Expert Systems With Applications 73 (2017) 220–239

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Learning from class-imbalanced data: Review of methods and applications

Guo Haixiang[a,b,c,1,*], Li Yijing[a,b,1,*], Jennifer Shang[d,*], Gu Mingyun[a], Huang Yuanyue[a], Gong Bing[e]

[a] College of Economics and Management, China University of Geosciences, Wuhan 430074, China
[b] Research Center for Digital Business Management, China University of Geosciences, Wuhan 430074, China
[c] Mineral Resource Strategy and Policy Research Center of China University of Geosciences(WUHAN), Wuhan 43007, China
[d] The Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA 15260, USA
[e] Department of Industrial Engineering, Business Administration and Statistic, E.T.S Industrial Engineering, Universidad Politécnica de Madrid, C/José Gutiérrez Abascal, 2- 20086, Madrid, Spain

ARTICLE INFO

Article history:
Received 6 September 2016
Revised 23 November 2016
Accepted 25 December 2016
Available online 30 December 2016

Keywords:
Rare events
Imbalanced data
Machine learning
Data mining

ABSTRACT

Rare events, especially those that could potentially negatively impact society, often require humans' decision-making responses. Detecting rare events can be viewed as a prediction task in data mining and machine learning communities. As these events are rarely observed in daily life, the prediction task suffers from a lack of balanced data. In this paper, we provide an in depth review of rare event detection from an imbalanced learning perspective. Five hundred and seventeen related papers that have been published in the past decade were collected for the study. The initial statistics suggested that rare events detection and imbalanced learning are concerned across a wide range of research areas from management science to engineering. We reviewed all collected papers from both a technical and a practical point of view. Modeling methods discussed include techniques such as data preprocessing, classification algorithms and model evaluation. For applications, we first provide a comprehensive taxonomy of the existing application domains of imbalanced learning, and then we detail the applications for each category. Finally, some suggestions from the reviewed papers are incorporated with our experiences and judgments to offer further research directions for the imbalanced learning and rare event detection fields.

1. Introduction

Rare events, unusual patterns and abnormal behavior are difficult to detect, but often require responses from various management functions in a timely manner. By definition, rare events refer to events that occur much less frequently than commonly occurring events (Maalouf and Trafalis, 2011). Examples of rare events include software defects (Rodriguez et al., 2014), natural disasters (Maalouf and Trafalis, 2011), cancer gene expressions (Yu et al., 2012), fraudulent credit card transactions (Panigrahi et al., 2009), and telecommunications fraud (Olszewski, 2012).

In the field of data mining, detecting events is a prediction problem, or, typically, a data classification problem. Rare events are difficult to detect because of their infrequency and casualness; however, misclassifying rare events can result in heavy costs. For financial fraud detection, invalid transactions may only emerge out of hundreds of thousands of transaction records, but failing to identify a serious fraudulent transaction would cause enormous losses. The scarce occurrences of rare events impair the detection task to imbalanced data classification problem. Imbalanced data refers to a dataset within which one or some of the classes have a much greater number of examples than the others. The most prevalent class is called the majority class, while the rarest class is called the minority class (Li et al., 2016c). Although data mining approaches have been widely used to build classification models to guide commercial and managerial decision-making, classifying imbalanced data significantly challenges these traditional classification models. As were discussed on existing surveys, the reasons are fivefold:

* Corresponding authors.
E-mail addresses: faterdumk0732@sina.com (G. Haixiang), liyijing024@hotmail.com, liyijing024@gmail.com (L. Yijing), shang@katz.pitt.edu (J. Shang), 550312686@qq.com (G. Mingyun), huangyuanyue1991@126.com (H. Yuanyue), gongbing1112@gmail.com (G. Bing).
[1] Guo Haixiang and Li Yijing contributed equally in this paper.

http://dx.doi.org/10.1016/j.eswa.2016.12.035
0957-4174/

(1) Standard classifiers such as logistic regression, Support Vector Machine (SVM) and decision tree are suitable for balanced training sets. When facing imbalanced scenarios, these models often provide suboptimal classification results, i.e. a good coverage of the majority examples, whereas the minority examples are distorted (López et al., 2013).
(2) The learning process guided by global performance metrics such as prediction accuracy induces a bias towards the majority class, while the rare episodes remain unknown even if the prediction model produces a high overall precision (Loyola-González et al., 2016). Some original discussion can be found in Weiss and Hirsh (2000) and Weiss (2004).
(3) Rare minority examples may possibly be treated as noise by the learning model. Contrarily, noise may be wrongly identified as minority examples, since both of them are rare patterns in the data space (Beyan and Fisher, 2015).
(4) Even though skewed sample distributions are not always difficult to learn (such as when the classes are separable), minority examples usually overlap with other regions where the prior probabilities of both classes are almost equal. Denil and Trappenberg (2010) has discussed overlapping problem under imbalanced scenario.
(5) Besides, small disjuncts (Jo and Japkowicz, 2004), a lack of density and small sample size with high feature dimensionality (Wasikowski and Chen, 2010) are challenges to imbalanced learning, which often cause learning models to fail in detecting rare patterns (Branco et al., 2016; López et al., 2013).

Many machine learning approaches have been developed in the past decade to cope with imbalanced data classification, most of which have been based on sample techniques, cost sensitive learning and ensemble methods (Galar et al., 2012; Krawczyk et al., 2014; Loyola-González et al., 2016). There is also one book in this area, see He and Ma (2013). Although several surveys related to imbalanced learning have been published (Branco et al., 2016; Fernández et al., 2013; Galar et al., 2012; He and Garcia, 2009; López et al., 2012; Sun et al., 2009), all of them focused on detailed techniques while application literature is neglected. For researchers from management, biology or other domains, rather than sophisticated algorithms, the problems that can be solved using imbalanced learning techniques and the building of imbalanced learning systems with mature yet effective methods may be of more concern.

In this paper, we aim to provide a thorough overview of the classification of imbalanced data that includes both techniques and applications. At the technical level, we introduce common approaches to deal with imbalanced learning and propose a general framework within which each algorithm can be placed. This framework is a unified data mining model and includes preprocessing, classification and evaluation. At the practical level, we review 162 papers that tried to build specific systems to tackle rare pattern detection problems and develop a taxonomy of existing imbalanced learning application domains. The existing application literature covers most research fields from medical to industry to management.

The rest of this paper is organized as follows. Section 2 describes the research methodology for this study, along with the initial statistics regarding recent trends in imbalanced learning. Section 3 presents approaches to address both binary and multiple class imbalanced data. In Section 4, we first categorize the existing imbalanced learning application literature into 13 domains and then introduce the respective research frameworks. In Section 5, we discuss our thoughts on future imbalanced learning research directions from both a technical and practical point of view. Finally, Section 6 present the conclusions of this paper.

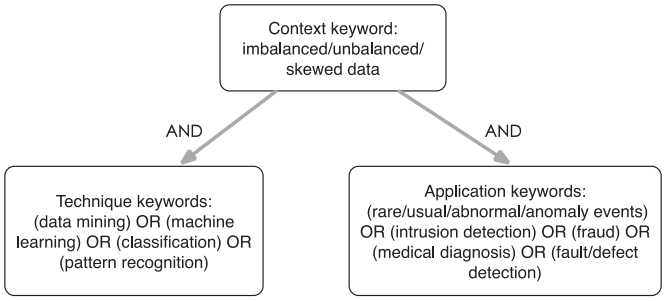

Fig. 1. Two-stage keywords tree structure (we use "/" to separate synonyms of the same concept).

2. Research methodology and initial statistics

2.1. Research methodology

For this study, based on the research methodology of Govindan and Jepsen (2016), a two-stage search procedure was conducted to compile relevant papers published from 2006 to October 2016. In the initial phase, seven library databases which covered most natural science and social science research fields were used to search for and collect literature: Elsevier, IEEExplore, Springer, ACM, Cambridge, Wiley and Sage. Full text search was used and the search term was designed following the search process outlined in Fahimnia et al. (2015). A two-level keywords tree was constructed to provide a comprehensive set of search terms to capture technical and application articles on rare events and imbalanced learning. Fig. 1 presents the search terms for this study. The search phase on the first level was restricted to imbalanced /unbalanced /skewed data, as the focus was on imbalanced data classification. The second level search terms were divided in twofold to cover both technical and practical articles. For techniques, keywords that referred to data mining approaches were used and for practical applications, key words were used that focused on event detection and prediction that included rare events, usual events, abnormal events, defective procedures, fraud, disease, and intrusion. Note that the corresponding inflected forms of a word and the synonyms (such as "anomaly" for "abnormal", "fraudulence" for "fraud") were also considered. The initial search yielded 657 papers, which were downloaded for the next filtering process.

After reviewing each paper manually, 464 papers were found to be relevant to this study. The second stage search was performed during the review and the relevant cross-references were searched for using Google Scholar. At this stage, attempts were made to access all cross-references in Google Scholar or were included in the accessible library databases. After the second stage, 63 papers were added to our review. Therefore, a total of 527 papers were included in this study.

2.2. Initial statistics

In this section, we present the initial statistics for the trends in imbalanced learning. Fig. 2 shows the publishing trends by plotting the quantity of publications from 2006 to 2016. A relatively stable growth in the number of publications can be observed after 2006. The only downward trend in the 2011/2013 interval was made up immediately by a sharp rise in the number of publications in 2013–2016. This trend suggests imbalanced learning hitherto remains a valuable research topic.

The initial statistics also show that 192 journals and conference proceedings published a total of 527 papers. The contribution of each journal was counted and the top 20 journals/conferences are shown in Fig. 3. These journals covered 43.5% of all published

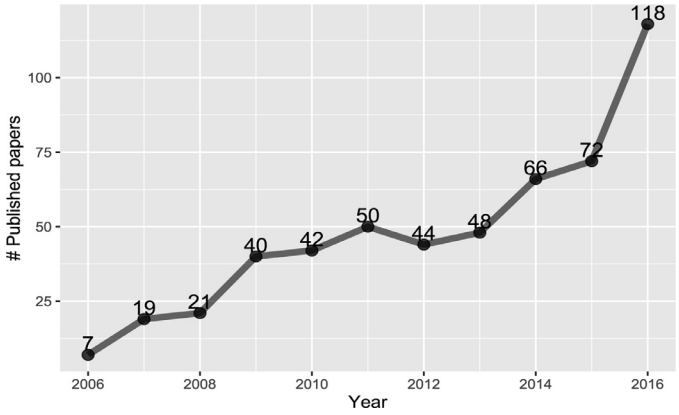

Fig. 2. Imbalanced learning publishing trends.

papers in the past decade, most of which are influential journals in the fields of computer science, operations research, management science, engineering and biotechnology. This also illustrates that imbalanced learning research has included both learning techniques from the computer science community and practical applications across a broad range of domains from natural science to management science.

We also collected the titles of all the papers involved in this study and generated a word cloud to capture the most studied imbalanced learning topics. To build the word cloud presented in Fig. 4, we first removed the stop words most commonly used in English such as "the" and "and", then NLTK tool was employed to lemmatize each word. As our goal was to uncover detailed topics in these imbalanced learning papers, some general and frequently appearing words were also eliminated.[1] Fig. 4 shows some of the specific techniques for the imbalanced data classification: resampling methods ("over-sampling", "sampling", "under-sampling", etc.), machine learning approaches ("cost-sensitve", "support vector machine", "neural network", "ensemble", etc.) and expert systems ("rule", "system", "fuzzy", etc.). However, some concrete application domains are also shown in the word cloud. Typical word clusters covered several domains, such as "patient", "fraud detection", "telecommunications", and "credit card", which implied that the main categories for the applied papers may lie in fields such as financial management, medical diagnosis, and telecommunications.

3. Imbalanced data classification approaches

Hundreds of algorithms have been proposed in the past decade to address imbalanced data classification problems. In this section, we give an overview of the state-of-the-art imbalanced learning techniques. These techniques are discussed in line with the basic machine learning model framework. In Section 3.1, two basic strategies for addressing imbalanced learning are introduced, which are, preprocessing and cost-sensitive learning. Preprocessing approaches include resampling methods conducted in the sample space and feature selection methods that optimize the feature space. The strategies introduced in Section 3.1 are then integrated into the classification models described in Section 3.2. The classifiers are further divided into ensemble classifiers and algorithmic modified classifiers. Section 3.3 discusses multi-class classification as a special imbalanced learning issue to clarify the extension of these binary classification algorithms into a multi-class case. In Section 3.4, the metrics for evaluating and selecting the models are introduced.

3.1. Basic strategies for dealing with imbalanced learning

3.1.1. Preprocessing techniques

Preprocessing is often performed before building learning model so as to attain better input data. Considering the representation spaces of data, two classical techniques are often employed as preprocessor:

3.1.1.1. Resampling. Resampling techniques are used to rebalance the sample space for an imbalanced dataset in order to alleviate the effect of the skewed class distribution in the learning process. Resampling methods are more versatile because they are independent of the selected classifier (López et al., 2013). Resampling techniques fall into three groups depending on the method used to balance the class distribution:

- Over-sampling methods: eliminating the harms of skewed distribution by creating new minority class samples. Two widely-used methods to create the synthetic minority samples are randomly duplicating the minority samples and SMOTE (Chawla et al., 2002).
- Under-sampling methods: eliminating the harms of skewed distribution by discarding the intrinsic samples in the majority class. The simplest yet most effective method is Random Under-Sampling (RUS), which involved the random elimination of majority class examples (Tahir et al., 2009).
- Hybrid methods: these are a combination of the over-sampling method and the under-sampling method.

