, 2015)	SMOTE+OVA	OVA						✓
(Sen, Islam, Murase, & Yao, 2016)	BBO	Boosting+OVA						✓
(Wang, Luo, Huang, Feng, & Liu, 2017)	BEBS	Bagging	✓		Range restricted			
(Gong & Kim, 2017)	RHSBoost	Boosting	✓	✓	Without-Smoothing			

oversampling and generation of synthetic examples; the reader can consult the specialized literature to review other ensembles proposed for imbalanced learning in which SMOTE does not take part in (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012; Fernández, López, Galar, Del Jesus, & Herrera, 2013; Hoens & Chawla, 2010). Specifically, it is important to point out that we may find several studies that show a good behavior for the undersampling-based approaches in synergy with ensemble learning (Khoshgoftaar, Hulse, & Napolitano, 2011; Galar et al., 2012; Blaszczynski & Stefanowski, 2015; Galar, Fernández, Barrenechea, Bustince, & Herrera, 2016).

The structure of the Table 2 is very similar to the previous one, Table 1. The dimensionality change and filtering are two properties not used in ensembles. Furthermore, we add a new column designating the type of ensemble method, namely if the method is a boosting, bagging or One-Versus-All (OVA) approach. The rest of properties are explained in Section 3.1.

3.4 Exhaustive Empirical Studies Involving SMOTE

SMOTE is established as the "de facto" standard or benchmark in learning from imbalanced dataset. Although it would be impossible to survey all the analytic studies that involve SMOTE in any step, in this brief section, we review some of the most influential empirical studies that studied SMOTE in depth. The first type of experimental studies emerged to check whether oversampling is more effective than undersampling and what rate of oversampling or undersampling rate should be used (Estabrooks, Jo, & Japkowicz, 2004). Several studies tackled this issue from a more general point of view (López et al., 2013) and specifically focused on SMOTE to ask about how to discover the proper amount and type of sampling (Chawla et al., 2008). In the work of Batista et al. (2004), some common resampling approaches are compared and the hybridizations of SMOTE with undersampling was shown to outperform the rest of resampling techniques. Later, Prati, Batista, and Silva (2015) designed a renewed experimental setup to answer some open-ended questions on the relationship and performance between learning paradigms, imbalance degrees and proposed solutions. More complex analytic studies can be found to analyze data intrinsic characteristics (López et al., 2013), data difficulty factors such as rare sub-concepts of minority instances, overlapping of classes (Luengo, Fernández, García, & Herrera, 2011; Stefanowski, 2016) and different types of minority class examples (Napierala & Stefanowski, 2016).

Table 3: List of SMOTE-based approaches for other learning paradigms

Reference	Algorithm Name	Learning Paradigm
(Ditzler et al., 2010)	Learn++.SMOTE	Data Streams
(Cao, Li, Woon, & Ng, 2011)	SPO	Time Series
(Palacios, Sánchez, & Couso, 2012)	SMOTE-LQD	Low Quality Data
(Piras & Giacinto, 2012)	< *no name* >	Image Retrieval
(Blagus & Lusa, 2013)	< *no name* >	High Dimensional Data
(Cao, Li, Woon, & Ng, 2013)	INOS	Time Series
(Ditzler & Polikar, 2013)	Learn++.NSE-SMOTE	Data Streams
(Ertekin, 2013)	VIRTUAL	Active Learning
(Iglesias, Vieira, & Borrajo, 2013)	COS-HMM	Text Classification
(Li, Yu, Yang, Xia, Li, & Kaveh-Yazdy, 2013a)	INNO	Semi-Supervised Learning
(Wang, Liu, Japkowicz, & Matwin, 2013)	Instance-SMOTE, Bag-SMOTE	Multi-Instance Learning
(Mera, Orozco-Alzate, & Branch, 2014)	< *no name* >	Multi-Instance Learning
(Moutafis & Kakadiaris, 2014)	GS4	Semi-Supervised Learning
(Park, Qi, Chari, & Molloy, 2014)	< *no name* >	Semi-Supervised Learning
(Barua, Islam, & Murase, 2015)	GOS-IL	Data Streams
(Charte, Rivera, del Jesus, & Herrera, 2015)	MLSMOTE	Multi-Label Learning
(Mera, Arrieta, Orozco-Alzate, & Branch, 2015)	Informative-Bag-SMOTE	Multi-Instance Learning
(Pérez-Ortiz, Gutiérrez, Hervás-Martínez, & Yao, 2015)	OGO-NI, OGO-ISP, OGO-SP	Ordinal Regression
(Torgo, Branco, Ribeiro, & Pfahringer, 2015)	SMOTER	Regression
(Triguero, García, & Herrera, 2015)	SEG-SSC	Semi-Supervised Learning
(Dong, Chung, & Wang, 2016)	OCHS-SSC	Semi-Supervised Learning
(Moniz, Branco, & Torgo, 2016)	SM_B, SM_T, SM_TPhi	Time Series

Another issue studied particularly in SMOTE is the relationship between data preprocessing and cost-sensitive learning. In the review by Lopez, Fernandez, Moreno-Torres, and Herrera (2012), an exhaustive empirical study was performed to this goal, concluding that both preprocessing and cost-sensitive learning are good and equivalent approaches to address the imbalance problem.

Regarding different typologies of algorithms, SMOTE has been deeply analyzed in combination with cost-sensitive neural networks (Zhou & Liu, 2006), SVMs (Tang, Zhang, Chawla, & Krasser, 2009), linguistic fuzzy rule based classification systems (Fernandez, Garcia, del Jesus, & Herrera, 2008) and genetics-based machine learning for rule induction (Fernandez, Garcia, Luengo, Bernado-Mansilla, & Herrera, 2010).

4. Variations of SMOTE to Other Learning Paradigms

In this section, we will introduce the SMOTE-based approaches that address other learning paradigms. In particular, the section will be divided into five subsections, each one providing an overview of each paradigm and the techniques devised to tackle it. Extensions of SMOTE have been applied other learning paradigms: (1) streaming data (see Section 4.1); (2) Semi-supervised and active learning (in Section 4.2); (3) Multi-instance and multi-label classification (Section 4.3); (4) Regression (in Section 4.4) and (5) Other and more complex prediction problems and such as text classification, low quality data classification, and so on (see Section 4.5).

Table 3 presents a summary of the SMOTE extensions by chronological order, indicating their references, algorithm names and learning paradigms they tackle. In the following, we will give a brief description of each learning paradigm and the associated developed techniques.

4.1 Streaming Data

Many applications for learning algorithms need to tackle dynamic environments where data arrive in a streaming fashion. The online nature of data creates some additional computational requirements for a classifier (Krawczyk et al., 2017). In addition, the prediction models are usually required to adapt to the concept drifts, which are phenomena derived from the non-stationary characteristics of data streams. In the offline version of imbalance classification, the classifier can estimate the relationship between the minority class and majority class before learning begins. Nevertheless, in online learning, it is not possible to do this due to the fact that classes can change their distribution over time, thus they have to cope with the dynamic of the data.

Two preprocessing techniques (Ramírez-Gallego et al., 2017) based on SMOTE have been proposed to deal with imbalanced data streams. The first is Learn++.NSE-SMOTE (Ditzler & Polikar, 2013), which is an extension of Learn++.SMOTE (Ditzler et al., 2010). First, the authors incorporated SMOTE within the algorithm Learn++.NSE and after they decided to replace SMOTE with a subensemble that makes strategic use of minority class data. The second technique is GOS-IL (Barua et al., 2015). It works by updating a base learner incrementally using standard Oversampling.

When a data stream is received over time and we have disposal of time information, we refer to time series classification. A time series data sample is an ordered set of real-valued variables coming from a continuous signal, which can be either in time or spatial domain. The variables close to each other are often highly correlated in time series. The methods SPO (Cao et al., 2011) and INOS (Cao et al., 2013) propose an integration of SMOTE in time series classification. INOS can be viewed as an extension of SPO and addresses the imbalanced learning issue by oversampling the minority class in the signal space. An hybrid technique was used to generate synthetic examples by means of estimating and maintaining the main covariance structure in the reliable eigen subspace and fixing the unreliable eigen spectrum.

A third family of techniques called SM_B, SM_T and SM_TPhi (Moniz et al., 2016) were also devised for time series, but for regression. Details for them will be given in Section 4.4.

4.2 Semi-supervised and Active Learning

An important limitation of supervised learning is the great effort to obtain enough labeled data to train predictive models. In a perfect situation, we want to train classifiers using diverse labeled data with a good representation of all classes. However, in many real applications, there is a huge amount of unlabeled data and the obtaining of a representative subset is a complex process. Active learning produces training data incrementally by identifying the most informative data to label. When external supervision is involved (humans or other system), we are referring to real active learning in which the new examples are selected and then labeled by the expert. If this is not the case, we refer to semi-supervised classification, which utilizes unlabeled data to improve the predictive performance, modifying the learned hypothesis obtained from labeled examples. Different perspectives are employed to tackle semi-supervised classification, such as self-training, graph-based approaches, generative models, and so on (Zhu, Goldberg, Brachman, & Dietterich, 2009).

Several methods based on SMOTE have been developed for this learning paradigm:

- VIRTUAL (Ertekin, 2013) is designed for active learning problems and SVMs and it adaptively creates instances from the real positive support vectors selected in each active learning step.

- INNO (Li et al., 2013a) is a technique for graph-based semi-supervised learning and performs an iterative search to generate a few unlabeled samples around known labeled samples.

- GS4 (Moutafis & Kakadiaris, 2014), SEG-SSC (Triguero et al., 2015) and OCHS-SSC (Dong et al., 2016) generate synthetic examples to diminish the drawbacks produced by the absence of labeled examples. Several learning techniques were checked and some properties such as the common hidden space between labeled samples and the synthetic sample were exploited.

- The technique proposed by Park et al. (2014) is a semi-supervised active learning method in which labels are incrementally obtained and applied using a clustering algorithm.

4.3 Multi-class, Multi-instance and Multi-label Classification

Although the original SMOTE technique can be applied to multi-class problems by identifying the minority class against the remaining ones (One-versus-all approach), there are some extensions specifically employed for tackling multi-class imbalanced classification problems (Wang & Yao, 2012): the work from Fernández-Navarro et al. (2011), Alejo et al. (2015) and Abdi and Hashemi (2016).