We found 156 reviewed papers have utilized resampling techniques, accounting for 29.6% of all papers reviewed. This indicates that resampling is a popular strategy for dealing with imbalanced data. Under-sampling methods are employed 39 times, whereas over-sampling methods are used 84 times; and hybrid-sampling methods are chosen 33 times. Instead using existing resampling methods, 52 papers developed novel re-sampling methods. As cluster-based methods (e.g. k-means), distance based methods (e.g. nearest neighbors) and evolutionary methods (e.g. generic algorithm) were the most frequently used strategies to generate or eliminate examples, these have been summarized in Table 1. For method that did not employ the above strategies, we give a brief introduction in the last column.

It should be noted that all of the re-sampling methods allow to re-sample to any desired ratio, and it is not necessary and sufficient to exactly balance the number of majority and minority classes. Zhou (2013) recommended different sample ratio for different data size, Napierala and Stefanowski (2015) studied types of minority class examples and their influences on learning classifiers from imbalanced data. Some papers tried to automatically decide best sampling rate for different Imbalanced Ratios (IR) and problem settings (Lu et al., 2016; Moreo et al., 2016; Ren et al., 2016a; Tomek, 1976; Yun et al., 2016; Yeh et al., 2016; Zhang et al., 2016a).

Three papers studied the performances of different re-sampling methods (Loyola-González et al., 2016; Napierala and Stefanowski, 2015; Zhou 2013). There are some insights from these papers: 1). When there are hundreds of minority observations in the dataset, an under-sampling method was superior to an over-sampling method in terms of computational time. 2). When there are only a few dozen minority instances, the over-sampling method SMOTE was found to be a better choice. 3). If the training sample size is too large, a combination of SMOTE and under-sampling is suggested as an alternative. 4). SMOTE is slightly more effective in recognizing outliers.

[1] These words are "class-imbalanced", "imbalanced", "classification", "class-unbalanced", "unbalanced", "approaches", "approach", "based", "using", "data-sets", "datasets", "sets", "data".

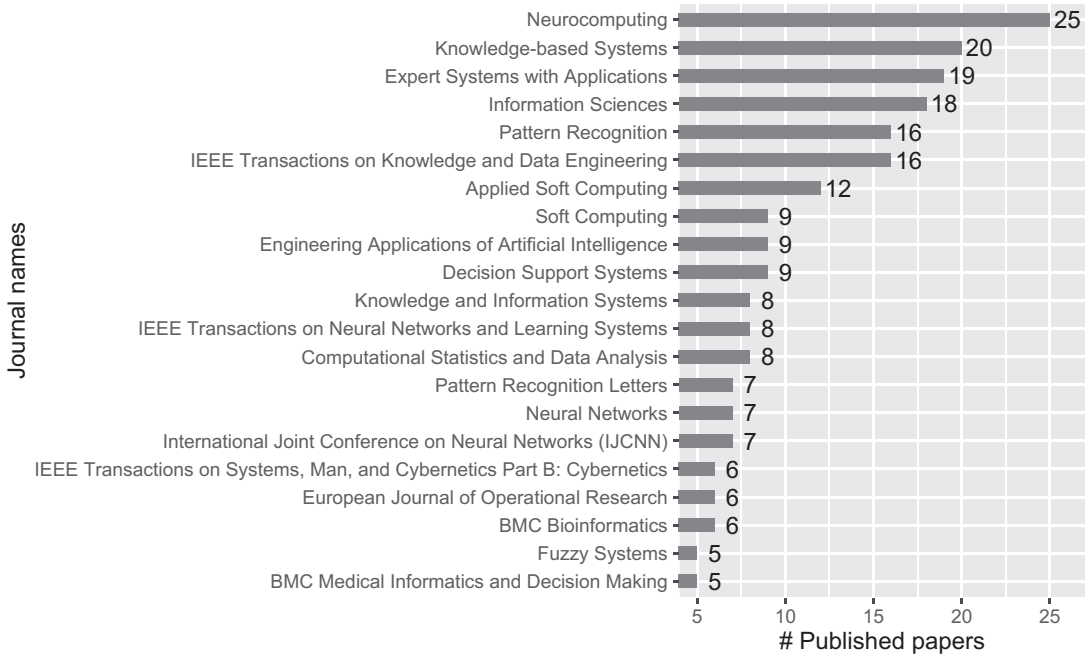

Fig. 3. Top 20 journals/conferences in which most imbalanced learning papers were published.

Fig. 4. Word cloud of the title words from the collected papers.

3.1.1.2. Feature selection and extraction. Compared to the progress achieved in resampling methods, significantly fewer papers considered feature selection. Under imbalanced scenarios, minority class samples can easily be discarded as noise; however, if the irrelevant features in the feature space are removed, this risk is reduced (Li et al., 2016c).

The goal of feature selection, in general, is to select a subset of k features from the entire feature space that allows a classifier to achieve optimal performance, where k is a user-specified or adaptively selected parameter. Feature selection can be divided into filters, wrappers, and embedded methods (Guyon and Elisseeff, 2003). The advantages and disadvantages of these methods can be found in Saeys et al. (2007).

Another way to deal with dimensionality is feature extraction. Feature extraction is related to dimensionality reduction, which transforms data into a low-dimensional space. It should be noted, however, that feature selection techniques are different from feature extraction. Feature extraction creates new features from the original features using functional mapping, whereas feature selection returns a subset of the original features (Motoda and Liu, 2002). There are a variety of techniques for feature extraction, such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and f Non-negative Matrix Factorization (NMF) (see Hartmann, 2004). Feature extraction methods tend to be employed more frequently for unstructured data such as images, text and speech.

Table 2 shows 31 reviewed articles employed feature selection or feature extraction. It was found that filter and wrapper feature selection methods were most regularly used. For filter methods, various metrics were used to rank the features, and for wrapper

Table 1
Summary of articles employing re-sampling methods.

Category	Strategy	Articles	Basic ideas behind the novel algorithms
Over-sampling	Cluster-based	Nekooeimehr and Lai-Yuen (2016), Cao et al. (2014)	
	Distance based (e.g. SMOTE and a modified version)	15 papers such as Li et al. (2016a), Yun et al. (2016), Zhang et al. (2015a), García et al. (2012), Zhai et al. (2015)	
	Evolutionary based	Li et al. (2016a), Yang et al. (2009)	
	Other novel method	Chetchotsak et al. (2015)	Created new data with a self-organizing map
		Menardi and Torelli (2014)	Bootstrap form of random over-sampling using a kernel method
		Das et al. (2015)	Used joint probability distribution of data attributes and Gibbs sampling to generate new minority class samples
		Cao et al. (2013)	Over-sampling time series data.
Under-sampling	Cluster-based	6 papers such as Sun et al. (2015), Li et al. (2013a), Kumar et al. (2014)	
	Distance based	D'Addabbo and Maglietta (2015), Anand et al. (2010)	
	Evolutionary based	6 papers such as Ha and Lee (2016), Galar et al. (2013), Krawczyk et al. (2014)	
Hybrid-sampling	Random/distance based over-sampling + under-sampling	8 papers such as Cateni et al. (2014), Dubey et al. (2014), Díez-Pastor et al. (2015a)	
	Cluster-based under-sampling + distance based over-sampling	8 papers such as Peng et al. (2014), Sáez et al. (2015), Song et al. (2016)	
	Other novel methods	Jian et al. (2016)	Over-sampling support-vectors and under-sampling non-support- vectors

Table 2
Summary of articles employing feature selection or extraction methods.

Article	Method	Detail strategy
Lusa (2010), Wei et al. (2013b), Lane et al. (2012), Li et al. (2013a), Lima and Pereira (2015)	Filter	Feature ranking based on correlation evaluation (two tailed Student *t*-test, Relief-F, Gini index, Information gain, Chi-Square)
Bae and Yoon (2015), Alibeigi et al. (2012), Yang et al. (2016a), Li et al. (2016c), Gong and Huang (2012)	Filter	Feature ranking based on a probability density estimation/ pairwise sequence similarity
Beyan and Fisher (2015) Chen and Wasikowski (2008), Maldonado et al. (2014), Trafalis et al. (2014), Casañola-Martin et al. (2016), Al-Ghraibah et al. (2015)	Wrapper	Heuristic search (sequential forward/backward selection, hill-climbing search)
Wei et al. (2013a), Li et al. (2016c), Guo et al. (2016)	Wrapper	Stochastic search (random selection, evolutionary method)
Dubey et al. (2014)	Embedded	Sparse logistic regression with stability selection
Song et al. (2014), Moepya et al. (2014), Vong et al. (2015), Zhang (2016)	Feature extraction	Traditional methods (PCA, NMF, SVD, etc.)
Braytee et al. (2016) / Ng et al. (2016)	Novel feature extraction methods	Cost-sensitive PCA and cost-sensitive NMF / Yielded useful feature representation using stacked auto-encoders

methods, heuristic search was a common choice. Another interesting finding is, that feature selection and extraction were frequently used for solving real-world problem such as disease diagnosis (Casañola-Martin et al., 2016; Dubey et al., 2014; Lusa, 2010; Yang et al., 2016a; Zhang, 2016), textual sentiment analysis (Lane et al., 2012; Zhang et al., 2015a), fraud detection (Li et al., 2013a; Lima and Pereira, 2015 Moepya et al., 2014; Wei et al., 2013b) and other rare events detection problems (Al-Ghraibah et al., 2015; Bae and Yoon, 2015; Gong and Huang, 2012; Guo et al., 2016; Vong et al., 2015), etc.

3.1.2. Cost-sensitive learning

By assuming higher costs for the misclassification of minority class samples with respect to majority class samples, cost-sensitive learning can be incorporated both at the data level (e.g. re-sampling and feature selection) and at the algorithmic level (see Section 3.2, López et al., 2012, 2013). The costs are often specified as cost matrices, where C_{ij} represents the misclassification cost of assigning examples belong to class i to class j. Given a specific domain, cost matrices can be determined using expert opinion, or in data stream scenarios, they can vary for each record or vary in a dynamic imbalanced status (Ghazikhani et al., 2013b). Compared with re-sampling methods, cost-sensitive learning is more computationally efficient, thus may be more suitable for big data streams. However, this method was used by only 39 of the reviewed papers, which is much less popular than re-sampling methods. There may be two potential reasons, one is that, as stated in Krawczyk et al. (2014), it is difficult to set values in the cost matrix. In most cases, as the misclassification cost is unknown from the data and cannot be given by an expert. Nevertheless, there is an alternative way to address this difficulty, by setting the majority class misclassification cost at 1 while setting the penalty minority class value as equal to the IR (Castro and Braga, 2013; López et al., 2015). Another reason, which may be more reasonable for this observation, is that re-sampling is a common choice in practical for those researchers who are not expert in machine learning. Different from cost-sensitive learning which often needs to modify learning algorithm, re-sampling methods are much easier to be directly implemented in both single and ensemble models. Throughout our study, most application papers employed re-sampling methods instead of cost-sensitive learning.

Three main approaches for dealing with cost-sensitive problems were found in the 39 related papers, as listed in Table 3.

Table 3
Summary of articles employing cost-sensitive learning.

Method	Detail strategy	Articles
Methods based on training data modification	Modifying the decision thresholds or assigning weights to instance when resampling the training dataset according to the cost decision matrix	Zhang et al. (2016a), Boyu Wang (2016), Zou et al. (2016), Yu et al. (2015), Yu et al. (2016), Voigt et al. (2014), Zhou and Liu (2006)
Changing the learning process or learning objective to build a cost-sensitive classifier	Modifying the objective function of SVM/ELM using a weighting strategy	Cheng et al. (2016), Wu et al. (2016), Casañola-Martin et al. (2016), Phoungphol et al. (2012), Maldonado and López (2014)
	Tree-building strategies that could minimize misclassification costs	del Río et al. (2014), Krawczyk et al. (2014), Sahin et al. (2013)
	Integrating a cost factor into the fuzzy rule-based classification system	López et al. (2015), Vluymans et al. (2015)
	Cost sensitive error function on neural network	Oh (2011), Castro and Braga (2013); Ghazikhani et al. (2013b)
	Cost-sensitive boosting methods	Sun et al. (2007), Wang et al. (2010)
Methods based on Bayes decision theory	Incorporating cost matrix into Bayes based decision boundary	Ali et al. (2016), Datta and Das (2015), Bahnsen et al. (2013), Moepya et al. (2014)

3.2. Classification algorithms for imbalanced learning

Imbalanced learning attempts to build a classification algorithm that can provide a better solution for class imbalance problems than traditional classifiers such as SVM, KNN, decision trees and neural networks. Two methods for solving imbalanced learning problems have been reported in the literature; ensemble methods and algorithmic classifier modifications. In Sections 3.2.1 and 3.2.2, we review the imbalanced data classification algorithms proposed in the past decade. While most of these proposed methods have targeted binary class problems, multi-class issues are commonly seen in many rare event detection domains, such as in machinery fault detection and disease diagnosis. For this reason, multi-class imbalanced learning is discussed as a special issue in Section 3.2.3, which gives a brief introduction to the current solutions.

3.2.1. Ensemble methods

Ensemble-based classifiers, also named multiple classifier systems (Krawczyk and Schaefer, 2013), are known to improve the performance of a single classifier by combining several base classifiers that outperform every independent one (López et al., 2013). Classifier ensembles have become a popular solution method for class imbalance problems. In the 527 reviewed papers, 218 papers proposed novel ensemble models or applied existing ensemble models to solve practical tasks. Galar et al. (2012) thoroughly surveyed imbalanced data learning using ensemble methods, in which ensemble methods were categorized into cost-sensitive ensembles and data pre-processing ensembles. However, as only bagging, boosting and hybrid ensembles (a combination of bagging and boosting) were considered in this study, the field was not fully covered. For example, Sun et al. (2015) and Tian et al. (2011) proposed two ensemble models that trained multiple base classifiers by balancing different datasets created using re-sampling methods without the need for bagging or boosting algorithms.

Note that the training process for the re-sampling based ensemble and bagging can be conducted in parallel, while boosting and some evolutionary based ensemble methods can only be trained using iterative processes. Therefore, in this study, we categorized the ensemble models into two categories, i.e. iterative based ensembles and parallel based ensembles.

3.2.1.1. Iterative based ensemble. Boosting is the most common and most effective method for ensemble learning. We found 63 reviewed papers employed boosting in their ensemble framework, most of which were based on the first applicable boosting algorithm, Adaboost, proposed by Freund and Schapire (1996). The benefit of Adaboost is that samples that fail to be assigned to the correct class are given higher weights, which forces the future classifier to focus more on learning these failed classified samples. Adaboost has had several extensions: AdaBoost.M1, AdaBoost.M2, AdaBoost.MR and AdaBoost.MH (Freund and Schapire, 1997; Schapire and Singer, 1999); which were designed for solving multi-class and multi-label problems. Other typical iterative ensemble methods include Gradient Boosting Decision Tree (GBDT) (Friedman, 2001) and some evolutionary algorithm (EA) based ensemble algorithms.

As mentioned in Galar et al. (2012), boosting algorithms are usually combined with cost-sensitive learning and re-sampling techniques. Table 4 lists some of the technical articles that have proposed novel iterative based ensemble algorithms. It can be seen that Adaboost and Adaboost.M2 have been the most popular iterative based ensemble schemes. Most ensemble models have considered cost-sensitive and re-sampling strategies.

3.2.1.2. Parallel based ensembles. In this study, parallel based ensembles refer to ensemble models in which each base classifier can be trained in parallel. Parallel based ensemble schemes include bagging, re-sampling based ensembles and feature selection based ensembles. The basic framework for parallel ensemble methods is shown in Fig. 5, in which the dashed boxes and lines represent optional processes. Galar et al. (2012) and López et al. (2013) suggested that bagging and the associated hybridizations with data preprocessing techniques have had good results compared with boosting. Through our research, it was found that parallel ensemble methods dominated two types of papers. First, parallel based ensemble methods were more popular than iterative based ensembles in application-oriented papers (such as in Hao et al., 2014, Wei et al., 2013b, Dai, 2015, etc.), and novel re-sampling methods were often combined with parallel based ensemble schemes (such as in Peng et al., 2014, Sun et al., 2015, Li et al., 2013a, etc.). Since parallel ensembles have time-saving and ease-of-development advantages, they are recommended for solving practical problems.

3.2.1.3. Base classifier choice in ensemble models. When implementing iterative or parallel ensemble methods, a base classifier is needed, which can be any of the classical models such as SVM and neural networks. Fig. 6 summarizes the main base classifiers selected by proposed ensemble learning algorithms. Note that some papers examined multiple base classifiers. We found that SVM, NN (neural network), NB (naïve Bayes), rule-based classifiers, and decision tree based classifiers (includes C4.5, CART, random forest and other novel tree classifiers) have been the most selected in the literature. Sun et al. (2009) summarized the difficulties of some base classifiers when learning from imbalanced data and pointed out

Table 4
Representatives of reviewed iterative ensemble methods.

Algorithms and reference	Ensemble scheme	Combined strategies
AdaC1, AdaC2, and AdaC3 (Sun et al., 2007), BoostingSVM-IB (Zięba et al., 2014), Boosting-SVM (Wang and Japkowicz, 2010), Adaboost.NC (Wang et al., 2010)	Adaboost (slightly modified re-weighting strategy)	Cost-sensitive
RUSBoost (Seiffert et al., 2010) /GESuperPBoost (García-Pedrajas and García-Osorio, 2013)	AdaBoost.M2 / Adaboost (the weighted error is modified)	RUS(Under-sampling)
EUSBoost (Galar et al., 2013), RAMOBoost (Chen et al., 2010), BNU-SVMs (Bao et al., 2016b)	AdaBoost.M2 or Adaboost	Novel under-sampling
MSMOTEBoost (Hu et al., 2009; Ren et al., 2016a)	Adaboost	Novel over-sampling
EasyEnsemble and BalanceCascade (Liu et al. 2009; Hassan and Abraham, 2016)	Adaboost	Bagging, RUS(Under-sampling)
BSIA (Zięba and Tomczak, 2015)	Adaboost	Under-sampling, cost-sensitive
Prusa et al. (2016) and Guo et al. (2016)	Adaboost, Adaboost.M1	Under-sampling, Feature selection
Boyu Wang (2016), Ditzler and Polikar (2013), Wang et al. (2013), Dal Pozzolo et al. (2015)	Adaboost	Online boosting, re-sampling
Li et al. (2014)	Adaboost	Weighted ELM
RB-Boost (Díez-Pastor et al., 2015a), ESM-UP-DW (Bae and Yoon, 2015)	Adaboost or Adaboost.M2	Hybrid-sampling
RankCost (Wan et al., 2014; Lusa, 2016)	GBDT	Cost sensitive
Can-CSC-GBE (Ali et al., 2016)	Gentleboost	Cost-sensitive
DyS (Lin et al., 2013a)	Novel iterative learning scheme	Dynamic sampling
M-AdaBoost cascade decision tree (Yi, 2010)	Novel cascaded structure based on Adaboost	
Krawczyk et al. (2014), Krawczyk et al. (2016) and Folino et al. (2016)	Evolutionary algorithm based ensemble	Cost sensitive, Feature selection
LEAT (Park and Ghosh, 2014)	Novel boosting mechanism	

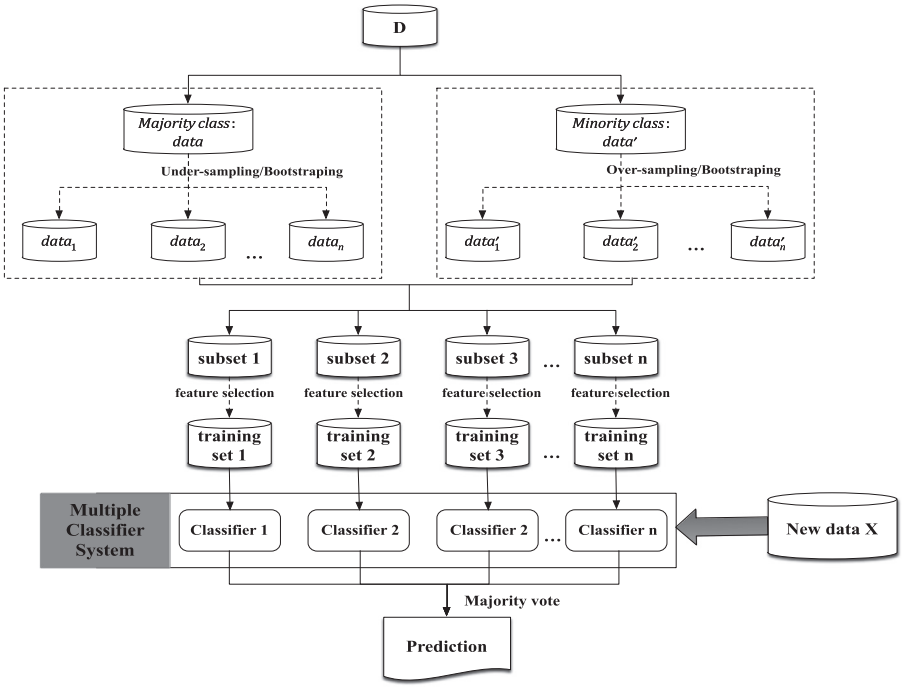

Fig. 5. Parallel ensemble framework (dashed frames and lines stand for optional operations).

that there were dozens of classifiers for machine learning, each with their own strengths and weaknesses. Users need to choose a proper base classifier grounded in reality. For example, SVM is robust and precise, but can be sensitive to missing values and difficult to train for large scale data. In contrast, decision trees are good at handling missing value cases but can fail to model small size data (Li et al., 2016c).

3.2.2. Algorithmic classifier modifications

Improving the learning ability of existing classification algorithms to improve the classification performance for imbalanced data is another main imbalanced learning research direction. Over 160 novel improved classifiers have been proposed in the past decades in the class imbalanced learning and rare event detection fields. SVM, decision tree, NN, KNN, ELM, rule-based classifiers, and

Table 5
Representatives of reviewed algorithmic classifier modifications.

Strategy	Detailed description	Representative articles
Kernel and activation function transformation method (Mostly based on SVM, ELM, and NN)	Enhances the discriminatory power of the classifiers using kernel transformation to separate two classes close to the boundary as far from each other as possible Designing mixed kernel/basis functions to increase the separability of the original training space	Logistic regression: Maalouf and Trafalis (2011); SVM: Zhang et al. (2014), Maratea et al. (2014), Zhao et al. (2011); ELM: Gao et al. (2016), Wang et al. (2016b); NN: Raj et al. (2016), da Silva and Adeodato (2011) SVM: Chen et al. (2012); NN: Pérez-Godoy et al. (2010); ELM: Wu et al. (2016)
Objective transformation method	Converting the training goal to a well-deigned objective function that more severely penalize errors in the minority examples Multi-objective optimization	Cost-sensitive learning: See Section 3.1.2; Weighted KNN: Kim et al. (2016); SVM: Duan et al. (2016b); SVM: Shao et al. (2014), Datta and Das (2015), Xu et al. (2015); Bayesian NN: Hong et al. (2007), Lan et al. (2009); KNN: Ando (2015), Huang et al. (2006b)
Fuzzy based method	Fuzzy rule-based classifiers (FRBCs) and decision trees to extract classification rules from the imbalanced data Other fuzzy based classifiers	Alshomrani et al. (2015), Fernández et al. (2010b), Ramentol et al. (2015), Vluymans et al. (2015), Bagherpour et al. (2016) SVM: Cheng and Liu (2015); KNN: Liao (2008) NN: Gao et al. (2014), Tan et al. (2015b)
Clustering algorithm	Unsupervised learning	Liang et al. (2012), Vigneron and Chen (2015), Fan et al. (2016), Zhang (2016)
Task decomposition strategies	A hierarchical decomposition strategy to globally and locally learn from the imbalanced data	Naïve Bayes and NN: Lerner et al. (2007); Rule based classifier: Napierała and Stefanowski (2015), Jacques et al. (2015); KNN: Garcia-Pedrajas et al. (2015), Beyan and Fisher (2015) ELM: Xiao et al. (2016), Mao et al. (2016)

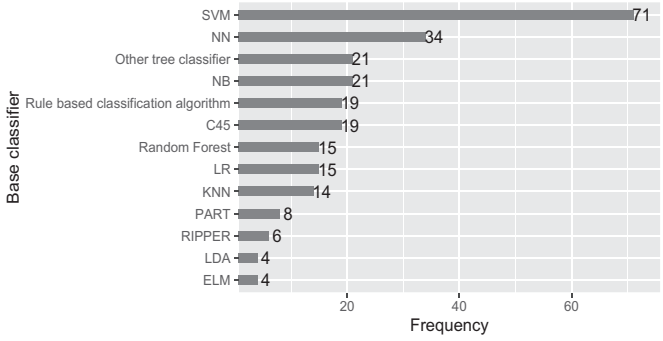

Fig. 6. Statistics for base classifiers. (NN: Neural Network, NB: Naïve Bayes, LR: Logistic Regression, PART: Projective ART (Cao and Wu, 2002), RIPPER: Repeated Incremental Pruning to Produce Error Reduction (Natwichai et al., 2005), LDA: Linear Discriminant Analysis, ELM: Extreme learning machine (Huang et al., 2006a).

naïve Bayes modifications were the most used at 54, 33, 24, 15, 13, 11, and 9 papers, respectively. Some common techniques for improving these 6 classifiers are summarized in Table 5.

3.2.3. Multi-class imbalanced learning

Multi-class learning has been seen as a difficult task for classification algorithms as multi-class classification may have a significantly lower performance than binary cases. This issue becomes more complex when facing imbalanced data, as the boundaries between the classes can severely overlap (Fernández et al., 2013). Multi-class imbalanced learning has attracted significant attention in recent years. A total of 32 of the reviewed papers generalized binary class imbalanced solutions into multi-class cases. Two generalization methods appeared to be the most used; a one-versus-one approach (OVO) and a one-versus-all approach (OVA), both of which are based on decomposition techniques. The OVO and OVA decomposition schemes are shown in Fig. 7 (Zhou, 2016), in which C_i represents all examples labeled i, and f_j is the hypothesis generated by classifier j.

Fernández et al. (2013) studied the OVO and OVA decompositions, and ad-hoc learning algorithms, which are natural for addressing multiple class learning problems. The experimental results showed that OVO outperforms OVA. However, there is no significant differences between decomposition methods and standard ad-hoc learning algorithms. Similar conclusions can be found in Wang and Yao (2012). They believe that it is unnecessary to use class decomposition, while learning from the entire data set directly is sufficient for multi-class imbalance classification. They concluded that it is unwise to integrate class decomposition with class imbalance techniques without considering the class distribution globally.

Although the above papers suggested that OVA is not the best available tool, it is still popular as it requires less decomposition thus is time efficient. Relative to OVA, OVO was less popular. A summary of OVO, OVA, ad-hoc related articles are shown in Table 6.