In multi-instance learning, the structure of the data is more complex than in single-instance learning (Dietterich, Lathrop, & Lozano-Pérez, 1997; Herrera et al., 2016b). Here, a learning sample is called a bag. The main feature in this paradigm is that a bag is associated with multiple instances or descriptions. Each instance is described by a feature vector, like in single-instance learning, but associated output is unknown. An instance, apart from its feature values, only knows its membership relationship to a bag.

Several ideas based on SMOTE have been proposed to tackle multi-instance learning. The first ones were Instance-SMOTE and Bag-SMOTE (Wang et al., 2013). The Instance-SMOTE algorithm creates synthetic minority instances in each bag, without creating new bags. Besides, Bag-SMOTE creates new synthetic minority bags with new instances. In the work of Mera, Orozco-Alzate, and Branch (2014) first, and next in Mera, Arrieta, Orozco-Alzate, and Branch (2015), the Informative-Bag-SMOTE technique was presented and improved. It uses a model of the negative population to find the best instances in the minority class to be oversampled. The new synthetic bags created support the target concept in the minority class.

In multilabel classification (Herrera et al., 2016a) each instance of the data has associated a vector of outputs, instead of only one value. This vector has a fixed size according to the number of different labels in the dataset. The vector is composed by binary values based elements which indicate whether or not the corresponding label is compatible to the instance. Of course, several labels can be active at once, showing different combinations of labels, which is known as labelset.

MLSMOTE (Charte et al., 2015) is the most popular extension of SMOTE designed for multilabel classification. Its objective is to produce synthetic instances related to minority labels. The subset of minority labels within the labelset is identified by two proposed measures. Input features of the synthetic examples are obtained using SMOTE, but the labelsets of these new instances are also gathered from the nearest neighbors, taking advantage of label correlation information in the neighborhood.

4.4 Regression

Regression tasks consider the output variable as continuous and hence, the values are represented by real numbers. Unlike standard classification, they are ordered. The imbalance learning correspondence for regression tasks is the correct prediction of rare extreme values of a continuous target variable. In the work of Torgo et al. (2015), several techniques for resampling were successfully applied for regression. Among them, SMOTER is the SMOTE-based contribution of Oversampling regression. SMOTER employs a user-defined threshold to define the rare cases as extreme high and low values, dealing both types as separate cases. Another major difference is the way the target value for the new cases is generated, in which a weighted average between two seed cases is used. SMOTER has been extended to tackle time series forecasting in the study of Moniz et al. (2016). Here, three methods are derived from SMOTER: SM_B, SM_T, SM_TPhi. They take into account the characteristics of the bins of the time series and manage the temporal and relevance bias.

The ordinal regression (or classification) problem is half way between the standard classification and regression. There exists a predefined order among the categories of the output variable, but the distance between two consecutive categories is unknown. Thus, the penalization of misclassification errors can be greater or lower depending on the difference between the real category and the predicted category. An imbalance scenario of classes may be usual in this kind of domains when addressing real applications. In the research conducted by Pérez-Ortiz et al. (2015), an approach of Oversampling from a graph-based perspective is used to balance the ordinal information. Three schemes of generation were proposed, namely OGO-NI, OGO-ISP, OGO-SP; depending on the use of intra-class edges, shortest paths and interior shortest paths of the graph constructed.

4.5 Other and More Complex Prediction Problems

Other problems in which a variant of SMOTE has been applied are the following:

- Imbalanced classification with imprecise datasets. This problem refers to the presence of vagueness in the data, preventing the values of the classes to be precisely known. SMOTE-LQD (Palacios et al., 2012) is a generalized version of SMOTE for this environment. It delivers the selection of minority instances assuming that the imbalance ratio is not precisely known and the computation of the nearest neighbors and generation of synthetic instances is carried out with fuzzy arithmetic operators.

- Image retrieval and semantic search of images is a challenging problem nowadays. In the work of Piras and Giacinto (2012), the authors proposed a technique that address the imbalance problem in image retrieval tasks by generating synthetic patterns according to nearest neighbor information.

- In bioinformatics problems, it is usual to have high-dimensional classification problems. In the work of Blagus and Lusa (2013), SMOTE was tested in such scenarios in both theoretical and empirical perspectives. Among the conclusions achieved, the most important was that SMOTE has hardly any effect on most classifiers trained on high-dimensional data. Other techniques such as Undersampling may be preferable on high-simensional settings.

- A variation of SMOTE based on document content to manage the class imbalance problem in text classification was proposed by Iglesias et al. (2013). The method called COS-HMM incorporates an Hidden Markov Model that is trained with a corpus in order to create new samples according to current documents.

5. Challenges in SMOTE-Based Algorithms

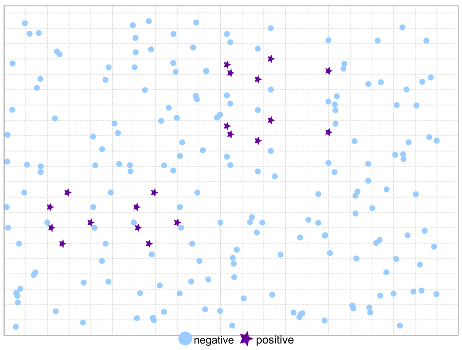
When working in the scenario of imbalanced classification, we must be aware that the skewed class distribution is not the only drawback for the performance degradation. Instead, its conjunction with several data intrinsic characteristic is the cause for the achievement of sub-optimal models (López et al., 2013). For example, if the two classes, despite having severely imbalanced data distribution are easily separable in two clusters or segments, then it becomes easy for any classifier to learn to discriminate between them. It is when the classes are interspersed when the challenges become profound, as is often the case with the real-world applications.

Throughout this section, we will discuss in detail several of these issues and their relationship with SMOTE. Particularly, we will first study the problems related to those areas where minority class are represented as small disjuncts (Orriols-Puig, Bernadó-Mansilla, Goldberg, Sastry, & Lanzi, 2009; Weiss & Provost, 2003), and their relationship with the lack of data (Raudys & Jain, 1991) and noisy instances (Seiffert, Khoshgoftaar, Hulse, & Folleco, 2014) (Section 5.1). Next, we will consider an issue that hinders the performance in imbalanced classification, i.e. overlapping or class separability (García et al., 2008) (Section 5.2). In addition, since SMOTE applies an interpolation procedure to generate new synthetic data on the feature space, we will analyze the curse of dimensionality (Blagus & Lusa, 2013) as well as different aspects for the interpolation process (Section 5.4). We must also take into account that a different data distribution between the training and test partitions, i.e. the dataset shift (Moreno-Torres, Sáez, & Herrera, 2012b), can also alter the validation of the results in these cases (Section 5.3).

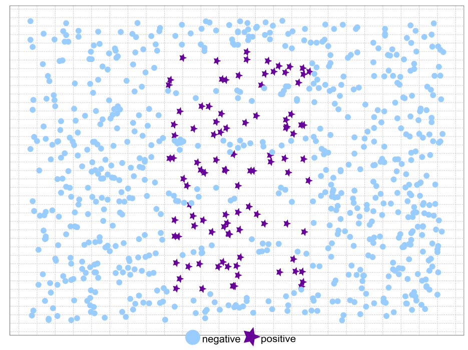
Finally, we will consider two significant novel scenarios for addressing imbalanced classification. On the one hand, we focus on real time processing, and more specifically data streams imbalanced classification (Nguyen, Cooper, & Kamei, 2011; Wang, Minku, & Yao, 2013) (Section 5.5). Next, we analyze the topic of Big Data (Fernández, Río, López, Bawakid, del Jesus, Benítez, & Herrera, 2014) and the constraints associated with the skewed class distribution (Río, López, Benítez, & Herrera, 2014) (Section 5.6).

5.1 Small Disjuncts, Noise and Lack of Data

We refer to a dataset containing small disjuncts when some concepts (disregard their class) are represented within small clusters (Orriols-Puig et al., 2009; Weiss & Provost, 2003). In the case of imbalanced classes, this problem occurs very often as underrepresented concepts are usually located in small areas of the dataset. This situation is represented in Figure 3, where we show two cases. First, Figure 3a depicts an artificially generated dataset with small disjuncts for the minority class. Then, Figure 3b shows the *"Subclus"* problem created in the work of Napierala, Stefanowski, and Wilk (2010), where we can find small disjuncts for both classes: the majority class samples are underrepresented with respect to the minority class samples in the central region of minority class rectangular areas, whereas the minority samples only cover a small part of the whole dataset and are placed inside the negative class.



(a) Artificial dataset: small disjuncts for the minority classt



(b) Subclus dataset: small disjuncts for both classes

Figure 3: Example of small disjuncts on imbalanced data

This situation increases the complexity in the search for quality solutions. This is due to the common working procedure of standard learning models which aim at achieving a good generalization ability. As such, most classification algorithms may consider these examples to be in the category of class-noise (Kubat & Matwin, 1997; Jo & Japkowicz, 2004), just because they are located in the "safe-area" of the contrary class. Taking into account that classification algorithms are more sensitive to noise than imbalance (Seiffert et al., 2014), different overfitting management techniques are often used to cope with this problem, i.e. pruning for decision trees. However, and as stated previously, this may cause to ignore correct clusters of minority class examples.

The problem of small disjuncts affects to a higher degree, those learning algorithms whose procedure is based on a divide-and-conquer strategy. Since the original problem is divided into different subsets, in several iterations this can lead to data fragmentation (Friedman, 1996). Some clear examples of this behavior are decision trees (Rokach, 2016), and the well-known MapReduce programming model that is used for Big Data applications (Dean & Ghemawat, 2008; Fernández et al., 2014).

The small sample size (lack of data) (Raudys & Jain, 1991) and small disjuncts are two closely related topics. This synergy is straightforward as information is barely represented in those small disjuncts. Therefore, learning classifiers cannot carry out a good generalization when there is not enough data to represent the boundaries of the problem (Jo & Japkowicz, 2004; Wasikowski & Chen, 2010). This way, small disjuncts, noisy data and lack of data are three inter-related problems that comprise a challenge to the research community in imbalanced classification.

Simpler oversampling approaches based on instance replication do not cope well with such data intrinsic problems. On the contrary, SMOTE based algorithms implicitly consider a mechanism to counteract both the class imbalance and the small disjuncts. By means of creating new instances in between close examples, it allows to reinforce the representation within the clusters. The premise for the good behavior of SMOTE is related to the fact that the nearest examples should be selected within that very area. Of course than depends on the number of elements composing the small disjuncts and the value of K selected for the oversampling. In addition, if the cluster with the small disjunct also contains any example from the contrary class, i.e. overlapping, SMOTE will not be able to correct this issue of the within-class imbalance. This is the main reason for using SMOTE hybridizations with cleaning techniques.