3.4. Model evaluation in the presence of rare classes

Model selection and model evaluation are two crucial processes in machine learning. Performance measures, therefore, are key indicators for both evaluating the effectiveness and guiding the learning of a classifier. Accuracy is the most commonly used evaluation metric for classification. However, under imbalanced scenarios, accuracy may not be a good choice because of the bias toward the majority class. Performance metrics adapted into imbalanced data problems, such as Receiver Operating Characteristics (ROC), G-Mean(GM), and F-measure(F_m), are less likely to suffer from imbalanced distributions as they take class distribution into account. Since these measures are widely used in imbalanced learning field and their detailed formulas can be found in most imbalanced learning related papers (such as Branco et al., 2016 and López et al., 2013), we only introduce these metrics in Appendix (see Supplementary Material). There are also some works focused on proposing novel evaluation metrics for imbalanced data, such as Adjusted F-measure (Maratea et al., 2014) and a probabilistic thresholding method (Su and Hsiao, 2007).

The most frequently used metrics include Accuracy, AUC/ROC, F-Measure, G-mean, Precision, Sensitivity, Specificity, Balanced accuracy, and Matthews Correlation Coefficient (MCC). Two papers

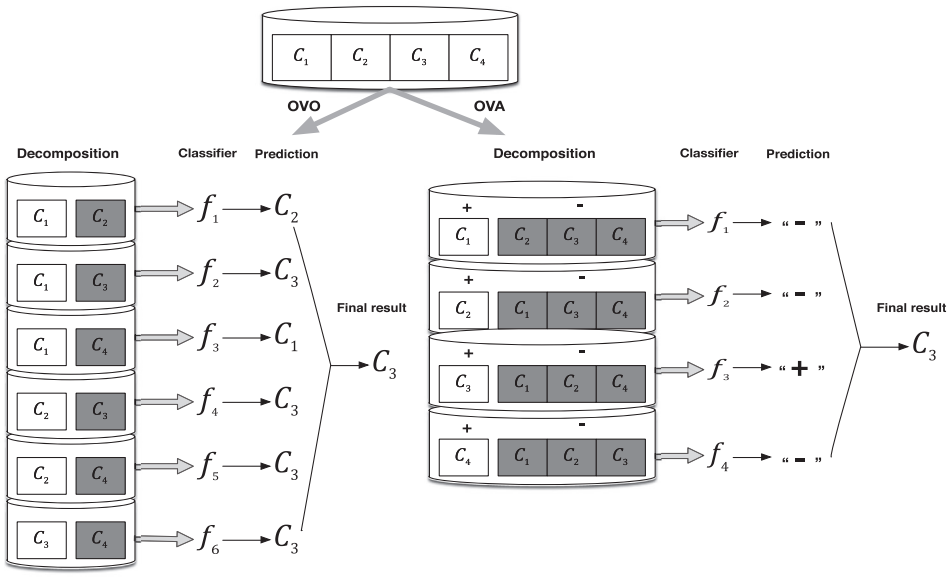

Fig. 7. OVO and OVA decomposition schemes. (Note: In the OVA scheme, one class is treated as a positive class and denoted "+", while the other three classes are merged as a negative class and denoted as "-". Based on the predictive labels for the four classifiers on the right side, it is clear that final result is C_3).

Table 6
Representatives of reviewed multi-class imbalanced learning algorithms.

Strategy	Detailed description	Representative articles
OVA	Re-sampling techniques with OVA Combining OVA with ensemble learning	Liao (2008), Zhao et al. (2008), Chen et al. (2006) Zhang et al. (2015d), Yang et al. (2016a)
OVO	Re-sampling with OVO Combining OVO with ensemble learning	Fernández et al. (2010a) Zhang et al. (2016b), Cerf et al. (2013)
Ad-hoc	Directly generalizing learning algorithms in to multi-class cases	SVM: Duan et al. (2016a), Shao et al. (2014); ELM: Yang et al. (2016a), Yu et al. (2016), Mirza et al. (2015a), Mirza et al. (2013)

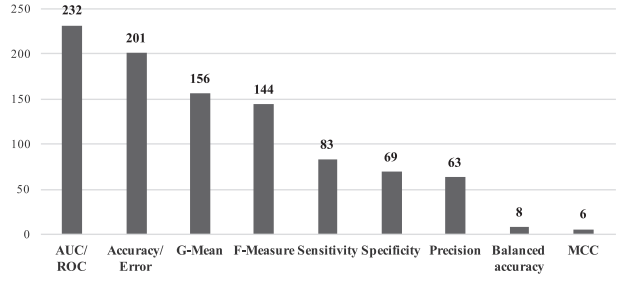

Fig. 8. The number papers using the evaluation metrics.

investigated the effectiveness of these metrics for imbalanced learning and made some recommendations (Gu et al., 2009; Jeni et al., 2013). A usage summary for all the metrics is given in Fig. 8. Note that although accuracy based metrics have been proven to be biased towards the majority class, these are still frequently used in research because they are the most general and intuitionistic metrics for classification tasks. AUC, the G-mean, and the F-measure have also been often utilized as evaluation measures for both model comparisons and model selections.

It should be noted that AUC/ROC has been questioned by Hand (2009), who argued that ROC depends on cutoffs the model produces and the cutoffs are related to the misclassification costs when only optimal thresholds are considered, hence the ROC of a model is incoherent. However, there are also some counter arguments about this interpretation (Ferri et al., 2011). Generally, AUC/ROC is recognized as a useful measure of ranking performance.

Metrics mentioned above are only applicable to binary classification problems. A natural way to extend these metrics to multi-class cases is to employ decomposition methods (the OVA and OVA schemes described in 3.2.3) and take the average of each pair-wise metric (Cerf et al., 2013). MAUC (Hand and Till, 2001) is an example of a mean AUC and in Guo et al. (2016) and Li et al. (2016c), a derivation of AUC was used. Another AUC based multi-class metric, Volume Under ROC (VUC) was employed in Phoungphol et al. (2012) and an extension of the G-mean was adapted by Sun et al. (2006). For the F-measure, two types of averages, micro-averages and macro-averages were introduced in Phoungphol et al. (2012).

4. Imbalanced data classification application domains

There is currently a great deal of interest in utilizing automated methods—particularly data mining and machine learning methods—to analyze the enormous amount of data that is routinely being collected. An important class of problems involves predicting future events based on past events. Event prediction often involves predicting rare events (Weiss and Hirsh, 2000). Rare events are events that occur with low frequency but may cause far-reaching impact and disrupt the society (King and Zeng, 2001). Rare events lurk in many shapes, including natural disaster (e.g. earthquakes, solar flares, tornado), anthropogenic hazards (e.g. financial fraud, industrial accidents, violent conflict) and diseases. Since such data are usually imbalanced, many efforts have been directed to build rare events detection systems with the help of imbalanced learning methods. Among the literature, we find 162 articles are application-oriented, ranging from management to engineering. In Section 4.1, we develop a taxonomy scheme to categorize these 162 articles into 13 application fields of rare events detection. In Sections 4.2–4.14, each category and its sub-categories are described and a few examples are given.

Table 7
Application domain categories.

Application domains	No. of papers
(1) Chemical, biomedical engineering	47
(2) Financial management	37
(3) Information technology	24
(4) Energy management	8
(5) Security management	7
(6) Electronics and communications	6
(7) Infrastructure and industrial manufacturing	9
(8) Business management	7
(9) Emergency management	4
(10) Environmental management	5
(11) Policy, social and education	4
(12) Agriculture and horticulture	1
(13) Other areas and non-specific areas	3

The main purpose of this section is to show researchers from different areas how imbalanced learning has been applied to detect rare patterns/events in their research domains. As learning techniques have already been presented in Section 3, we try to omit technical details in this section to avoid overlapping and focus more on describing specific problems and corresponding data collection and feature engineering processes in this section. Also, we found some interesting observations about which methods tend to be popularly used in different domains through our study, which are concluded in Section 6.

4.1. Taxonomy of application domains

Excluding popular benchmarks such as UCI and KEEL, we categorize the real-world applications into 13 categories. Each category contains several topics. We divided managerial applications into six categories: finance, energy, security, emergency, environment, and business management. The remaining categories involve engineering and humanity fields and are divided into: chemical, biological and medical engineering; information technology; electronics and communications; infrastructure and industrial manufacturing; policy, social and education; agriculture and horticulture; other areas. Table 7 shows the taxonomy and frequency distribution of the papers. Note that some applications may be interdisciplinary, we simply categorized them into the groups best describing their domain.

4.2. Chemical and biomedical engineering

Biomedical engineering applies engineering principles and design concepts to medicine and biology for healthcare purposes (e.g. diagnosis, monitoring and therapy), while chemical engineering tries to convert chemical materials or cells into useful forms and products. Both biomedical and chemical engineering employ physical and life sciences together with applied mathematics, computer science and economics. The application often involves building a decision support system to detect and predict abnormal structures in chemical processes and biomedical activities. These include disease diagnoses, disease early warning, protein detection, chemical repulsion and drug resistance. The related research problems and corresponding references are shown in Table 8.

The most researched topics related to imbalanced chemical and biomedical data classification problems were protein detection, gene expression prediction, and diseases diagnoses. Protein data sets are usually imbalanced data sets, and protein detection tries to identify protein structures and functions based on their sequence expressions (Dai, 2015). Feature extraction is essential in protein detection tasks to convert the non-numerical attributes in the sequences to numerical attributes (Vani and Sravani, 2014; Wasikowski and Chen, 2010). Similar properties can be found in gene expression identification (Yu et al., 2012) and DNA identification (Song et al., 2014). Image data are also utilized frequently to analyze abnormal biomedical events: Bria et al. (2012) used 198 images to detect microcalcifications on digital mammograms; Lerner et al. (2007) analyzed abnormal genes using FISH signal images; Bae and Yoon (2015) focused on finding the locations and sizes of polyps in endoscopic or colonoscopic images.

4.3. Financial management

Financial management is a sub-area of business management. Our review found 37 papers have tackled financial problems; we thus separated financial management from business management. Financial management is an activity of management which is concerned with the planning, procuring and controlling of the firm's financial resources. Sanz et al. (2015) tested imbalanced data classification methods in 11 financial problems including stock market prediction, credit card/loans approval application system and fraud detection. Other papers and their application setting are given in Table 9.

Most papers in this category were focused on financial fraud detection, which included e-payment fraud, credit and plastic card fraud, fraud in firm's activity, insurance fraud, and fraudulent company financial statements. Krivko (2010) highlighted the several challenges in building an effective fraud detection system; large volumes of daily transaction records, low frequency fraud occurrence and information delay. Detecting fraud events is a typical imbalanced learning problem as transaction records are highly skewed. In general, the dataset used to train a fraud detection system includes customer profiles and transaction records (type of transaction, date, location, amount, etc.). Kim et al. (2012) described some representative features of delinquency information used in credit card and loan fraud detection. Other interesting research topics also have been studied. Based on San Diego real estate information, Gong and Huang (2012) collected 37 variables on a property's status and whether it was refinanced to predict property refinancing. Alfaro et al. (2008) and Zhou (2013) used various features as the factors to judging corporate life cycle stages or even to predict corporate bankruptcy: dividend payout ratio, sales growth, capital expenditure, the firm age and other firm profiles such as recorded liabilities, assets, sales, legal structures and tax information. Other sources of data like network behavior information, social network information and other web information were also utilized in the literature (Abeysinghe et al., 2016).

4.4. Information technology

Information technology (IT) is the application of computers to store, retrieve, transmit and manipulate data. With the explosive growth in web data, detecting interesting events from information devices and platforms is crucial to business decision-making and strategy formation. Traditionally, software defect detection, network intrusion detection and other anomaly detection methods are implemented under imbalanced scenarios. The collected articles in this domain were divided into three parts according to the research subject, as shown in Table 10.

Predicting software defects and quality were the two primary research topics in software engineering. Module attributes are commonly used to predict defects or evaluate the quality of software. On the other hand, network intrusion detection often needs to make predictions online, which leads to an online imbalanced learning problem, detailed techniques was introduced in Wang et al. (2013). With the rapid development of internet technology, web data has become an important resource to analyze customer preferences. Vorobeva (2016) used web-post to identify web author, and sentiment analysis has become a popular

Table 8
Applications in chemical and biomedical engineering fields.

Research topic	Detailed application	The literature
Diseases diagnosis	Alzheimer's disease	Dubey et al. (2014)
	Transient ischemic attack prediction	Cheng and Hu (2009)
	Lung nodule detection	Peng et al. (2014)
	Cancer diagnosis	11 papers such as Krawczyk et al. (2016), Yang et al. (2016a), Lusa (2010), Yu et al. (2012)
	Coating on tongue	Li et al. (2013b)
	Diabetes	Vo and Won (2007), Oh et al. (2011)
	Hepatitis virus detection	Oh et al. (2011), Richardson and Lidbury (2013)
	Traumatic events detection	Niehaus et al. (2014)
	Polyp detection	Bae and Yoon (2015)
	pediatric acute conditions	Wilk et al. (2016)
	Parkinson's Disease	Yeh et al. (2016)
Protein/DNA/Gene related identifications	Protein/DNA identification	7 papers such as Herndon and Caragea (2016), Song et al. (2014), Vani and Sravani (2014), Zou et al. (2016)
	Protein-protein interaction	Bao et al. (2016a), Li et al. (2009), Zhang et al. (2012)
	Gene expression prediction	Yu et al. (2012), Blagus and Lusa (2013)
	Genetic abnormalities	Lerner et al. (2007)
Monitoring and therapy	Length of stay for appendectomy patients	Cheng and Hu (2009)
	Microcalcification detection	Bria et al. (2012)
	Liveability prediction	Zięba et al. (2014), Li et al. (2015)
	Risk scoring for the prediction of acute cardiac complications	Liu et al. (2014)
	Tuberculosis early warning	Li et al. (2016b)
	Elucidate therapeutic effects in uremia patients	Chen (2016)
Chemical engineering	Small molecules detection	Hao et al. (2014)
	Drug resistance	Raposo et al. (2016)
	Identifying potential new drugs	de Souza et al. (2016)
	Biological absorption, distribution, metabolism, and excretion prediction	Casañola-Martin et al. (2016),
Biology	Animal behavioral task and chronic neural recording	Vigneron and Chen (2015)

Table 9
Applications related to financial management.