Fortunately, and as introduced in Section 3.2, there are several SMOTE extensions that try to analyze these clusters of data. This way, cluster-based approaches based on local densities in conjunction with SMOTE are of high interest for a two-fold reason. On the one hand, they focus on those areas that truly need the instance generation, i.e. those with lack of representation. On the other hand, they avoid the overgeneralization problem increasing the density of examples on the cores of the minority class, and making them sparse far from the centroid. Finally, recent works suggest that changing the representation of the problem, i.e. taking into account the pairwise differences among the data (Pekalska & Duin, 2005) may somehow overcome the issue of small disjuncts (García, Sánchez, de J. Ochoa Domínguez, & Cleofas-Sánchez, 2015). However, we must point out that the problem of finding such class areas is still far from being properly addressed, as most of the clustering techniques previously described make several simplified assumptions to address real complex distribution problems.

Another approach is to apply a synergy of preprocessing models, i.e. filtering and/or instance generation to remove those instances that are actually noisy prior to the SMOTE application (Sáez et al., 2015; Verbiest, Ramentol, Cornelis, & Herrera, 2014). Some studies shown that simple undersampling techniques such as random undersampling and cleaning techniques are known to be robust for different levels of noise and imbalance (Seiffert et al., 2014). This way, many hybrid approaches between filtering techniques and SMOTE have been developed so far, since this allow to improve the quality of the data either a priori (from the original data), a posteriori (from the preprocessed data) or iteratively while creating new synthetic instances.

The cooperation between boosting algorithms and SMOTE can successfully address the problem of the small disjuncts. These learning algorithms are iterative and they apply different weights to the data instances dynamically as the procedure evolves (Schapire, 1999). Specifically, incorrectly classified instances have their weights increased, so that in further steps the generated model will be focused on them. Because instances in the small disjuncts are known to be difficult to predict, it is reasonable to believe that boosting will improve their classification performance. Following this idea, many approaches have been developed modifying the standard boosting weight-update mechanism to improve the performance of the minority class and the small disjuncts (Galar, Fernández, Barrenechea, Bustince, & Herrera, 2011), and those involving a derivative of SMOTE were mentioned in Section 3.3. However, we must take into account that in case that several data intrinsic characteristics (overlapping, small disjuncts, noise, among others) converge in the same problem, even ensemble learning algorithm will find quite difficult to carry out a proper class discrimination.

5.2 Overlapping or Class Separability

Among all data intrinsic characteristics, the overlapping between classes is possibly the most harmful issue (García et al., 2008). It is defined as those regions of the data space in which the representation of the classes is similar. This situation leads to develop an inference with almost the same a priori probabilities in this overlapping area, which makes very hard or even impossible the distinction between the two classes. Indeed, any "linearly separable" problem can be solved by a naïve classifier, regardless of the class distribution (Prati & Batista, 2004).

The common occurrence of overlapping and class imbalance implies a harder restriction for the learning models. This issue was pointed out in the research conducted by Luengo et al. (2011), in which authors depicted the performance of several datasets ordered with respect to different data complexity measures in order to search for some regions of interesting good or bad behavior. The findings in this work show that the metrics which measure the overlap between the classes can better characterize the degree of final precision obtained, in contrast to the imbalance ratio.

The widest use metric to compute the degree of overlap for a given dataset is known as *maximum Fisher's discriminant ratio*, or simply $F1$ (Ho & Basu, 2002) (it must not be confused with the F1-score performance metric). It is obtained for every individual feature (one dimension) as:

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

being μ_1, μ_2, σ_1^2, σ_2^2 the means and variances of the two classes respectively. Finally, $F1$ is obtained as the maximum value for all features.

Datasets with a small value for the F1 metric will have a high degree of overlapping. Figures 4 to 7 show an illustrative example of this behavior, which have been built with synthetic data, using two variables within the range $[0.0; 1.0]$ and two classes.

The overlapping areas are directly related to the concept of "borderline examples" (Napierala et al., 2010). As its name suggests, these are defined as those instances that are located in the area surrounding class boundaries, where the minority and majority classes overlap. The main issue is again trying to determine whether these examples are simply noise or they represent useful information. Thus, it is of special importance being able to identify among different types of instances for a given problem, i.e. linearly separable, borderline, and overlapping data (Vorraboot, Rasmequan, Chinnasarn, & Lursinsap, 2015). This way, we will be able to discard "misleading" instances and to focus on those areas that are hard to discriminate, carrying out an informed oversampling process. Therefore, a similar procedure to that used in small disjuncts can be followed in this case, i.e. combining filtering techniques, clustering, and analyzing the neighborhood of each instance to determine their actual contribution to the problem.

Additionally, feature selection or feature weighting can be combined with SMOTE preprocessing (Martín-Félez & Mollineda, 2010; ?). In this sense, SMOTE preprocessing will deal with class distribution and small disjuncts ("IR part") and feature preprocessing somehow reduces the degree of overlapping ("F1 part"). A recent approach proposed a synergy between SMOTE and both feature and instance selection (Fernandez, Carmona, del Jesus, & Herrera, 2017). The basis of this novel methodology is similar to the previous ones, but

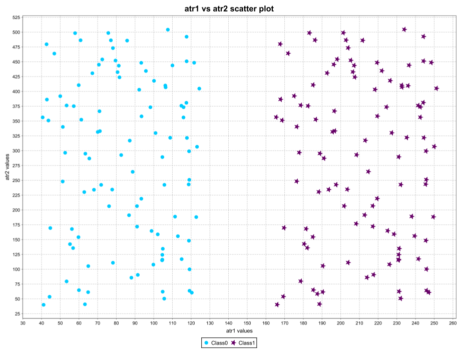


Figure 4: $F1 = 12.5683$

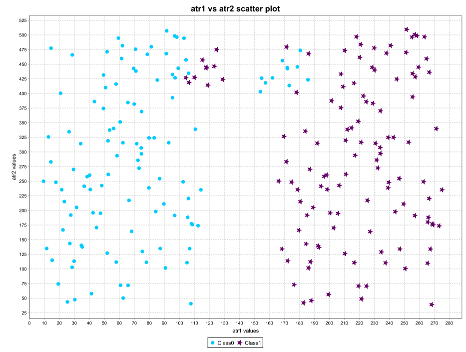


Figure 5: $F1 = 5.7263$

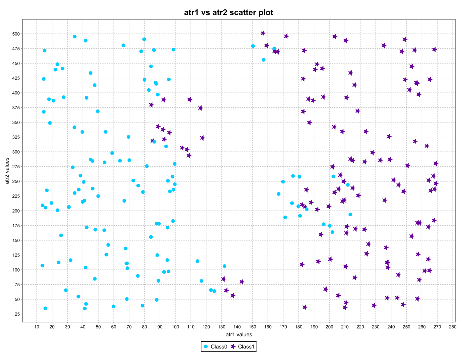


Figure 6: $F1 = 3.3443$

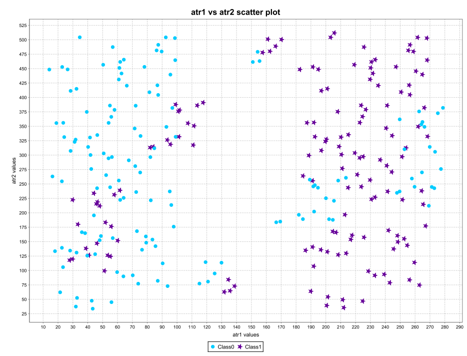


Figure 7: $F1 = 0.6094$

instead of learning a single solution, it provides a Multi-Objective Evolutionary Algorithm (Zhou, Qu, Li, Zhao, Suganthan, & Zhangd, 2011) to achieve a diverse set of classifiers under different training sets, i.e. considering different features and instances. The key is to specialize several classifiers in different areas of the problem, leading to a robust ensemble scheme.

5.3 Dataset Shift

The problem of dataset shift (Moreno-Torres, Raeder, Alaiz-Rodriguez, Chawla, & Herrera, 2012a) is defined as the case where training and test data follow different distributions. There are three potential types of dataset shift:

1. *Prior Probability Shift*: when the class distribution is different between the training and test sets (Storkey, 2009). This case can be directly addressed by applying a stratified cross validation scheme so that the same number of instances per class are represented in both sets.

2. *Covariate Shift*: when the input attribute values that have different distributions between the training and test sets (Shimodaira, 2000). The incidence of this issue mainly depends on the partitioning of the data for validation purposes. The widest used procedure for this task, the stratified k-fold cross validation may lead to this type

of induced dataset shift, as the instances are randomly shuffled among the different folds.

3. *Concept Shift*: when the relationship between the input and class variables changes (Yamazaki, Kawanabe, Watanabe, Sugiyama, & Müller, 2007). This represents the hardest challenge among the different types of dataset shift. In the specialized literature it is usually referred to as "Concept Drift" (Webb, Hyde, Cao, Nguyen, & Petitjean, 2016).

As described above, dataset shift comprises a general and common problem that can affect all kind of classification problems. In other words, it is not a condition intrinsically related to data streams or real time processing. Particularly, in imbalanced domains this issue can be especially sensitive due to the low number of examples for the minority class (Moreno-Torres & Herrera, 2010). In the most extreme cases, a single misclassified example of the minority class can create a significant drop in performance.

In the case of covariate shift, it is necessary to combine the SMOTE oversampling technique with a suitable validation technique. In particular, we may find in the work of Moreno-Torres et al. (2012b) a novel approach that is not biased to this problem. Named as DOB-SCV, this partitioning strategy aims at assigning close-by examples to different folds, so that each fold will end up with enough representatives of every region. Lopez, Fernandez, and Herrera (2014) considered the use of the former procedure in the scenario of imbalanced classification and they found to be an stable performance estimator. Avoiding different data distribution inside each fold will allow researchers on imbalanced data to concentrate their efforts on designing new learning models based only on the skewed data, rather than seeking for complex solutions when trying to overcome the gaps between training and test results.

Finally, regarding concept shift more sophisticated solutions must be applied. As we mentioned in Section 4.1, Ditzler and Polikar (2013) integrated the SMOTE preprocessing within a novel ensemble boosting approach that applies distribution weights among the instances depending on their distribution at each time step.

5.4 Curse of Dimensionality and Interpolation Mechanisms

Classification problems with a large number of attributes imply a significant handicap for the correct development of the final models. First, because most of the learning approaches take into account the whole feature space to build the system, it is harder to find a real optimal solution. Second, because of the overlap between classes for some of these attributes, which can cause overfitting, as pointed out previously.

In addition to the former, we must take into account that the dimensionality problem also gives rise to the phenomenon of hubness (Radovanovic, Nanopoulos, & Ivanovic, 2010), defined as a small number of points that become most of the observed nearest neighbors. In the case of the SMOTE procedure, this affects the quality of the new synthetic examples for two inter-related reasons (Blagus & Lusa, 2013). On the one hand, the computation of the neighborhood becomes skewed to the actual one. On the other hand, the variance for the new created instances becomes higher.

One way to overcome this problem can be to predict and rectify the detrimental hub point occurrences, for example using methods based on naive bayes to avoid borderline examples and outliers (Tomasev & Mladenic, 2013). Another simpler solution is to benefit from the use of a feature selection approach prior to the application of the SMOTE oversampling, as suggested in several works (Lin & Chen, 2013; Yin & Gai, 2015). Some studies also show that k-NN classifiers obtain a higher benefit from this synergy (Blagus & Lusa, 2013). However, we may find other works in which authors follow the contrary procedure, i.e. they first rebalance the data and then apply the feature selection scheme (Gao, Khosgoftaar, & Wald, 2014a; Lachheta & Bawa, 2016), also achieving very good results.

The use of different interpolation mechanisms can provide some interesting insight to this problem. Additionally, there is a need to add more variability to the new synthetic instances, and this could be achieved by means of a partial extrapolation. Therefore, the generalization will be positively biased, leading to a better coverage of the "possibly-sparse" minority examples.

Another interesting perspective to obtain more relevant synthetic instances is to analyze different distance measures to obtain the nearest neighbors. One example is the Malanahobis distance that creates an elliptic area of influence that could be better suited in case of overlapping (Abdi & Hashemi, 2016). The Hellinger distance metric, being based on probability distributions and strongly skew insensitive, have been also applied in the context of imbalanced learning, although rather focused on feature selection (Yin, Ge, Xiao, Wang, & Quan, 2013). Finally, we must consider the case of mixed attributes in which metrics such as HOEM or HVDM are mandatory in order to find neighbor instances (Wilson & Martinez, 1997).

Finally, feature extraction to transform the problem into a lower dimensional space is another way to address this issue. When this process is carried out before the application of SMOTE, the new clusters of this transformed dataset may allow a better generation of instances (Xie et al., 2015). It also can be applied after the dataset is rebalanced (Hamid, Sugumaran, & Journaux, 2016). In this latter case, the feature extraction is suggested for a better learning process of the classifier.

5.5 Real-Time Processing

As it has been reported in this manuscript, the problem of imbalanced classification has been commonly focused on stationary datasets. However, there is large number of applications in which data arrive continuously and where queries must be answered in real time. We are referring to the topic of online learning of classifiers from data streams (Last, 2002). In this scenario, the uneven distribution of examples occurs in many case studies, such as video surveillance (Radtke, Granger, Sabourin, & Gorodnichy, 2014), or fault detection (Wang, Minku, & Yao, 2013b). The hitch related to this issue is that it demands a mechanism to intensify the underrepresented class concepts to provide a high overall performance (Wang et al., 2013).

In addition to the former, the dynamical structure of the problem itself also implies the management of unstable class concepts, i.e. concept drifts (Wang, Minku, Ghezzi, Caltabiano, Tiño, & Yao, 2013a). To this end, several methods have been proposed to deal with both obstacles from the point of view of preprocessing (Nguyen et al., 2011; He & Chen, 2011; Wang, Minku, & Yao, 2015), particularly using SMOTE (Ditzler & Polikar, 2013), and/or cost-sensitive learning via ensembles of classifiers (Mirza, Lin, & Liu, 2015; Ghazikhani, Monsefi, & Sadoghi Yazdi, 2013; Pan, Wu, Zhu, & Zhang, 2015).

The adaptation of SMOTE to this framework is not straightforward. The windowing process implies that only a subset of the total data is feed to the preprocessing algorithm, limiting the quality of the generated data. But if we could even store a history of the data, the issue of concept drift, both from the point of view of data and class distribution, diminishes the optimal performance that could be achieved. Therefore, the correlation between the generated synthetic instances along time, and the new incoming minority class instances should be computed. In case of finding a high variance, an update process must be carried out.

5.6 Imbalanced Classification in Big Data Problems

The significance of the topic of Big Data is related to the large advantage from knowledge extraction for these types of problems with huge Volume, high Velocity, and large in Variety (Fernández et al., 2014; Zikopoulos, Eaton, deRoos, Deutsch, & Lapis, 2011).

This implies the need for a novel framework that allows the scalability of the traditional learning approaches. This framework is MapReduce (Dean & Ghemawat, 2008) and its open source implementation (Hadoop-MapReduce). This new execution paradigm carries out a "divide-and-conquer" distributed procedure in a fault-tolerant way to adapt for commodity hardware. To allow computational algorithms to be embedded into this framework, programmers must implement two simple functions, namely Map and Reduce. In general terms, Map tasks are devoted to work with a subset of the original data and to produce partial results. Reduce tasks take as input the output from the Maps (all of which must share the same "key" information) and carry out a fusion or aggregation process.

At present, few research has been developed on the topic of imbalanced classification for Big Data problems (Fernandez et al., 2017). Among all research studies, we must first emphasize the one carried out by Río et al. (2014) in which the first SMOTE adaptation to Big Data was adapted to the MapReduce work-flow. Particularly, each Map task was responsible for the data generation for its chunk of data, whereas a unique Reduce stage joined the outputs from the former to provide a single balanced dataset. We may also find a couple of SMOTE extensions to MapReduce, the first one based on Neighborhood Rough Set Theory (Hu & Li, 2013; Hu, Li, Lou, & Dai, 2014), and the latter on ensemble learning and data resampling (Zhai, Zhang, & Wang, 2017). However, none of these works are actual Big Data solutions as their scalability is limited. Finally, a recent approach based on the use of Graphics Processing Units (GPUs) for the parallel computation of SMOTE has been proposed by Gutierrez, Lastra, Benitez, and Herrera (2017). The preprocessing technique is adapted to commodity hardware by means of a smart use of the main memory, i.e. by including only the minority class instances, and the neighborhood computation via a fast GPU implementation of the kNN algorithm (Gutierrez, Lastra, Bacardit, Benitez, & Herrera, 2016).

One of the reasons of such few works on the topic is probably due to the technical difficulties associated to the adaptation of standard solutions to the MapReduce programming style. Regarding this issue, the main point is to focus on the development and adoption of global and exact parallel techniques in MapReduce (Ramírez-Gallego, Fernández, García, Chen, & Herrera, 2018). Focusing on SMOTE, the problem is mainly related to the use of a fast and exact kNN approach, considering that all minority class instances should be considered for the task.

In addition, the use of streaming processors with GPUs is not a straightforward solution. The technical capabilities of the programmer, in conjunction with the restrictions of memory and data structures of the GPU implementation, imply a significant challenge. Finally, we must also take into account the availability of proper hardware equipment for an experimental study with such Big Data.

We must also point out that the data repartition applied to overcome the scalability problem implies additional sources of complexity. We must keep in mind the lack of data and the small disjuncts (Jo & Japkowicz, 2004; Wasikowski & Chen, 2010), which may become more severe in this scenario. As we have already pointed out, these problems have a strong influence for the behavior of the SMOTE algorithm. This has been stressed as one possible issue for the low performance of SMOTE in comparison with simpler techniques such as random oversampling and random undersampling in Big Data problems (Fernandez et al., 2017). This fact implies the necessity of carrying out a thorough design of the data generation procedure to improve the quality of the new synthetic instances. Additionally, it is recommended to study different possibilities related to the fusion of models or the management of an ensemble system with respect to the final Reduce task.

6. Conclusion

This paper presented a state-of-the-art of SMOTE algorithm in its 15th year anniversary, celebrating the abundant research and developments. It provided a summative analysis of the variations of SMOTE devised with respect to both the improvements on different drawbacks detected on the original idea and its potential application to more complex prediction problems such as streaming data, semi-supervised learning, multi-instance and multi-label learning and regression. In the context of current challenges outlined, we highlighted the need for enhancing the treatment of small disjuncts, noise, lack of data, overlapping, dataset shift and the curse of dimensionality. To do so, the theoretical properties of SMOTE regarding these data characteristics, and its relationship with the new synthetic instances, must be further analyzed in depth. Finally, we also posited that it is important to focus on data sampling and pre-processing approaches (such as SMOTE and its extension) within the framework of Big Data and real-time processing.

Developments and applications to new fields of more refined data preprocessing approaches which follow a SMOTE-inspired similar oversampling strategy based on the artificial generation of data is still a demanding issue for the next 15 years. To inspire this purpose, we wanted provide a valuable overview to this respect for both, beginners and researchers working in any perspective of data mining, and especially in imbalance learning scenarios.

Acknowledgments

This work have been partially supported by the Spanish Ministry of Science and Technology under projects TIN2014-57251-P, TIN2015-68454-R and TIN2017-89517-P; the Project

BigDaP-TOOLS - Ayudas Fundación BBVA a Equipos de Investigación Científica 2016; and the National Science Foundation (NSF) Grant IIS-1447795.

References

Abdi, L., & Hashemi, S. (2016). To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE Transactions on Knowledge and Data Engineering*, *28*(1), 238–251.

Alejo, R., García, V., & Pacheco-Sánchez, J. H. (2015). An efficient over-sampling approach based on mean square error back-propagation for dealing with the multi-class imbalance problem. *Neural Processing Letters*, *42*(3), 603–617.

Almogahed, B. A., & Kakadiaris, I. A. (2015). NEATER: filtering of over-sampled data using non-cooperative game theory. *Soft Computing*, *19*(11), 3301–3322.

Anand, R., Mehrotra, K. G., Mohan, C. K., & Ranka, S. (1993). An improved algorithm for neural network classification of imbalanced training sets. *IEEE Transactions on Neural Networks*, *4*(6), 962–969.

Bach, M., Werner, A., Zywiec, J., & Pluskiewicz, W. (2017). The study of under- and over-sampling methods utility in analysis of highly imbalanced data on osteoporosis. *Information Sciences*, *384*, 174–190.

Barua, S., Islam, M. M., & Murase, K. (2011). A novel synthetic minority oversampling technique for imbalanced data set learning. In *Neural Information Processing - 18th International Conference (ICONIP)*, pp. 735–744.

Barua, S., Islam, M. M., & Murase, K. (2013). ProWSyn: Proximity weighted synthetic oversampling technique for imbalanced data set learning. In *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference (PAKDD)*, pp. 317–328.