Research topic	Detail application	The literature
Fraud detection	Online banking/web payment system fraud	Zhang et al. (2008), Wei et al. (2013b), Lima and Pereira (2015)
	Credit/plastic card fraud detection	13 papers such as Zakaryazad and Duman (2016), Sanz et al. (2015), Sahin et al. (2013)
	Occupational fraud detection	Mardani and Shahriari (2013)
	Insurance fraud detection	6 papers such as Hassan and Abraham (2016), Pan et al. (2011), Kirlidog and Asuk (2012)
	Fraudulent financial statement detection(Firms)	6 papers such as Li and Wong (2015), Moepya et al. (2014), Pai et al. (2011)
	Tax declaration	Zhang et al. (2009)
Property refinance prediction	Predict whether or not a particular property would refinance	Gong and Huang (2012)
Loan default prediction	Predict whether a lender would repay in time	Abeysinghe et al. (2016), Brown and Mues (2012), Sanz et al. (2015)
Corporate bankruptcy prediction	Pre-warning of whether a corporate will fall into a decline stage or even bankruptcy	Lin et al. (2013b), Alfaro et al. (2008), Zhou (2013)
Credit card approval	Whether to approve a new applicant	Li and Wong (2015), Sanz et al. (2015)

Table 10
Information technology applications.

Research subject	Detail application	The literature
Software	Software defect prediction	Wang and Yao (2013), Rodriguez et al. (2014), Tan et al. (2015a)
	Software quality evaluation	Drown et al. (2009)
	Automated identification of high impact bug reports	Yang et al. (2016b)
Network	Network intrusion detection	8 papers such as Folino et al. (2016), Hajian et al. (2011), Engen et al. (2008)
	Peer-to-peer traffic classification	Zhong et al. (2013)
Web	Data disclosure prediction	Zhang et al. (2015c)
	Web author identification	Vorobeva (2016)
	Web service Qos prediction	Xiong et al. (2014)
	Click fraud in online advertising	Taneja et al. (2015)
	Fake website detection	Abbasi and Chen (2009)
	Sentiment analysis	6 papers such as Prusa et al. (2016), Lane et al. (2012), Zhang et al. (2015a)

topic which processes and analyzes users' preferences from User-Generated Content (UGC) on the web. Since users' opinions on a specific topic tend to be consistent, UGC data for sentiment analysis is usually imbalanced. As sentiment analysis and other web data mining models are often built from heterogeneous data sources such as text, image and numeric data, feature engineering techniques are needed to generate the features before the imbalanced learning models can be designed. For example, word embedding (Mikolov et al., 2013) is an efficient technique to build word representations for text, and Convolutional Neural Networks (CNN, or ConvNet) are popular for generating features from raw images (Razavian et al., 2014).

4.5. Energy management

Energy management includes the planning and operation of energy production and energy consumption units. There were 8 papers related to this area. Guo et al. (2016) and Li et al. (2016c) used oil well logging data to recognize the oil-bearing formations of each layer in the well. Xu et al. (2007a, b) focused on power distribution outage identification to enhance the availability and reliability of power distribution systems. Historical power distribution outage data as well as environmental attributes were utilized to extract fault patterns caused by trees, animals and lightning. A condition assessment model was built by Ashkezari et al. (2013) to evaluate the healthiness(fitness) level of power transformers. The model was trained by dissolved gas analysis and insulating oil test data. Qing (2015) focused on predicting the inventory material consumption demand of an electric power system. They found project-based demand and operational/maintenance-based demand in electric power systems have skewed frequencies; therefore, building an imbalanced learning model able to forecast possible material consumption demand can have better results than traditional machine learning models. Fraud detection techniques were also applied in the energy field. In particular, electricity fraud, which is the dishonest or illegal use of electricity equipment or service with the intention to avoid billing charges (Fabris et al., 2009; Nagi et al., 2008). Since electricity customer consumption data is made up of time series records, these attributes need to be treated in a special manner to extract all the relevant information (Fabris et al., 2009).

4.6. Security management

To implement effective controls, organizations use security management procedures such as potential risk detection, risk assessment and risk analysis to identify threats, crimes, and other anomalies. Insider threats such as infringing intellectual property and malicious sabotage are crucial events that need to be detected by a security department. Azaria et al. (2014) analyzed the behavior of insider threat. Experiments using Amazon Mechanical Turk (AMT) were conducted to distinguish normal behaviors and of those who intend to leak private data from an organization. Another common application in security management was detecting threats and unusual events from surveillance video (Mandadi and Sethi, 2013; Wang et al., 2016a; Wang et al., 2015b; Wen et al., 2015; Xu et al., 2016). Automatic event detection systems based on surveillance video typically contained feature extraction and pattern classification components (Xu et al., 2016). With regards to feature extraction, approaches such as discrete optical flow descriptors, trajectory based approaches (Xu et al., 2014) and sparse spatio-temporal (ST) corner descriptors were some of widely-used methods to extract features from local image regions and video clips (Mandadi and Sethi, 2013). For the pattern classification phase, imbalanced data classification algorithms can be used to recognize abnormal actions and events.

4.7. Electronics and communications

Five papers were related to electronics and telecommunications in our study, four of which focused on detecting telecommunications fraud. In brief, telecommunications fraud can be simply defined as any activity in which a telecommunications service is obtained without any payment intention (Hilas and Mastorocostas, 2008). Hilas and Mastorocostas (2008) examined several telecommunication fraud categories such as technical fraud, contractual fraud, hacking fraud and procedural fraud. Similarly, Farvaresh and Sepehri (2011) identified subscription fraud, dial through fraud, free phone fraud, premium rate service fraud, handset theft, and roaming fraud, and sought to detect residential and commercial subscription fraud based on call detail recording and bill data. Other studies such as Olszewski (2012) and Subudhi and Panigrahi (2015) distinguish normal and fraudulent behavior based on user profiles.

Other papers in this category were Kwak et al. (2015) and Tan et al. (2015b) both of which examined the detection of circuitry defects or other anomalies in a wafer of electrical and electronic devices. Tan et al. (2015b) trained a machine learning model on an encrypted dataset generated from three main semiconductor manufacturing operational processes (etests, sort and class test).

4.8. Infrastructure and industrial manufacturing

In this category, eight papers applied imbalanced learning methods to solve industrial manufacturing problems. Cateni et al. (2014) used resampling method for two metal industry problems. The first problem concerned the detection of defects in an automatic inspection system of the product surface during production. The other industrial application from the steel-making field was designed to identify nozzles to determine final product quality. Sun et al. (2010) focused on a detection process for variations in the nano-CMOS circuits used widely in manufacturing. The proposed method was tested on six instrument indicators and realized with a 45 nm CMOS. Liao (2008) proposed a multi-class imbalanced data classification algorithm to identify the different types of weld flaws that might be unequally distributed. In their experiments, each weld flaw sample had 12 features describing the shape, size, location, and intensity information, which were extracted from a radiographic image. The goal of their study was to classify weld flaws to identify a lack of fusion, a lack of penetration, gas holes, porosities or cracks. Tajik et al. (2015) proposed a fault detection system for industrial gas turbines. Applying imbalanced learning to machinery fault diagnoses has been studied recently: Duan et al. (2016a) and Mao et al. (2017) classified multiple faults that could occur in roller bearings; Jin et al. (2014) and Zhang et al. (2015b) constructed several features to represent different health conditions of machines in order to detect potential machinery faults.

There is also one paper related to building modeling. Xin et al. (2011) suggested that detecting building foot-points from LIDAR data was the foundation and one of the difficulties in building modeling and edge detection applications. They therefore attempted to detect building points from a non-ground points dataset, which consisted of two unbalanced datasets generated from a built-up area (with dense buildings and small tress) and a rural area (with dense trees and low houses).

4.9. Business management

Business management is a broad concept that includes planning, organizing, staffing, leading, and controlling an organization to accomplish a goal or target. As financial management was introduced in Section 4.3, only 7 papers that focused on other busi-

ness functions are discussed in this category, most of which were related to Customer Relationship Management (CRM). Data mining is an essential part of CRM to analyze large data streams and gain insight into customer behavior, needs and preferences (Lessmann and Voß, 2009).

Sultana (2012) used customer data from an insurance company to identify the potential customers that preferred to buy caravan insurance. The features of interest were socio-demographic variables derived from the customer's ZIP area code, and variables regarding the ownership of other insurance policies. Chen et al. (2012), Verbeke et al. (2012), Wu and Meng (2016), and Yi (2010) chose purchase time, purchase amount, rebates, buyer's credit rating, payment integral and demographic information as features to detect customer churning behavior. Chang and Chang (2012) applied an imbalanced learning model to monitor online auctions, in which attributes such as density of ratings, time information and historical records were used to detect significant abnormal and fraudulent behaviors. Bogina et al. (2016) leveraged the temporal features of both the session and the items clicked in a session to predict whether it ends up with a purchase.

4.10. Emergency management

To our surprise, few reviewed papers casted attentions on predicting emergency events. An emergency is a situation that poses an immediate risk to health, life, property or the environment (Anderson and Adey, 2012). Emergency events are typically rare events considering their infrequency. Due to their sudden destructiveness, predicting emergency events is a valuable yet difficult research topic. Existing research in emergency event detection has focused on natural disasters. Maalouf and Trafalis (2011), Maalouf and Siddiqi (2014), and Trafalis et al. (2014) built imbalanced learning models to forecast tornados. The tornado dataset had 83 attributes, including radar-derived velocity parameters that described aspects of the mesocyclone, monthly attributes, features that described the pre-storm environment and the predisposition of the atmosphere to explosively lift air over specific heights. Kim et al. (2016) applied their imbalanced learning model to detect several emergencies such as earthquake, fire, flood, landslide, nuclear event, and volcano using text documents collected from the Hungarian National Association of Radio Distress-Signaling and Infocommunications (RSOE, monitors extraordinary risk events that occur all over the world, 24 h per day).

4.11. Environmental management

Environmental resource management is the management of the interaction and impact of human societies on the environment. Vong et al. (2015) suggested that air pollution index forecasts were a time series problem. In their study, an online sequential learning method was used to predict the PM_{10} level (Good, Moderate, and Severe). Air pollution data collected by the Macau government meteorological center (including atmospheric pressure, temperature, mean relative humidity, wind speed, etc.) was selected as the case study. Topouzelis (2008) focused on ocean oil spills, which can seriously affecting the marine ecosystem. The amount of pollutant discharges and the associated effects on the marine environment was employed to evaluate sea water quality. Similarly, in order to monitor oil spill events, Brekke and Solberg (2008) used Synthetic Aperture Radar images (SAR) to distinguish oil spill from other natural phenomena.

Other two papers focused on predicting pollution threshold exceedances. Lu and Wang (2008) employed a cost sensitive algorithm with SVM to predict ozone threshold exceedances (pollutant day), and Tsai et al. (2009) applied cost sensitive neural network methods to forecast the ozone in an episode day. Both studies proved that cost-sensitive algorithms can effectively address imbalanced data issues and obtain a better forecast of rare samples (pollution days) in environmental applications. However, the application of re-sampling or ensembles to environmental management has not yet been researched.

4.12. Policy, social and education

Three key concepts related to public utilities were integrated in this category, but only four papers focused on social and educational issues. Márquez-Vera et al. (2013) posited that detecting student failure was an effective way to better understand why so many youths fail to complete their school studies. This was a difficult task as there are many possible factors that could affect school failure; further, a majority of students pass. Therefore, this failure was recognized as a prediction problem with high dimensional, imbalanced data. In their study, 77 attributes (e.g. socioeconomic factors; personal, social family and school factors; previous and current marks) were selected to build the imbalanced data classification model to predict whether a student would pass or fail high school. Other interesting studies include: Huang et al. (2016) analyzed video data to predict crowd counting; Ren et al. (2016b) designed a comprehensive feature engineering process to predict potential red light running in real time, features they used include occupancy time, time gap, used yellow time, vehicle passing, etc. Sensor data was generated from wearable devices by Gao et al. (2016) to monitor the occurrence of fall accidents.

4.13. Agriculture and horticulture

Agriculture and horticulture is an important area in science. However, we found only one paper in this group. D'Este et al. (2014) tackled a shellfish farm closure problem by predicting the desired shortest possible closure time based on water quality information. The dataset used was 18,692 manual water samples taken by the Tasmanian Shellfish Quality Assurance Program from 38 growing zones in Australia.