Barua, S., Islam, M. M., & Murase, K. (2015). GOS-IL: A generalized over-sampling based online imbalanced learning framework. In *Neural Information Processing - 22nd International Conference (ICONIP)*, pp. 680–687.

Barua, S., Islam, M. M., Yao, X., & Murase, K. (2014). MWMOTE-Majority weighted minority oversampling technique for imbalanced data set learning. . *IEEE Transactions on Knowledge Data Engineering*, *26*(2), 405–425.

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

Bellinger, C., Drummond, C., & Japkowicz, N. (2016). Beyond the boundaries of SMOTE - A framework for manifold-based synthetically oversampling. In *Machine Learning and Knowledge Discovery in Databases - European Conference (ECML PKDD)*, pp. 248–263.

Bellinger, C., Japkowicz, N., & Drummond, C. (2015). Synthetic oversampling for advanced radioactive threat detection. In *14th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 948–953.

Bhagat, R. C., & Patil, S. S. (2015). Enhanced SMOTE algorithm for classification of imbalanced big-data using random forest. In *Advance Computing Conference (IACC), 2015 IEEE International*, pp. 403–408.

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*, 106.

Blaszczynski, J., Deckert, M., Stefanowski, J., & Wilk, S. (2012). Iivotes ensemble for imbalanced data. *Intelligent Data Analysis*, *16*(5), 777–801.

Blaszczynski, J., & Stefanowski, J. (2015). Neighbourhood sampling in bagging for imbalanced data. . *Neurocomputing*, *150*, 529–542.

Borowska, K., & Stepaniuk, J. (2016). Imbalanced data classification: A novel re-sampling approach combining versatile improved SMOTE and rough sets. In *Computer Information Systems and Industrial Management - 15th IFIP TC8 International Conference (CISIM)*, pp. 31–42.

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modelling under imbalanced distributions. *ACM Computing Surveys*, *49*(2), 31:1–31:50.

Bruzzone, L., & Serpico, S. B. (1997). Classification of imbalanced remote-sensing data by neural networks. *Pattern Recognition Letters*, *18*(11-13), 1323–1328.

Brzezinski, D., & Stefanowski, J. (2017). Prequential AUC: Properties of the area under the roc curve for data streams with concept drift. *Knowledge and Information Systems*, *52*(2), 531–562.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, (PAKDD '09), pp. 475–482.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2012). DBSMOTE: Density-based synthetic minority over-sampling TEchnique. *Applied Intelligence*, *36*(3), 664–684.

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

Cao, H., Li, X., Woon, D. Y., & Ng, S. (2011). SPO: structure preserving oversampling for imbalanced time series classification. In *11th IEEE International Conference on Data Mining (ICDM)*, pp. 1008–1013.

Cao, H., Li, X., Woon, D. Y., & Ng, S. (2013). Integrated oversampling for imbalanced time series classification. *IEEE Transactions on Knowledge and Data Engineering*, *25*(12), 2809–2822.

Cao, P., Liu, X., Yang, J., & Zhao, D. (2017a). A multi-kernel based framework for heterogeneous feature selection and over-sampling for computer-aided detection of pulmonary nodules. *Pattern Recognition*, *64*, 327–346.

Cao, P., Liu, X., Zhang, J., Zhao, D., Huang, M., & Zaïane, O. R. (2017b). l21 norm regularized multi-kernel based joint nonlinear feature selection and over-sampling for imbalanced data classification. *Neurocomputing*, *234*, 38–57.

Cao, Q., & Wang, S. (2011). Applying over-sampling technique based on data density and cost-sensitive svm to imbalanced learning. In *International Conference on Information Management, Innovation Management and Industrial Engineering*, pp. 543–548.

Cateni, S., Colla, V., & Vannucci, M. (2011). Novel resampling method for the classification of imbalanced datasets for industrial and other real-world problems. In *11th International Conference on Intelligent Systems Design and Applications (ISDA)*, pp. 402–407.

Cervantes, J., García-Lamont, F., Rodríguez-Mazahua, L., Chau, A. L., Ruiz-Castilla, J. S., & Trueba, A. (2017). PSO-based method for SVM classification on skewed data sets. *Neurocomputing*, *228*, 187–197.

Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F. (2015). MLSMOTE: approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, *89*, 385–397.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over–sampling technique. *Journal of Artificial Intelligent Research*, *16*, 321–357.

Chawla, N. V., Japkowicz, N., & Kolcz, A. (2004). Editorial: Special issue on class imbalances. *SIGKDD Explorations*, *6*(1), 1–6.

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

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. In *Proceedings of 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'03)*, pp. 107–119.

Chen, L., Cai, Z., Chen, L., & Gu, Q. (2010a). A novel differential evolution-clustering hybrid resampling algorithm on imbalanced datasets. . In *WKDD*, pp. 81–85.

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

Chen, S., Guo, G., & Chen, L. (2010c). A new over-sampling method based on cluster ensembles. In *7th International Conference on Advanced Information Networking and Applications Workshops*, pp. 599–604.

Cieslak, D. A., & Chawla, N. V. (2008). Start globally, optimize locally, predict globally: Improving performance on imbalanced data. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pp. 143–152. IEEE.

Cieslak, D. A., Hoens, T. R., Chawla, N. V., & 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., & Geissbühler, A. (2006). Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence in Medicine*, *37*(1), 7–18.

Dal Pozzolo, A., Caelen, O., & Bontempi, G. (2015). When is undersampling effective in unbalanced classification tasks?. In Appice, A., Rodrigues, P. P., Costa, V. S., Soares, C., Gama, J., & Jorge, A. (Eds.), *ECML/PKDD*, Vol. 9284 of *Lecture Notes in Computer Science*, pp. 200–215. Springer.

Dang, X. T., Tran, D. H., Hirose, O., & Satou, K. (2015). SPY: A novel resampling method for improving classification performance in imbalanced data. In *2015 Seventh International Conference on Knowledge and Systems Engineering (KSE)*, pp. 280–285.

Das, B., Krishnan, N. C., & Cook, D. J. (2015). RACOG and wRACOG: Two probabilistic oversampling techniques. *IEEE Transactions on Knowledge and Data Engineering*, *27*(1), 222–234.

de la Calleja, J., & Fuentes, O. (2007). A distance-based over-sampling method for learning from imbalanced data sets. In *Proceedings of the Twentieth International Florida Artificial Intelligence*, pp. 634–635.

de la Calleja, J., Fuentes, O., & González, J. (2008). Selecting minority examples from misclassified data for over-sampling. In *Proceedings of the Twenty-First International Florida Artificial Intelligence*, pp. 276–281.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107–113.

Deepa, T., & Punithavalli, M. (2011). An E-SMOTE technique for feature selection in high-dimensional imbalanced dataset. In *3rd International Conference on Electronics Computer Technology (ICECT)*, pp. 322–324.

Dietterich, T. G., Lathrop, R. H., & Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, *89*(1), 31–71.

Ditzler, G., & Polikar, R. (2013). Incremental learning of concept drift from streaming imbalanced data. *IEEE Transactions on Knowledge Data Engineering*, *25*(10), 2283–2301.

Ditzler, G., Polikar, R., & Chawla, N. V. (2010). An incremental learning algorithm for non-stationary environments and class imbalance. In *20th International Conference on Pattern Recognition (ICPR)*, pp. 2997–3000.

Dong, A., Chung, F.-l., & Wang, S. (2016). Semi-supervised classification method through oversampling and common hidden space. *Information Sciences*, *349*, 216–228.

Dong, Y., & Wang, X. (2011). A new over-sampling approach: Random-SMOTE for learning from imbalanced data sets. In *Knowledge Science, Engineering and Management - 5th International Conference (KSEM)*, pp. 343–352.

Douzas, G., & Bacao, F. (2017). Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning. *Expert Systems with Applications*, *82*, 40–52.

Ertekin, S. (2013). Adaptive oversampling for imbalanced data classification. In *Information Sciences and Systems 2013 - Proceedings of the 28th International Symposium on Computer and Information Sciences (ISCIS)*, pp. 261–269.

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

Fan, X., Tang, K., & Weise, T. (2011). Margin-based over-sampling method for learning from imbalanced datasets. In *Advances in Knowledge Discovery and Data Mining - 15th Pacific-Asia Conference (PAKDD)*, pp. 309–320.

Farquad, M. A. H., & Bose, I. (2012). Preprocessing unbalanced data using support vector machine. *Decision Support Systems*, *53*(1), 226–233.

Fernandez, A., del Rio, S., Chawla, N. V., & Herrera, F. (2017). An insight into imbalanced big data classification: Outcomes and challenges. *Complex and Intelligent Systems*, *3*(2), 105–120.

Fernandez, A., Garcia, S., del Jesus, M. J., & 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.

Fernandez, A., Garcia, S., Luengo, J., Bernado-Mansilla, E., & Herrera, F. (2010). Genetics-based machine learning for rule induction: State of the art, taxonomy and comparative study. *IEEE Transactions on Evolutionary Computation*, *14*(6), 913–941.

Fernández, A., López, V., Galar, M., Del Jesus, M. J., & Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-based systems*, *42*, 97–110.

Fernández, A., Río, S., López, V., Bawakid, A., del Jesus, M. J., Benítez, J. M., & Herrera, F. (2014). Big data with cloud computing: An information sciencesight on the computing environment, mapreduce and programming framework. *WIREs Data Mining and Knowledge Discovery*, *4*(5), 380–409.

Fernandez, A., Carmona, C. J., del Jesus, M. J., & Herrera, F. (2017). A pareto based ensemble with feature and instance selection for learning from multi-class imbalanced datasets. *International Journal of Neural Systems*, *27*(6), 1–21.

Fernández-Navarro, F., Hervás-Martínez, C., & Gutiérrez, P. A. (2011). A dynamic over-sampling procedure based on sensitivity for multi-class problems. *Pattern Recognition*, *44*(8), 1821–1833.

Frank, E., & Pfahringer, B. (2006). Improving on bagging with input smearing. In *Advances in Knowledge Discovery and Data Mining, 10th Pacific-Asia Conference (PAKDD)*, pp. 97–106.

Friedman, J. H. (1996). Another approach to polychotomous classification. Tech. rep., Department of Statistics, Stanford University.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, *44*(8), 1761–1776.

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

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2016). Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced datasets. . *Information Sciences*, *354*, 178–196.

Gao, K., Khosgoftaar, T. M., & Wald, R. (2014a). the use of under- and oversampling within ensemble feature selection and classification for software quality prediction. *International Journal of Reliability, Quality and Safety Engineering*, *21*(1).

Gao, M., Hong, X., Chen, S., Harris, C. J., & Khalaf, E. (2014b). PDFOS: PDF estimation based over-sampling for imbalanced two-class problems. *Neurocomputing*, *138*, 248–259.

García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, *98*, 1–29.

García, V., Mollineda, R. A., & Sánchez, J. S. (2008). On the k-nn performance in a challenging scenario of imbalance and overlapping. *Pattern Analysis and Applications*, *11*(3-4), 269–280.

García, V., Sánchez, J. S., de J. Ochoa Domínguez, H., & Cleofas-Sánchez, L. (2015). Dissimilarity-based learning from imbalanced data with small disjuncts and noise. . In Paredes, R., Cardoso, J. S., & Pardo, X. M. (Eds.), *IbPRIA*, Vol. 9117 of *Lecture Notes in Computer Science*, pp. 370–378. Springer.

Gazzah, S., & Amara, N. E. B. (2008). New oversampling approaches based on polynomial fitting for imbalanced data sets. In *The Eighth IAPR International Workshop on Document Analysis Systems*, pp. 677–684.

Gazzah, S., Hechkel, A., & Amara, N. E. B. (2015). A hybrid sampling method for imbalanced data. In *12th International Multi-Conference on Systems, Signals and Devices*, pp. 1–6.

Ghazikhani, A., Monsefi, R., & Sadoghi Yazdi, H. (2013). Ensemble of online neural networks for non-stationary and imbalanced data streams. *Neurocomputing*, *122*, 535–544.

Gong, C., & Gu, L. (2016). A novel SMOTE-based classification approach to online data imbalance problem. *Mathematical Problems in Engineering*, *Article ID 5685970*, 14.

Gong, J., & Kim, H. (2017). RHSBoost: improving classification performance in imbalance data. *Computational Statistics and Data Analysis*, *111*(C), 1–13.

Gu, Q., Cai, Z., & Zhu, L. (2009). Classification of imbalanced data sets by using the hybrid re-sampling algorithm based on isomap. . In *ISICA*, Vol. 5821 of *Lecture Notes in Computer Science*, pp. 287–296.

Guo, H., & Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach. *SIGKDD Explorations Newsletter*, *6*, 30–39.

Gutierrez, P. D., Lastra, M., Bacardit, J., Benitez, J. M., & Herrera, F. (2016). GPU-SME-kNN: Scalable and memory efficient kNN and lazy learning using GPUs. *Information Sciences*, *373*, 165–182.

Gutierrez, P. D., Lastra, M., Benitez, J. M., & Herrera, F. (2017). SMOTE-GPU: Big data preprocessing on commodity hardware for imbalanced classification. *Progress in Artificial Intelligence*, *6*(4), 347–354.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. with Applicat.* , *73*, 220 – 239.

Hamid, Y., Sugumaran, M., & Journaux, L. (2016). A fusion of feature extraction and feature selection technique for network intrusion detection. *International Journal of Security and its Applications*, *10*(8), 151–158.

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline–SMOTE: A new over–sampling method in imbalanced data sets learning. In *Proceedings of the 2005 International Conference on Intelligent Computing (ICIC'05)*, Vol. 3644 of *Lecture Notes in Computer Science*, pp. 878–887.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the 2008 IEEE International Joint Conference Neural Networks (IJCNN'08)*, pp. 1322–1328.

He, H., & Chen, S. (2011). Towards incremental learning of nonstationary imbalanced data stream: A multiple selectively recursive approach. *Evolving Systems*, *2*(1), 35–50.

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

Herrera, F., Charte, F., Rivera, A. J., & del Jesús, M. J. (2016a). *Multilabel Classification - Problem Analysis, Metrics and Techniques*. Springer.

Herrera, F., Ventura, S., Bello, R., Cornelis, C., Zafra, A., Tarragó, D. S., & Vluymans, S. (2016b). *Multiple Instance Learning - Foundations and Algorithms*. Springer.

Ho, T. K., & Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(3), 289–300.

Hoens, T. R., & Chawla, N. V. (2010). Generating diverse ensembles to counter the problem of class imbalance. . In Zaki, M. J., Yu, J. X., Ravindran, B., & Pudi, V. (Eds.), *Proceedings in Advances in Knowledge Discovery and Data Mining (PAKDD)*, Vol. 6119 of *Lecture Notes in Computer Science*, pp. 488–499. Springer.

Hoens, T. R., & Chawla, N. V. (2013). *Imbalanced datasets: from sampling to classifiers*, pp. 43–59. Wiley.

Hoens, T. R., Chawla, N. V., & Polikar, R. (2011). Heuristic updatable weighted random subspaces for non-stationary environments. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pp. 241–250. IEEE.

Hoens, T. R., Polikar, R., & Chawla, N. V. (2012a). Learning from streaming data with concept drift and imbalance: an overview. *Progress in Artificial Intelligence*, *1*(1), 89–101.

Hoens, T. R., Qian, Q., Chawla, N. V., & Zhou, Z.-H. (2012b). Building decision trees for the multi-class imbalance problem. . In Tan, P.-N., Chawla, S., Ho, C. K., & Bailey, J. (Eds.), *Proceedings in Advances in knowledge discovery and data mining (PAKDD)*, Vol. 7301 of *Lecture Notes in Computer Science*, pp. 122–134. Springer.

Hoens, T. R., & Chawla, N. V. (2012). Learning in non-stationary environments with class imbalance. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–176. ACM.

Hu, F., & Li, H. (2013). A novel boundary oversampling algorithm based on neighborhood rough set model: NRSBoundary-SMOTE. *Mathematical Problems in Engineering*, *Article ID 694809*, 10.

Hu, F., Li, H., Lou, H., & Dai, J. (2014). A parallel oversampling algorithm based on nrsboundary-smote. *Journal of Information and Computational Science*, *11*(13), 4655–4665.

Hu, F., & Li, H. (2013). A novel boundary oversampling algorithm based on neighborhood rough set model: Nrsboundary-smote. *Mathematical Problems in Engineering*, *2013*, 694809.

Hukerikar, S., Tumma, A., Nikam, A., & Attar, V. (2011). SkewBoost: An algorithm for classifying imbalanced datasets. In *Proceedings of the International Conference on Computer Communication Technology (ICCCT)*, pp. 46–52.

Hwang, D., Fotouhi, F., Finley Jr., R. L., & Grosky, W. I. (2003). Predictive model for yeast protein functions using modular neural approach. . In *Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering (BIBE03)*, pp. 436–441. IEEE Computer Society.

Iglesias, E. L., Vieira, A. S., & Borrajo, L. (2013). An HMM-based over-sampling technique to improve text classification. *Expert Systems with Applications*, *40*(18), 7184–7192.

Japkowicz, N., & Holte, R. (2000). Workshop report: AAAI2000 workshop on learning from imbalanced data-sets. *AI Magazine*, *22*(1), 127–136.

Jeatrakul, P., & Wong, K. W. (2012). Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Jiang, K., Lu, J., & Xia, K. (2016). A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE. *Arabian Journal for Science and Engineering*, *41*(1), 3255–3266.

Jiang, L., Qiu, C., & Li, C. (2015). A novel minority cloning technique for cost-sensitive learning. *International Journal of Pattern Recognition and Artificial Intelligence*, *29*(4).

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

Kang, Y.-I., & Won, S. (2010). Weight decision algorithm for oversampling technique on class-imbalanced learning. In *ICCAS*, pp. 182–186.

Khan, S. H., Bennamoun, M., Sohel, F. A., & Togneri, R. (2018). Cost sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, *in press, doi: 10.1109/TNNLS.2017.2732482*, 1–15.

Khoshgoftaar, T. M., Hulse, J. V., & Napolitano, A. (2011). Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, *41*(3), 552–568.

Koto, F. (2014). SMOTE-out, SMOTE-cosine, and selected-SMOTE: An enhancement strategy to handle imbalance in data level. In *6th International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 193–197.

Koziarski, M., Krawczyk, B., & Wozniak, M. (2017). Radial-based approach to imbalanced data oversampling. . In de Pisón, F. J. M., Urraca-Valle, R., Quintián, H., & Corchado, E. (Eds.), *HAIS*, Vol. 10334 of *Lecture Notes in Computer Science*, pp. 318–327. Springer.

Krawczyk, B., Galar, M., Jelen, L., & Herrera, F. (2016). Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Applied Soft Computing Journal*, *38*, 714–726.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232.

Krawczyk, B., Minku, L. L., Gama, J. a., Stefanowski, J., & Woźniak, M. (2017). Ensemble learning for data stream analysis: a survey. *Information Fusion*, *37*, 132–156.

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

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: one–sided selection. In *Proceedings of the 14th International Conference on Machine Learning (ICML'97)*, pp. 179–186.

Lachheta, P., & Bawa, S. (2016). Combining synthetic minority oversampling technique and subset feature selection technique for class imbalance problem. In *ACM International Conference Proceeding Series*, pp. 1–8.

Last, M. (2002). Online classification of nonstationary data streams. *Intelligent Data Analysis*, *6*(2), 129–147.

Lee, J., Kim, N., & Lee, J. (2015). An over-sampling technique with rejection for imbalanced class learning. In *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pp. 102:1–102:6.

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*, 1–5.

Li, F., Yu, C., Yang, N., Xia, F., Li, G., & Kaveh-Yazdy, F. (2013a). Iterative nearest neighborhood oversampling in semisupervised learning from imbalanced data. . *The Scientific World Journal.* , *Article ID 875450*.

Li, H., Zou, P., Wang, X., & Xia, R. (2013b). A new combination sampling method for imbalanced data. In *Proceedings of 2013 Chinese Intelligent Automation Conference*, pp. 547–554.

Li, J., Fong, S., Wong, R. K., & Chu, V. W. (201). Adaptive multi-objective swarm fusion for imbalanced data classification. *Information Fusion*, *39*, 1–24.

Li, J., Fong, S., & Zhuang, Y. (2015). Optimizing SMOTE by metaheuristics with neural network and decision tree. In *2015 3rd International Symposium on Computational and Business Intelligence (ISCBI)*, pp. 26–32.

Li, K., Zhang, W., Lu, Q., & Fang, X. (2014). An improved SMOTE imbalanced data classification method based on support degree. In *International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI)*, pp. 34–38.

Liang, Y., Hu, S., Ma, L., & He, Y. (2009). MSMOTE: Improving classification performance when training data is imbalanced. In *Computer Science and Engineering, International Workshop on*, Vol. 2, pp. 13–17.