4.14. Other areas and non-specific applications

Applications that did not fit any of the 12 categories were assigned to this category. In detail, we found two papers related to astronomical research. For instance, Voigt et al. (2014) studied gamma-ray astronomy detection problem, where hadron observations are 100 to 1000 times more common than gamma events. Data from the MAGIC experiment in astronomy was collected to choose an optimal threshold for signal-background-separation. Al-Ghraibah et al. (2015) attempted to detect solar flares by predicting flare activity from quantitative measures of the solar magnetic field. Finally, Vajda and Fink (2010) and Alsulaiman et al. (2012) proposed a handwritten recognition system under an imbalanced scenario to identify verification.

5. Future research directions of imbalanced learning

In this section, we propose possible research directions of imbalanced learning based on our survey. In particular, imbalanced techniques we think still need to be considered are proposed in Section 5.1. Some application domains that imbalanced data are frequently observed but not well-studied are pointed out in Section 5.2.

5.1. At the technical level

5.1.1. Diversity within ensembles

As a good way to improve the classification performance of weak learners, ensemble based algorithms have been employed to solve many imbalanced learning tasks.

Wang and Yao (2009) maintained that the performance of an ensemble model was decided by the accuracies of the individual classifiers and the diversity between all classifiers. Diversity is the degree to which classifiers make different decisions on one problem. Diversity allows voted accuracy to be greater than that of single classifier. They demonstrated how diversity affected classification performance, especially on minority classes. Their empirical studies have shown that a larger diversity results in a better recall for the minority but harms the majority classes as it enhances the probability of classifying the examples as minority when the accuracy is not high enough. In addition, multi-classes are more flexible and beneficial when increasing diversity within an ensemble model. A similar study can be found in Błaszczyński and Lango (2016).

Several previous works have taken diversity into consideration when building the ensemble model, such as Díez-Pastor et al. (2015b), Krawczyk and Schaefer (2013),and Lin et al. (2013a), in which diversity measures or evolutionary methods were employed to prune the classifiers in an ensemble model to maintain the diversity. However, diversity problems still need to be studied carefully as most existing applications tend to first learn accurate base classifiers which are then integrated into the ensemble. Wang and Yao (2009) suggested that the status in an ensemble model with medium accuracy and medium diversity could lead to better performance, but the trade-off between accuracy and diversity remains unclear. Further, related to this, while pruning classifiers can be powerful in increasing ensemble diversity and avoiding overfitting, many base classifiers still need to be trained and evaluated before the pruning process, which is time-consuming. Building an ensemble model that integrates diversified and precise weak learners more efficiently needs to be addressed in future studies.

5.1.2. Adaptive learning

Hundreds of algorithms have been proposed to deal with imbalanced data classification problems, and they are shown to outperform others in some dimension. However, from the technical papers, we found no specific algorithm that was superior in all benchmark tests. Most proposed algorithms treated all imbalanced data consistently and handled it using a versatile algorithm. Yet, as imbalanced data has variations in the imbalanced ratio, the number of features and the number of classes, the classifier performances when learning from different types of datasets are different. This uncertainty in a learning model becomes more obvious when building ensemble models. Li et al. (2016c) suggested that using a specific ensemble classifier to tackle all kinds of imbalanced data was inefficient. The learning quality of a model can be affected by the way the training samples are selected, the options of the base classifiers and the final ensemble rules. While building a unified ensemble framework has been well-studied in the past decade, each component within an ensemble framework is often decided by the users. This raises another question as to how to adaptively choose a detailed algorithm to fit each component in an ensemble framework for different types of imbalanced data.

Apart from adaptive learning for ensemble models, other papers have studied the ways of adaptively selecting informative instances to re-sample from and learning the best sample rate automatically (Lu et al., 2016; Moreo et al., 2016; Ren et al., 2016a; Yun et al., 2016; Yeh et al., 2016; Zhang et al., 2016a). Furthermore, Krawczyk et al. (2014) tried to learn cost matrix of cost-sensitive learning from data. Noting that these are all most recent studies, which also support that adaptive learning could be another research topic for imbalanced learning.

5.1.3. Online learning for imbalanced data stream classification

The sheer volume and accessibility of data draws much enthusiasm to big data analytics; one of its challenges is to tackle and respond to streaming and fast-moving input data. Online learning, which aims to process one example at a time, thus has gained increasing attention in data mining community. First, it receives an example and then makes a prediction. If the prediction is wrong, it suffers a loss and updates its parameters (Maurya et al., 2015). Skewed class distributions can be seen in many data stream applications, such as in the fault diagnosis of control monitoring systems and intrusion detection in network and spam identification (Hoens et al., 2012). When learning data streams online, three main difficulties may arise: a). the underlying data distribution often changes considerably over time, which is referred to as concept drift (or non-stationary) learning (Ghazikhani et al., 2014). b). Online class imbalanced learning has significant difficulties because there is a lack of prior knowledge about which data classes should be regarded as the minority or the majority and the uncertainty imbalance status (Wang et al., 2014b, 2015a). c). Data sparsity problem is commonly found in data streams (Maurya et al., 2016). These encourage research into dynamically determining the class imbalance status in data streams and effectively adapting online learners to the class imbalance (Ghazikhani et al., 2013a).

Boyu Wang (2016), Dal Pozzolo et al. (2015), Ditzler and Polikar (2013), Wang et al. (2013) designed ensemble models with re-sampling for learning imbalanced data streams. However, cost-sensitive was rarely seen in the existing literature, only three models are found in Ghazikhani et al. (2013b), Maurya et al. (2016) and Wang et al. (2014a). When classifying big data streams, cost sensitive learning is computationally more efficient than data sampling techniques. We thus recommend researchers pay more attention to cost-sensitive online learning. Moreover, ELM based online learning algorithms have gained popularity, as the efficiency of ELM meets the real-time prediction requirement (Mao et al., 2017; Mirza et al. 2015a,b)." Since the requirements of quick and accurate responses for any data that may arrive at any time is increasing in the era of big data, online learning in a dynamic and imbalanced scenario may become a popular new research topic.

5.1.4. Semi-supervised learning and active learning

In some data analysis domains, massive data is cheap to collect; however, it is expensive to obtain labeled examples to train a classifier. Massive corpora with a few labeled instances (typically a minority) and abundant unlabeled instances are common in big data. Semi-supervised learning techniques have attempted to leverage the intrinsic information in unlabeled instances to improve classification models (Zhu and Goldberg, 2009); however, these techniques have assumed that the labeled instances cover all the learning classes, which often are not the case. Moreover, when there is imbalanced class distribution, extracting labeled instances from minority classes might be very costly. One way to gather more labeled examples is to ask experts or users for extensive labeling, which could lead to a specific semi-supervised learning method, called active learning. Active learning allows an expert to label new instances based on criteria to reduce the labeling effort (Frasca et al., 2013). The general idea in active learning is to estimate the value of labeling one unlabeled instance. The most valuable queries, given the goals of the classification task, are selected by the learning algorithm instead of being randomly selected as is the case in passive supervised learning (Escudeiro and Jorge, 2012). Few active learning algorithms have been proposed to address imbalanced data (Dong et al., 2016; Fu and Lee, 2013; Oh et al., 2011). More efforts are needed to investigate the selection

and utilization of informative examples when an imbalanced data distribution exists.

5.2. At a practical level

Revisiting the application distribution summarized in Section 4, two research domains closely related to management science and decision-making had adopted few imbalanced learning techniques. The first was emergency management. Four references were found that sought to predict natural disasters under imbalanced distributed data, as natural disasters are typical rare events. However, other types of emergency events, including accidents (such as forest fires), public health incidents (outbreaks of diseases such as cholera, Ebola, and malaria) and social security incidents (such as terrorist attacks) were rarely discussed in the imbalanced learning area. With the development of the Internet of Things, affluent monitoring data collected by sensors is accessible to researchers and scientists. Since large scaled multi-source and heterogeneous data can be easily collected in the big data era, it may be probable to develop feature engineering techniques to fuse multi-source data such as sensor data, text on the Internet and surveillance videos to build machine learning systems to detect other types of emergency events. Imbalanced learning techniques are crucial when designing learning models as the collected data related to emergency events may be imbalanced.

Another valuable research direction, from our point of view, is adapting imbalanced learning to security management issues. In particular, internet security management has had increased attention in recent years because of the rapid development of social networks. People tend to express their loves and hates on anything from a movie to a political strategy in social media, which also makes it possible for extremists to offend public order. Sentiment analysis and rumor detection may be powerful methods to monitor social networks and prevent the occurrence of risky events. Mining risk statements in massive user generated content is a rare event detection problem, which could be solved using imbalanced learning techniques.

6. Conclusions

In this paper, we attempted to provide a thorough review of rare event detection techniques and its applications. In particular, a data mining and a machine learning perspective was taken to view rare event detection as a class imbalanced data classification problem. We collected 527 papers that are related to imbalanced learning and rare event detection for this study. Unlike other surveys that have been published in the imbalanced learning field, we reviewed all papers from both a technical and a practical point of view.

Through our review, we also found some insights about commonly-used methods in some domains:

1) In chemical and biomedical engineering areas, re-sampling based ensemble classifiers are widely-employed. Since data that used in these areas are usually clinical data with fixed structure, feature engineering is rarely considered. However, for those high-dimensional data (such as protein data), feature selection is a popular choice.
2) A sophisticated feature engineering process is important for some management tasks such as financial management and business management. The features used for coping with a specific task are usually well-designed by the experts. Different from other domains, the goal of prediction in such field is often profit-driven instead of accuracy-driven. Therefore, cost-sensitive learning is often utilized and the cost of misclassification can be decided by experts or managers. Some widely-used classifiers in management fields are rule-based classifiers, such as decision tree and expert systems, in which fuzzy theory is often incorporated. This may due to that, apart from making wise decisions, understanding the criteria of decision-making is also essential for companies.
3) The main challenge of rare events detection in IT is the complexity of the data. Network log and unstructured data such as text and image usually need data cleaning and feature engineering processes. Besides, data streams are widely existed in IT area, which require online learning instead of traditional offline learning.

At the last of this paper, we incorporated some future suggestions from the reviewed papers with our thoughts to propose some future research directions for imbalanced learning and rare event detection, which will also be the focus of our future research projects.

Acknowledgements

This research has been supported by National Natural Science Foundation of China under Grant No.71103163, No.71573237; New Century Excellent Talents in University of China under Grant No. NCET-13-1012; Research Foundation of Humanities and Social Sciences of Ministry of Education of China No.15YJA630019; Special Funding for Basic Scientific Research of Chinese Central University under Grant No.CUG120111, CUG110411, G2012002A, CUG140604, CUG160605; Open Foundation for the Research Center of Resource Environment Economics in China University of Geosciences (Wuhan) under Grant No. H2015004B.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2016.12.035.

References

Abbasi, A., & Chen, H. (2009). A comparison of fraud cues and classification methods for fake escrow website detection. *Information Technology and Management, 10*(2-3), 83–101.
Abeysinghe, C., Li, J., & He, J. (2016). A Classifier Hub for Imbalanced Financial Data. In *Australasian Database Conference*. Springer.
Al-Ghraibah, A., Boucheron, L. E., & McAteer, R. J. (2015). A Study of Feature Selection of Magnetogram Complexity Features in an Imbalanced Solar Flare Prediction Data-set. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE.
Alfaro, E., García, N., Gámez, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. *Decision Support Systems, 45*(1), 110–122.
Ali, S., Majid, A., Javed, S. G., & Sattar, M. (2016). Can-CSC-GBE: Developing Cost-sensitive Classifier with Gentleboost Ensemble for breast cancer classification using protein amino acids and imbalanced data. *Computers in biology and medicine, 73*, 38–46.
Alibeigi, M., Hashemi, S., & Hamzeh, A. (2012). DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets. *Data & Knowledge Engineering, 81*, 67–103.
Alshomrani, S., Bawakid, A., Shim, S.-O., Fernández, A., & Herrera, F. (2015). A proposal for evolutionary fuzzy systems using feature weighting: Dealing with overlapping in imbalanced datasets. *Knowledge-Based Systems, 73*, 1–17.
Alsulaiman, F. A., Valdes, J. J., & El Saddik, A. (2012). Identity verification based on haptic handwritten signatures: Genetic programming with unbalanced data. In *Computational Intelligence for Security and Defence Applications (CISDA), 2012 IEEE Symposium on*. IEEE.
Anand, A., Pugalenthi, G., Fogel, G. B., & Suganthan, P. (2010). An approach for classification of highly imbalanced data using weighting and undersampling. *Amino acids, 39*(5), 1385–1391.
Anderson, B., & Adey, P. (2012). Governing events and life:'Emergency'in UK Civil Contingencies. *Political Geography, 31*(1), 24–33.
Ando, S. (2015). Classifying imbalanced data in distance-based feature space. *Knowledge and Information Systems*, 1–24.
Ashkezari, A. D., Ma, H., Saha, T. K., & Ekanayake, C. (2013). Application of fuzzy support vector machine for determining the health index of the insulation system of in-service power transformers. *Dielectrics and Electrical Insulation, IEEE Transactions on, 20*(3), 965–973.

Azaria, A., Richardson, A., Kraus, S., & Subrahmanian, V. (2014). Behavioral Analysis of Insider Threat: A Survey and Bootstrapped Prediction in Imbalanced Data. *Computational Social Systems, IEEE Transactions on, 1*(2), 135–155.

Bae, S.-H., & Yoon, K.-J. (2015). Polyp Detection via Imbalanced Learning and Discriminative Feature Learning. *Medical Imaging, IEEE Transactions on, 34*(11), 2379–2393.

Bagherpour, S., Nebot, À., & Mugica, F. (2016). FIR as Classifier in the Presence of Imbalanced Data. In *International Symposium on Neural Networks*. Springer.

Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2013). Cost sensitive credit card fraud detection using Bayes minimum risk. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*. IEEE.

Bao, F., Deng, Y., & Dai, Q. (2016a). ACID: association correction for imbalanced data in GWAS. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Bao, L., Juan, C., Li, J., & Zhang, Y. (2016b). Boosted Near-miss Under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets. *Neurocomputing, 172*, 198–206.

Beyan, C., & Fisher, R. (2015). Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition, 48*(5), 1653–1672.

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC bioinformatics, 14*(1), 1–16.

Błaszczyński, J., & Lango, M. (2016). Diversity Analysis on Imbalanced Data Using Neighbourhood and Roughly Balanced Bagging Ensembles. In *International Conference on Artificial Intelligence and Soft Computing*. Springer.

Bogina, V., Kuflik, T., & Mokryn, O. (2016). Learning Item Temporal Dynamics for Predicting Buying Sessions. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM.

Boyu Wang, J. P. (2016). Online Bagging and Boosting for Imbalanced Data Streams. *IEEE Transactions on Knowledge and Data Engineering, 28*(12), 3353–3366.

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A Survey of Predictive Modeling on Imbalanced Domains. *ACM Computing Surveys (CSUR), 49*(2).

Braytee, A., Liu, W., & Kennedy, P. (2016). A Cost-Sensitive Learning Strategy for Feature Extraction from Imbalanced Data. In *International Conference on Neural Information Processing*. Springer.

Brekke, C., & Solberg, A. H. (2008). Classifiers and confidence estimation for oil spill detection in ENVISAT ASAR images. *Geoscience and Remote Sensing Letters, IEEE, 5*(1), 65–69.

Bria, A., Marrocco, C., Molinara, M., & Tortorella, F. (2012). A ranking-based cascade approach for unbalanced data. In *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446–3453.

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

Cao, H., Tan, V. Y., & Pang, J. Z. (2014). A parsimonious mixture of Gaussian trees model for oversampling in imbalanced and multimodal time-series classification. *Neural Networks and Learning Systems, IEEE Transactions on, 25*(12), 2226–2239.

Cao, Y., & Wu, J. (2002). Projective ART for clustering data sets in high dimensional spaces. *Neural Networks, 15*(1), 105–120.

Casañola-Martin, G., Garrigues, T., Bermejo, M., González-Álvarez, I., Nguyen-Hai, N., Cabrera-Pérez, M. Á., et al. (2016). Exploring different strategies for imbalanced ADME data problem: case study on Caco-2 permeability modeling. *Molecular diversity, 20*(1), 93–109.

Castro, C. L., & Braga, A. P. (2013). Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. *Neural Networks and Learning Systems, IEEE Transactions on, 24*(6), 888–899.

Cateni, S., Colla, V., & Vannucci, M. (2014). A method for resampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing, 135*, 32–41.

Cerf, L., Gay, D., Selmaoui-Folcher, N., Crémilleux, B., & Boulicaut, J.-F. (2013). Parameter-free classification in multi-class imbalanced data sets. *Data & Knowledge Engineering, 87*, 109–129.

Chang, J.-S., & Chang, W.-H. (2012). A cost-effective method for early fraud detection in online auctions. In *ICT and Knowledge Engineering (ICT & Knowledge Engineering), 2012 10th International Conference on*. IEEE.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 321–357.

Chen, K., Lu, B.-L., & Kwok, J. T. (2006). Efficient classification of multi-label and imbalanced data using min-max modular classifiers. In *Neural Networks, 2006. IJCNN'06. International Joint Conference on*. IEEE.

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

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

Chen, Y.-S. (2016). An empirical study of a hybrid imbalanced-class DT-RST classification procedure to elucidate therapeutic effects in uremia patients. *Medical & biological engineering & computing, 54*(6), 983–1001.

Chen, Z.-Y., Fan, Z.-P., & Sun, M. (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of Operational Research, 223*(2), 461–472.

Cheng, F., Zhang, J., & Wen, C. (2016). Cost-Sensitive Large margin Distribution Machine for classification of imbalanced data. *Pattern Recognition Letters, 80*, 107–112.

Cheng, J., & Liu, G.-Y. (2015). Affective detection based on an imbalanced fuzzy support vector machine. *Biomedical Signal Processing and Control, 18*, 118–126.

Cheng, T.-H., & Hu, P.-H. (2009). A data-driven approach to manage the length of stay for appendectomy patients. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 39*(6), 1339–1347.

Chetchotsak, D., Pattanapairoj, S., & Arnonkijpanich, B. (2015). Integrating new data balancing technique with committee networks for imbalanced data: GRSOM approach. *Cognitive neurodynamics, 9*(6), 627–638.

D'Este, C., Timms, G., Turnbull, A., & Rahman, A. (2014). Ensemble aggregation methods for relocating models of rare events. *Engineering Applications of Artificial Intelligence, 34*, 58–65.

D'Addabbo, A., & Maglietta, R. (2015). Parallel selective sampling method for imbalanced and large data classification. *Pattern Recognition Letters, 62*, 61–67.

da Silva, I. B. V., & Adeodato, P. J. (2011). PCA and Gaussian noise in MLP neural network training improve generalization in problems with small and unbalanced data sets. In *Neural networks (IJCNN), the 2011 international joint conference on*. IEEE.

Dai, H.-L. (2015). Imbalanced Protein Data Classification Using Ensemble FTM-SVM. *NanoBioscience, IEEE Transactions on, 14*(4), 350–359.

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2015). Credit card fraud detection and concept-drift adaptation with delayed supervised information. In *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE.

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

Datta, S., & Das, S. (2015). Near-Bayesian Support Vector Machines for imbalanced data classification with equal or unequal misclassification costs. *Neural Networks, 70*, 39–52.

de Souza, J. C. S., Claudino, S. G., da Silva Simões, R., Oliveira, P. R., & Honório, K. M. (2016). Recent advances for handling imbalancement and uncertainty in labelling in medicinal chemistry data analysis. In *SAI Computing Conference (SAI), 2016*. IEEE.

del 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.

Denil, M., & Trappenberg, T. (2010). Overlap versus Imbalance. In *Canadian Conference on Advances in Artificial Intelligence*.

Díez-Pastor, J. F., Rodríguez, J. J., García-Osorio, C., & Kuncheva, L. I. (2015a). Random balance: ensembles of variable priors classifiers for imbalanced data. *Knowledge-Based Systems, 85*, 96–111.

Díez-Pastor, J. F., Rodríguez, J. J., García-Osorio, C. I., & Kuncheva, L. I. (2015b). Diversity techniques improve the performance of the best imbalance learning ensembles. *Information Sciences, 325*, 98–117.

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

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

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

Duan, L., Xie, M., Bai, T., & Wang, J. (2016a). A new support vector data description method for machinery fault diagnosis with unbalanced datasets. *Expert Systems with Applications, 64*, 239–246.

Duan, L., Xie, M., Bai, T., & Wang, J. (2016b). Support vector data description for machinery multi-fault classification with unbalanced datasets. In *Prognostics and Health Management (ICPHM), 2016 IEEE International Conference on*. IEEE.

Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., & Ye, J.A. s. D. N. Initiative. (2014). Analysis of sampling techniques for imbalanced data: An n= 648 ADNI study. *NeuroImage, 87*, 220–241.

Engen, V., Vincent, J., & Phalp, K. (2008). Enhancing network based intrusion detection for imbalanced data. *International Journal of Knowledge-Based and Intelligent Engineering Systems, 12*(5, 6), 357–367.

Escudeiro, N. F., & Jorge, A. M. (2012). D-Confidence: an active learning strategy to reduce label disclosure complexity in the presence of imbalanced class distributions. *Journal of the Brazilian Computer Society, 18*(4), 311–330.

Fabris, F., Margoto, L. R., & Varejao, F. M. (2009). Novel approaches for detecting frauds in energy consumption. In *Network and System Security, 2009. NSS'09. Third International Conference on*. IEEE.

Fahimnia, B., Tang, C. S., Davarzani, H., & Sarkis, J. (2015). Quantitative models for managing supply chain risks: A review. *European Journal of Operational Research, 247*(1), 1–15.

Fan, J., Niu, Z., Liang, Y., & Zhao, Z. (2016). Probability Model Selection and Parameter Evolutionary Estimation for Clustering Imbalanced Data without Sampling. *Neurocomputing*.

Farvaresh, H., & Sepehri, M. M. (2011). A data mining framework for detecting subscription fraud in telecommunication. *Engineering Applications of Artificial Intelligence, 24*(1), 182–194.

Fernández, A., Del Jesus, M. J., & Herrera, F. (2010a). Multi-class imbalanced data-sets with linguistic fuzzy rule based classification systems based on pairwise learning. *Computational Intelligence for Knowledge-Based Systems Design* (pp. 89–98). Springer.

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

Fernández, A., 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.

Ferri, C., Hernández-Orallo, J., & Flach, P. A. (2011). A coherent interpretation of AUC as a measure of aggregated classification performance. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*.

Folino, G., Pisani, F. S., & Sabatino, P. (2016). An Incremental Ensemble Evolved by using Genetic Programming to Efficiently Detect Drifts in Cyber Security Datasets. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*. ACM.

Frasca, M., Bertoni, A., Re, M., & Valentini, G. (2013). A neural network algorithm for semi-supervised node label learning from unbalanced data. *Neural Networks, 43*, 84–98.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *ICML*.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences, 55*(1), 119–139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Fu, J., & Lee, S. (2013). Certainty-based active learning for sampling imbalanced datasets. *Neurocomputing, 119*, 350–358.

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

Galar, M., Fernández, A., Barrenechea, E., & Herrera, F. (2013). EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition, 46*(12), 3460–3471.

Gao, M., Hong, X., & Harris, C. J. (2014). Construction of neurofuzzy models for imbalanced data classification. *Fuzzy Systems, IEEE Transactions on, 22*(6), 1472–1488.

Gao, X., Chen, Z., Tang, S., Zhang, Y., & Li, J. (2016). Adaptive weighted imbalance learning with application to abnormal activity recognition. *Neurocomputing, 173*, 1927–1935.

García, V., Sánchez, J. S., Martín-Félez, R., & Mollineda, R. A. (2012). Surrounding neighborhood-based SMOTE for learning from imbalanced data sets. *Progress in Artificial Intelligence, 1*(4), 347–362.

Garcia-Pedrajas, N., del Castillo, J. A. R., & Cerruela-Garcia, G. (2015). A Proposal for Local k Values for k-Nearest Neighbor Rule. *IEEE transactions on neural networks and learning systems*.

García-Pedrajas, N., & García-Osorio, C. (2013). Boosting for class-imbalanced datasets using genetically evolved supervised non-linear projections. *Progress in Artificial Intelligence, 2*(1), 29–44.

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

Ghazikhani, A., Monsefi, R., & Yazdi, H. S. (2013b). Online cost-sensitive neural network classifiers for non-stationary and imbalanced data streams. *Neural Computing and Applications, 23*(5), 1283–1295.

Ghazikhani, A., Monsefi, R., & Yazdi, H. S. (2014). Online neural network model for non-stationary and imbalanced data stream classification. *International Journal of Machine Learning and Cybernetics, 5*(1), 51–62.

Gong, R., & Huang, S. H. (2012). A Kolmogorov–Smirnov statistic based segmentation approach to learning from imbalanced datasets: With application in property refinance prediction. *Expert Systems with Applications, 39*(6), 6192–6200.

Govindan, K., & Jepsen, M. B. (2016). ELECTRE: A comprehensive literature review on methodologies and applications. *European Journal of Operational Research, 250*(1), 1–29.

Gu, Q., Zhu, L., & Cai, Z. (2009). Evaluation measures of the classification performance of imbalanced data sets. *International Symposium on Intelligence Computation and Applications*. Springer.

Guo, H., Li, Y., Yanan, L., Xiao, L., & Jinling, L. (2016). BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification. *Engineering Applications of Artificial Intelligence, 49*, 176–193.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research, 3*, 1157–1182.

Ha, J., & Lee, J.-S. (2016). A New Under-Sampling Method Using Genetic Algorithm for Imbalanced Data Classification. In *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*. ACM.

Hajian, S., Domingo-Ferrer, J., & Martinez-Balleste, A. (2011). Discrimination prevention in data mining for intrusion and crime detection. In *Computational Intelligence in Cyber Security (CICS), 2011 IEEE Symposium on*. IEEE.

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

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning, 45*(2), 171–186.

Hao, M., Wang, Y., & Bryant, S. H. (2014). An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data. *Analytica chimica acta, 806*, 117–127.

Hartmann, W. M. (2004). Dimension reduction vs. variable selection. Applied Parallel Computing. *State of the Art in Scientific Computing* (pp. 931–938). Springer.

Hassan, A. K. I., & Abraham, A. (2016). Modeling insurance fraud detection using imbalanced data classification. *Advances in Nature and Biologically Inspired Computing* (pp. 117–127). Springer.

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

He, H. and Y. Ma (2013). "Imbalanced learning. Foundations, algorithms, and applications."

Herndon, N., & Caragea, D. (2016). A Study of Domain Adaptation Classifiers Derived From Logistic Regression for the Task of Splice Site Prediction. *IEEE transactions on nanobioscience, 15*(2), 75–83.

Hilas, C. S., & Mastorocostas, P. A. (2008). An application of supervised and unsupervised learning approaches to telecommunications fraud detection. *Knowledge-Based Systems, 21*(7), 721–726.