Lichtenwalter, R. N., Lussier, J. T., & Chawla, N. V. (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 243–252. ACM.

Lin, W.-J., & Chen, J. J. (2013). Class-imbalanced classifiers for high-dimensional data. . *Briefings in Bioinformatics*, *14*(1), 13–26.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141.

Lopez, V., Fernandez, A., & Herrera, F. (2014). On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed. *Information Sciences*, *257*, 1–13.

Lopez, V., Fernandez, A., Moreno-Torres, J. G., & Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Systems with Applications*, *39*(7), 6585–6608.

López, V., Triguero, I., Carmona, C. J., García, S., & Herrera, F. (2014). Addressing imbalanced classification with instance generation techniques: Ipade-id. *Neurocomputing*, *126*, 15–28.

Luengo, J., Fernández, A., García, S., & Herrera, F. (2011). Addressing data complexity for imbalanced data sets: analysis of SMOTE–based oversampling and evolutionary undersampling. *Soft Computing*, *15*(10), 1909–1936.

Ma, L., & Fan, S. (2017). CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. *BMC Bioinformatics*, *18*, 169.

Maciejewski, T., & Stefanowski, J. (2011). Local neighbourhood extension of SMOTE for mining imbalanced data. . In *CIDM*, pp. 104–111.

Mahmoudi, S., Moradi, P., Ahklaghian, F., & Moradi, R. (2014). Diversity and separable metrics in over-sampling technique for imbalanced data classification. In *4th International eConference on Computer and Knowledge Engineering (ICCKE)*, pp. 152–158.

Mao, W., Wang, J., & Wang, L. (2015). Online sequential classification of imbalanced data by combining extreme learning machine and improved SMOTE algorithm. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Martín-Félez, R., & Mollineda, R. A. (2010). On the suitability of combining feature selection and resampling to manage data complexity. In *Proceedings of the Conferencia de la Asociación Española de Inteligencia Artificial (CAEPIA'09)*, Vol. 5988 of *Lecture Notes on Artificial Intelligence*, pp. 141–150.

Mathew, J., Luo, M., Pang, C. K., & Chan, H. L. (2015). Kernel-based SMOTE for SVM classification of imbalanced datasets. In *Industrial Electronics Society, IECON 2015-41st Annual Conference of the IEEE*, pp. 001127–001132.

Maua, G., & Galinac Grbac, T. (2017). Co-evolutionary multi-population genetic programming for classification in software defect prediction: An empirical case study. *Applied Soft Computing Journal*, *55*, 331–351.

Mease, D., Wyner, A. J., & Buja, A. (2007). Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research*, *8*, 409–439.

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discoveryd*, *28*(1), 92–122.

Mera, C., Arrieta, J., Orozco-Alzate, M., & Branch, J. (2015). A bag oversampling approach for class imbalance in multiple instance learning. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 20th Iberoamerican Congress (CIARP)*, pp. 724–731.

Mera, C., Orozco-Alzate, M., & Branch, J. (2014). Improving representation of the positive class in imbalanced multiple-instance learning. In *Image Analysis and Recognition - 11th International Conference (ICIAR)*, pp. 266–273.

Mirza, B., Lin, Z., & Liu, N. (2015). Ensemble of subset online sequential extreme learning machine for class imbalance and concept drift. . *Neurocomputing*, *149*, 316–329.

Moniz, N., Branco, P., & Torgo, L. (2016). Resampling strategies for imbalanced time series. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA) 2016*, pp. 282–291.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodriguez, R., Chawla, N. V., & Herrera, F. (2012a). A unifying view on dataset shift in classification. *Pattern Recognition*, *45*(1), 521–530.

Moreno-Torres, J. G., Sáez, J. A., & Herrera, F. (2012b). Study on the impact of partition-induced dataset shift on-fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(8), 1304–1312.

Moreno-Torres, J. G., & Herrera, F. (2010). A preliminary study on overlapping and data fracture in imbalanced domains by means of genetic programming-based feature extraction. In *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications (ISDA'10)*, pp. 501–506.

Moutafis, P., & Kakadiaris, I. A. (2014). GS4: generating synthetic samples for semi-supervised nearest neighbor classification. In *Trends and Applications in Knowledge Discovery and Data Mining (PAKDD)*, pp. 393–403.

Nakamura, M., Kajiwara, Y., Otsuka, A., & Kimura, H. (2013). LVQ-SMOTE - learning vector quantization based synthetic minority over-sampling technique for biomedical data. *BioData Mining*, *6*, 16.

Napierala, K., & Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, *46*(3), 563–597.

Napierala, K., Stefanowski, J., & Wilk, S. (2010). Learning from imbalanced data in presence of noisy and borderline examples. In *Proceedings of the 7th International Conference on Rough Sets and Current Trends in Computing (RSCTC'10)*, Vol. 6086 of *Lecture Notes on Artificial Intelligence*, pp. 158–167.

Nekooeimehr, I., & Lai-Yuen, S. K. (2016). Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets. *Expert Systems with Applications*, *46*(C), 405–416.

Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Online learning from imbalanced data streams. In *Proceedings of the 2011 International Conference of Soft Computing and Pattern Recognition, SoCPaR 2011*, pp. 347–352.

Orriols-Puig, A., Bernadó-Mansilla, E., Goldberg, D. E., Sastry, K., & Lanzi, P. L. (2009). Facetwise analysis of XCS for problems with class imbalances. *IEEE Transactions on Evolutionary Computation*, *13*, 260–283.

Palacios, A. M., Sánchez, L., & Couso, I. (2012). Equalizing imbalanced imprecise datasets for genetic fuzzy classifiers. *International Journal of Computational Intelligence Systems*, *5*(2), 276–296.

Pan, S., Wu, J., Zhu, X., & Zhang, C. (2015). Graph ensemble boosting for imbalanced noisy graph stream classification. *IEEE Transactions on Cybernetics*, *45*(5), 940–954.

Park, Y., Qi, Z., Chari, S. N., & Molloy, I. (2014). PAKDD'12 best paper: generating balanced classifier-independent training samples from unlabeled data. *Knowledge and Information Systems*, *41*(3), 871–892.

Pekalska, E., & Duin, R. P. W. (2005). *The Dissimilarity Representation for Pattern Recognition - Foundations and Applications*, Vol. 64 of *Series in Machine Perception and Artificial Intelligence*. World Scientific.

Peng, L., Zhang, H., Yang, B., Chen, Y., & Zhou, X. (2016). SMOTE-DGC: an imbalanced learning approach of data gravitation based classification. In *Intelligent Computing Theories and Application - 12th International Conference (ICIC)*, pp. 133–144.

Peng, Y., & Yao, J. (2010). AdaOUBoost: adaptive over-sampling and under-sampling to boost the concept learning in large scale imbalanced data sets. In *Multimedia Information Retrieval*, pp. 111–118.

Pérez-Ortiz, M., Gutiérrez, P. A., & Hervás-Martínez, C. (2013). Borderline kernel based over-sampling. In *Hybrid Artificial Intelligent Systems - 8th International Conference (HAIS)*, pp. 472–481.

Pérez-Ortiz, M., Gutiérrez, P. A., Hervás-Martínez, C., & Yao, X. (2015). Graph-based approaches for over-sampling in the context of ordinal regression. *IEEE Transactions on Knowledge Data Engineering*, *27*(5), 1233–1245.

Pérez-Ortiz, M., Gutiérrez, P. A., Tiño, P., & Hervás-Martínez, C. (2016). Oversampling the minority class in the feature space. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(9), 1947–1961.

Piras, L., & Giacinto, G. (2012). Synthetic pattern generation for imbalanced learning in image retrieval. *Pattern Recognition Letters*, *33*(16), 2198–2205.

Pourhabib, A., Mallick, B. K., & Ding, Y. (2015). Absent data generating classifier for imbalanced class sizes. *Journal of Machine Learning Research*, *16*(1), 2695–2724.

Prati, R. C., & Batista, G. E. A. P. A. (2004). Class imbalances versus class overlapping: an analysis of a learning system behavior. In *Proceedings of the 2004 Mexican International Conference on Artificial Intelligence (MICAI'04)*, pp. 312–321.

Prati, R. C., Batista, G. E. A. P. A., & Silva, D. F. (2015). Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. *Knowledge and Information Systems*, *45*(1), 247–270.

Puntumapon, K., & Waiyamai, K. (2012). A pruning-based approach for searching precise and generalized region for synthetic minority over-sampling. In *Advances in Knowledge Discovery and Data Mining - 16th Pacific-Asia Conference (PAKDD)*, pp. 371–382.

Radovanovic, M., Nanopoulos, A., & Ivanovic, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. . *Journal of Machine Learning Research*, *11*, 2487–2531.

Radtke, P. V. W., Granger, E., Sabourin, R., & Gorodnichy, D. O. (2014). Skew-sensitive boolean combination for adaptive ensembles - an application to face recognition in video surveillance. *Information Fusion*, *20*(1), 31–48.

Ramentol, E., Caballero, Y., Bello, R., & Herrera, F. (2012). 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., Gondres, I., Lajes, S., Bello, R., Caballero, Y., Cornelis, C., & Herrera, F. (2016). Fuzzy-rough imbalanced learning for the diagnosis of high voltage circuit breaker maintenance: The SMOTE-FRST-2T algorithm. *Engineering Applications of AI*, *48*, 134–139.

Ramírez-Gallego, S., Fernández, A., García, S., Chen, M., & Herrera, F. (2018). Big data: Tutorial and guidelines on information and process fusion for analytics algorithms with mapreduce. *Information Fusion*, *42*, 51–61.

Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., & Herrera, F. (2017). A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, *239*, 39–57.

Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(3), 252–264.

Río, S., López, V., Benítez, J. M., & Herrera, F. (2014). On the use of mapreduce for imbalanced big data using random forest. *Information Sciences*, *285*, 112–137.

Rivera, W. A. (2017). Noise reduction a priori synthetic over-sampling for class imbalanced data sets. *Information Sciences*, *408*, 146–161.

Rivera, W. A., & Xanthopoulos, P. (2016). A priori synthetic over-sampling methods for increasing classification sensitivity in imbalanced data sets. *Expert Systems with Appications.* , *66*, 124–135.

Rokach, L. (2016). Decision forest: Twenty years of research. *Information Fusion*, *27*, 111–125.

Rong, T., Gong, H., & Ng, W. W. Y. (2014). Stochastic sensitivity oversampling technique for imbalanced data. In *ICMLC (CCIS volume)*, Vol. 481 of *Communications in Computer and Information Science*, pp. 161–171.

Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). SMOTE-IPF: addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, *291*, 184–203.

Sánchez, A. I., Morales, E. F., & Gonzalez, J. A. (2013). Synthetic oversampling of instances using clustering. *International Journal on Artificial Intelligence Tools*, *22*(2).

Sandhan, T., & Choi, J. Y. (2014). Handling imbalanced datasets by partially guided hybrid sampling for pattern recognition. In *22nd International Conference on Pattern Recognition (ICPR)*, pp. 1449–1453.

Schapire, R. E. (1999). A brief introduction to boosting. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'99)*, pp. 1401–1406.

Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V., & Folleco, A. (2014). An empirical study of the classification performance of learners on imbalanced and noisy software quality data. *Information Sciences*, *259*, 571–595.

Sen, A., Islam, M. M., Murase, K., & Yao, X. (2016). Binarization with boosting and oversampling for multiclass classification. *IEEE Transactions on Cybernetics*, *46*(5), 1078–1091.

Shimodaira, H. (2000). Improving predictive inference under Covariate Shift by Weighting the Log-likelihood Function. *Journal of Statistical Planning and Inference*, *90*(2), 227–244.

Stefanowski, J. (2016). Dealing with data difficulty factors while learning from imbalanced data. . In Matwin, S., & Mielniczuk, J. (Eds.), *Challenges in Computational Statistics and Data Mining*, Vol. 605 of *Studies in Computational Intelligence*, pp. 333–363. Springer.

Stefanowski, J., & Wilk, S. (2008). Selective pre-processing of imbalanced data for improving classification performance. In *Data Warehousing and Knowledge Discovery, 10th International Conference*, pp. 283–292.

Storkey, A. (2009). When training and test sets are different: Characterizing learning transfer. In nonero Candela, J. Q., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.), *Dataset Shift in Machine Learning*, pp. 3–28. MIT Press.

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(4), 687–719.

Tang, B., & He, H. (2015). KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning. In *IEEE Congress on Evolutionary Computation (CEC)*, pp. 664–671.

Tang, S., & Chen, S. (2008). The generation mechanism of synthetic minority class examples. In *5th International Conference on Information Technology and Applications in Biomedicine (ITAB)*, pp. 444–447.

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

Thanathamathee, P., & Lursinsap, C. (2013). Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and adaboost techniques. *Pattern Recognition Letters*, *34*(12), 1339–1347.

Tomasev, N., & Mladenic, D. (2013). Class imbalance and the curse of minority hubs. . *Knowledge-Based Systems*, *53*, 157–172.

Torgo, L., Branco, P., Ribeiro, R. P., & Pfahringer, B. (2015). Resampling strategies for regression. *Expert Systems*, *32*(3), 465–476.

Torres, F. R., Carrasco-Ochoa, J. A., & Martínez Trinidad, J. F. (2016). SMOTE-D a deterministic version of SMOTE. In *Pattern Recognition - 8th Mexican Conference (MCPR)*, pp. 177–188.

Triguero, I., García, S., & Herrera, F. (2015). SEG-SSC: A framework based on synthetic examples generation for self-labeled semi-supervised classification. *IEEE Transactions on Cybernetics*, *45*(4), 622–634.

Verbiest, N., Ramentol, E., Cornelis, C., & Herrera, F. (2014). Preprocessing noisy imbalanced datasets using smote enhanced with fuzzy rough prototype selection. . *Applied Soft Computing*, *22*, 511–517.

Vorraboot, P., Rasmequan, S., Chinnasarn, K., & Lursinsap, C. (2015). Improving classification rate constrained to imbalanced data between overlapped and non-overlapped regions by hybrid algorithms. . *Neurocomputing*, *152*, 429–443.

Wang, J., Xu, M., Wang, H., & Zhang, J. (2006). Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. In *8th International Conference on Signal Processing (ICSP)*, Vol. 3, pp. 1–6. IEEE.

Wang, J., Yun, B., li Huang, P., & ao Liu, Y. (2013a). Applying threshold smote algorithm with attribute bagging to imbalanced datasets. In *International Conference on Rough Sets and Knowledge Technology*, pp. 221–228.

Wang, J., Yao, Y., Zhou, H., Leng, M., & Chen, X. (2013b). A new over-sampling technique based onsvm for imbalanced diseases data. In *International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)*, pp. 1224–1228.

Wang, Q., Luo, Z., Huang, J., Feng, Y., & Liu, Z. (2017). A novel ensemble method for imbalanced data learning: Bagging of extrapolation-SMOTE SVM. *Computational Intelligence and Neuroscience*, *2017*, 1827016:1–1827016:11.

Wang, S., Minku, L. L., & Yao, X. (2013). A learning framework for online class imbalance learning. In *Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Ensemble Learning, CIEL 2013 - 2013 IEEE Symposium Series on Computational Intelligence, SSCI 2013*, pp. 36–45.

Wang, S., & Yao, X. (2012). Multiclass imbalance problems: Analysis and potential solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *42*(4), 1119–1130.

Wang, S., Li, Z., Chao, W., & Cao, Q. (2012). Applying adaptive over-sampling technique based on data density and cost-sensitive SVM to imbalanced learning. In *The 2012 International Joint Conference on Neural Networks (IJCNN),*, pp. 1–8.

Wang, S., Minku, L. L., Ghezzi, D., Caltabiano, D., Tiño, P., & Yao, X. (2013a). Concept drift detection for online class imbalance learning. . In *IJCNN*, pp. 1–10. IEEE.

Wang, S., Minku, L. L., & Yao, X. (2013b). Online class imbalance learning and its applications in fault detection. . *International Journal of Computational Intelligence and Applications*, *12*(4).

Wang, S., Minku, L. L., & Yao, X. (2015). Resampling-based ensemble methods for online class imbalance learning. . *IEEE Transactions on Knowledge Data Engineering*, *27*(5), 1356–1368.

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

Wang, X., Liu, X., Japkowicz, N., & Matwin, S. (2013). Resampling and cost-sensitive methods for imbalanced multi-instance learning. In *13th IEEE International Conference on Data Mining Workshops (ICDM)*, pp. 808–816.

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

Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., & Petitjean, F. (2016). Characterizing concept drift. *Data Mining and Knowledge Discovery*, *30*(4), 964–994.

Weiss, G. M., & Provost, F. J. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, *19*, 315–354.

Wilson, D. R., & Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, *6*, 1–34.

Xie, Z., Jiang, L., Ye, T., & Li, X. (2015). A synthetic minority oversampling method based on local densities in low-dimensional space for imbalanced learning. In *Database Systems for Advanced Applications - 20th International Conference (DASFAA)*, pp. 3–18.

Xu, Y. H., Le, L. P., & Tian, X. Y. (2014). Neighborhood triangular synthetic minority over-sampling technique for imbalanced prediction on small samples of chinese tourism and hospitality firms. In *Seventh International Joint Conference on Computational Sciences and Optimization*, pp. 534–538.

Yamazaki, K., Kawanabe, M., Watanabe, S., Sugiyama, M., & Müller, K.-R. (2007). Asymptotic bayesian generalization error when training and test distributions are different. . In Ghahramani, Z. (Ed.), *ICML*, Vol. 227 of *ACM International Conference Proceeding Series*, pp. 1079–1086. ACM.

Yin, H., & Gai, K. (2015). An empirical study on preprocessing high-dimensional class-imbalanced data for classification. In *High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), 2015 IEEE 12th International Conferen on Embedded Software and Systems (ICESS), 2015 IEEE 17th International Conference on*, pp. 1314–1319.

Yin, L., Ge, Y., Xiao, K., Wang, X., & Quan, X. (2013). Feature selection for high-dimensional imbalanced data. *Neurocomputing*, *105*, 3–11.

Yongqing, Z., Min, Z., Danling, Z., Gang, M., & Daichuan, M. (2013). Improved SMOTE-Bagging and its application in imbalanced data classification. In *Conference Anthology, IEEE*, pp. 1–6.

Young, W. A., Nykl, S. L., Weckman, G. R., & Chelberg, D. M. (2015). Using voronoi diagrams to improve classification performances when modeling imbalanced datasets. *Neural Computing and Applications*, *26*(5), 1041–1054.

Yun, J., Ha, J., & Lee, J. (2016). Automatic determination of neighborhood size in SMOTE. In *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pp. 100:1–100:8.

Zhai, J., Zhang, S., & Wang, C. (2017). The classification of imbalanced large data sets based on mapreduce and ensemble of elm classifiers. . *International Journal of Machine Learning and Cybernetics*, *8*(3), 1009–1017.

Zhang, H., Yang, J., Xie, J., Qian, J., & Zhang, B. (2017). Weighted sparse coding regularized nonconvex matrix regression for robust face recognition. *Information Sciences*, *394-395*, 1–17.

Zhang, H., & Li, M. (2014). RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. *Information Fusion*, *20*, 99–116.

Zhang, H., & Wang, Z. (2011a). A normal distribution-based over-sampling approach to imbalanced data classification. In *Advanced Data Mining and Applications - 7th International Conference (ADMA)*, pp. 83–96.

Zhang, L., & Wang, W. (2011b). A re-sampling method for class imbalance learning with credit data. In *International Conference on Information Technology, Computer Engineering and Management Sciences (ICM)*, pp. 393–397.

Zhou, A., Qu, B. Y., Li, H., Zhao, S. Z., Suganthan, P. N., & Zhangd, Q. (2011). Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm and Evolutionary Computation*, *1*(1), 32–49.

Zhou, B., Yang, C., Guo, H., & Hu, J. (2013). A quasi-linear SVM combined with assembled SMOTE for imbalanced data classification. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7.

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

Zhu, X., Goldberg, A. B., Brachman, R., & Dietterich, T. (2009). *Introduction to Semi-Supervised Learning*. Morgan and Claypool Publishers.

Zieba, M., Tomczak, J. M., & Gonczarek, A. (2015). RBM-SMOTE: restricted boltzmann machines for synthetic minority oversampling technique. In *Intelligent Information and Database Systems - 7th Asian Conference (ACIIDS)*, pp. 377–386.

Zikopoulos, P. C., Eaton, C., deRoos, D., Deutsch, T., & Lapis, G. (2011). *Understanding Big Data - Analytics for Enterprise Class Hadoop and Streaming Data* (1st edition). McGraw-Hill Osborne Media.

Zuo, Y., Zhao, J., & Xu, K. (2016). Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, *48*(2), 379–398.