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

Hong, X., Chen, S., & Harris, C. J. (2007). A kernel-based two-class classifier for imbalanced data sets. *Neural Networks, IEEE Transactions on, 18*(1), 28–41.

Hu, S., Liang, Y., Ma, L., & He, Y. (2009). MSMOTE: improving classification performance when training data is imbalanced. In *2009 Second International Workshop on Computer Science and Engineering*. IEEE.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006a). Extreme learning machine: theory and applications. *Neurocomputing, 70*(1), 489–501.

Huang, K., Yang, H., King, I., & Lyu, M. R. (2006b). Imbalanced learning with a biased minimax probability machine. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 36*(4), 913–923.

Huang, X., Zou, Y., & Wang, Y. (2016). Cost-sensitive sparse linear regression for crowd counting with imbalanced training data. In *Multimedia and Expo (ICME), 2016 IEEE International Conference on*. IEEE.

Jacques, J., Taillard, J., Delerue, D., Dhaenens, C., & Jourdan, L. (2015). Conception of a dominance-based multi-objective local search in the context of classification rule mining in large and imbalanced data sets. *Applied Soft Computing, 34*, 705–720.

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing Imbalanced Data–Recommendations for the Use of Performance Metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE.

Jian, C., Gao, J., & Ao, Y. (2016). A new sampling method for classifying imbalanced data based on support vector machine ensemble. *Neurocomputing, 193*, 115–122.

Jin, X., Yuan, F., Chow, T. W., & Zhao, M. (2014). Weighted local and global regressive mapping: A new manifold learning method for machine fault classification. *Engineering Applications of Artificial Intelligence, 30*, 118–128.

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

Kim, J., Choi, K., Kim, G., & Suh, Y. (2012). Classification cost: An empirical comparison among traditional classifier, Cost-Sensitive Classifier, and MetaCost. *Expert Systems with Applications, 39*(4), 4013–4019.

Kim, S., Kim, H., & Namkoong, Y. (2016). Ordinal Classification of Imbalanced Data with Application in Emergency and Disaster Information Services. *IEEE Intelligent Systems, 31*(5), 50–56.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis, 9*(2), 137–163.

Kirlidog, M., & Asuk, C. (2012). A fraud detection approach with data mining in health insurance. *Procedia-Social and Behavioral Sciences, 62*, 989–994.

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

Krawczyk, B., & Schaefer, G. (2013). An improved ensemble approach for imbalanced classification problems. In *Applied Computational Intelligence and Informatics (SACI), 2013 IEEE 8th International Symposium on*. IEEE.

Krawczyk, B., Woźniak, M., & Schaefer, G. (2014). Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing, 14*, 554–562.

Krivko, M. (2010). A hybrid model for plastic card fraud detection systems. *Expert Systems with Applications, 37*(8), 6070–6076.

Kumar, N. S., Rao, K. N., Govardhan, A., Reddy, K. S., & Mahmood, A. M. (2014). Undersampled K-means approach for handling imbalanced distributed data. *Progress in Artificial Intelligence, 3*(1), 29–38.

Kwak, J., Lee, T., & Kim, C. O. (2015). An Incremental Clustering-Based Fault Detection Algorithm for Class-Imbalanced Process Data. *Semiconductor Manufacturing, IEEE Transactions on, 28*(3), 318–328.

Lan, J.-s., Berardi, V. L., Patuwo, B. E., & Hu, M. (2009). A joint investigation of misclassification treatments and imbalanced datasets on neural network performance. *Neural Computing and Applications, 18*(7), 689–706.

Lane, P. C., Clarke, D., & Hender, P. (2012). On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data. *Decision Support Systems, 53*(4), 712–718.

Lerner, B., Yeshaya, J., & Koushnir, L. (2007). On the classification of a small imbalanced cytogenetic image database. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 4*(2), 204–215.

Lessmann, S., & Voß, S. (2009). A reference model for customer-centric data mining with support vector machines. *European Journal of Operational Research, 199*(2), 520–530.

Li, H., & Wong, M.-L. (2015). Financial fraud detection by using Grammar-based multi-objective genetic programming with ensemble learning. In *Evolutionary Computation (CEC), 2015 IEEE Congress on*. IEEE.

Li, J., Fong, S., Mohammed, S., & Fiaidhi, J. (2015). Improving the classification performance of biological imbalanced datasets by swarm optimization algorithms. *The Journal of Supercomputing*, 1–21.

Li, J., Liu, L.-s., Fong, S., Wong, R. K., Mohammed, S., Fiaidhi, J., et al. (2016a). Adaptive Swarm Balancing Algorithms for rare-event prediction in imbalanced healthcare data. *Computerized Medical Imaging and Graphics*.

Li, K., Kong, X., Lu, Z., Wenyin, L., & Yin, J. (2014). Boosting weighted ELM for imbalanced learning. *Neurocomputing, 128*, 15–21.

Li, L., Jing, L., & Huang, D. (2009). Protein-protein interaction extraction from biomedical literatures based on modified SVM-KNN. In *Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on*. IEEE.

Li, Q., Yang, B., Li, Y., Deng, N., & Jing, L. (2013a). Constructing support vector machine ensemble with segmentation for imbalanced datasets. *Neural Computing and Applications, 22*(1), 249–256.

Li, S., Tang, B., & He, H. (2016b). An Imbalanced Learning based MDR-TB Early Warning System. *Journal of medical systems, 40*(7), 1–9.

Li, X., Shao, Q., & Wang, J. (2013b). Classification of tongue coating using Gabor and Tamura features on unbalanced data set. In *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*. IEEE.

Li, Y., Guo, H., Xiao, L., Yanan, L., & Jinling, L. (2016c). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems, 94*, 88–104.

Liang, J., Bai, L., Dang, C., & Cao, F. (2012). The-Means-Type Algorithms Versus Imbalanced Data Distributions. *Fuzzy Systems, IEEE Transactions on, 20*(4), 728–745.

Liao, T. W. (2008). Classification of weld flaws with imbalanced class data. *Expert Systems with Applications, 35*(3), 1041–1052.

Lima, R. F., & Pereira, A. C. (2015). A Fraud Detection Model Based on Feature Selection and Undersampling Applied to Web Payment Systems. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE.

Lin, M., Tang, K., & Yao, X. (2013a). Dynamic sampling approach to training neural networks for multiclass imbalance classification. *Neural Networks and Learning Systems, IEEE Transactions on, 24*(4), 647–660.

Lin, S.-J., Chang, C., & Hsu, M.-F. (2013b). Multiple extreme learning machines for a two-class imbalance corporate life cycle prediction. *Knowledge-Based Systems, 39*, 214–223.

Liu, N., Koh, Z. X., Chua, E. C.-P., Tan, L. M.-L., Lin, Z., Mirza, B., et al. (2014). Risk scoring for prediction of acute cardiac complications from imbalanced clinical data. *Biomedical and Health Informatics, IEEE Journal of, 18*(6), 1894–1902.

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

López, V., del Río, S., Benítez, J. M., & Herrera, F. (2015). Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. *Fuzzy Sets and Systems, 258*, 5–38.

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.

López, V., Fernández, 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.

Loyola-González, O., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., & García-Borroto, M. (2016). Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing, 175*, 935–947.

Lu, J., Zhang, C., & Shi, F. (2016). A Classification Method of Imbalanced Data Base on PSO Algorithm. In *International Conference of Young Computer Scientists, Engineers and Educators*. Springer.

Lu, W. Z., & Wang, D. (2008). Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme. *Science of the Total Environment, 395*(2-3), 109–116.

Lusa, L. (2010). Class prediction for high-dimensional class-imbalanced data. *BMC bioinformatics, 11*(1), 523.

Lusa, L. (2016). Gradient boosting for high-dimensional prediction of rare events. *Computational Statistics & Data Analysis*.

Maalouf, M., & Siddiqi, M. (2014). Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems, 59*, 142–148.

Maalouf, M., & Trafalis, T. B. (2011). Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics & Data Analysis, 55*(1), 168–183.

Maldonado, S., & López, J. (2014). Imbalanced data classification using second-order cone programming support vector machines. *Pattern Recognition, 47*(5), 2070–2079.

Maldonado, S., Weber, R., & Famili, F. (2014). Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines. *Information Sciences, 286*, 228–246.

Mandadi, B., & Sethi, A. (2013). Unusual event detection using sparse spatio-temporal features and bag of words model. In *Image Information Processing (ICIIP), 2013 IEEE Second International Conference on*. IEEE.

Mao, W., He, L., Yan, Y., & Wang, J. (2017). Online sequential prediction of bearings imbalanced fault diagnosis by extreme learning machine. *Mechanical Systems and Signal Processing, 83*, 450–473.

Mao, W., Wang, J., He, L., & Tian, Y. (2016). Two-Stage Hybrid Extreme Learning Machine for Sequential Imbalanced Data. In *Proceedings of ELM-2015: Volume 1* (pp. 423–433). Springer.

Maratea, A., Petrosino, A., & Manzo, M. (2014). Adjusted F-measure and kernel scaling for imbalanced data learning. *Information Sciences, 257*, 331–341.

Mardani, S., & Shahriari, H. R. (2013). A new method for occupational fraud detection in process aware information systems. In *Information Security and Cryptology (ISCISC), 2013 10th International ISC Conference on*. IEEE.

Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied intelligence, 38*(3), 315–330.

Maurya, C. K., Toshniwal, D., & Venkoparao, G. V. (2015). Online anomaly detection via class-imbalance learning. In *Contemporary Computing (IC3), 2015 Eighth International Conference on*. IEEE.

Maurya, C. K., Toshniwal, D., & Venkoparao, G. V. (2016). Online sparse class imbalance learning on big data. *Neurocomputing*.

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

Mikolov, T., K. Chen, G. Corrado and J. Dean (2013). "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781.

Mirza, B., Lin, Z., Cao, J., & Lai, X. (2015a). Voting based weighted online sequential extreme learning machine for imbalance multi-class classification. In *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE.

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

Mirza, B., Lin, Z., & Toh, K.-A. (2013). Weighted online sequential extreme learning machine for class imbalance learning. *Neural processing letters, 38*(3), 465–486.

Moepya, S. O., Akhoury, S. S., & Nelwamondo, F. V. (2014). Applying Cost-Sensitive Classification for Financial Fraud Detection under High Class-Imbalance. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*. IEEE.

Moreo, A., Esuli, A., & Sebastiani, F. (2016). Distributional Random Oversampling for Imbalanced Text Classification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM.

Motoda, H., & Liu, H. (2002). Feature selection, extraction and construction, *Communication of IICM: Vol 5* (pp. 67–72). Taiwan: Institute of Information and Computing Machinery.

Nagi, J., Yap, K., Tiong, S., Ahmed, S., & Mohammad, A. (2008). Detection of abnormalities and electricity theft using genetic support vector machines. In *TENCON 2008-2008 IEEE Region 10 Conference*. IEEE.

Napierala, K., & Stefanowski, J. (2015). Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, 1–35.

Napierała, K., & Stefanowski, J. (2015). Addressing imbalanced data with argument based rule learning. *Expert Systems with Applications, 42*(24), 9468–9481.

Natwichai, J., Li, X., & Orlowska, M. (2005). Hiding classification rules for data sharing with privacy preservation. *Data Warehousing and Knowledge Discovery* (pp. 468–477). Springer.

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

Ng, W. W., Zeng, G., Zhang, J., Yeung, D. S., & Pedrycz, W. (2016). Dual autoencoders features for imbalance classification problem. *Pattern Recognition, 60*, 875–889.

Niehaus, K. E., Clark, I. A., Bourne, C., Mackay, C. E., Holmes, E. A., Smith, S. M., et al. (2014). MVPA to enhance the study of rare cognitive events: An investigation of experimental PTSD. In *Pattern Recognition in Neuroimaging, 2014 International Workshop on*. IEEE.

Oh, S., Lee, M. S., & Zhang, B.-T. (2011). Ensemble learning with active example selection for imbalanced biomedical data classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 8*(2), 316–325.

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

Olszewski, D. (2012). A probabilistic approach to fraud detection in telecommunications. *Knowledge-Based Systems, 26*, 246–258.

Pai, P.-F., Hsu, M.-F., & Wang, M.-C. (2011). A support vector machine-based model for detecting top management fraud. *Knowledge-Based Systems, 24*(2), 314–321.

Pan, J., Fan, Q., Pankanti, S., Trinh, H., Gabbur, P., & Miyazawa, S. (2011). Soft margin keyframe comparison: Enhancing precision of fraud detection in retail surveillance. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*. IEEE.

Panigrahi, S., Kundu, A., Sural, S., & Majumdar, A. K. (2009). Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. *Information Fusion, 10*(4), 354–363.

Park, Y., & Ghosh, J. (2014). Ensembles of $({alpha}) $-Trees for Imbalanced Classification Problems. *Knowledge and Data Engineering, IEEE Transactions on, 26*(1), 131–143.

Peng, Yang, J., Li, W., Zhao, D., & Zaiane, O. (2014). Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD. *Computerized Medical Imaging and Graphics, 38*(3), 137–150.

Pérez-Godoy, M. D., Fernández, A., Rivera, A. J., & del Jesus, M. J. (2010). Analysis of an evolutionary RBFN design algorithm, CO 2 RBFN, for imbalanced data sets. *Pattern Recognition Letters, 31*(15), 2375–2388.

Phoungphol, P., Zhang, Y., & Zhao, Y. (2012). Robust multiclass classification for learning from imbalanced biomedical data. *Tsinghua Science and technology, 17*(6), 619–628.

Prusa, J. D., Khoshgoftaar, T. M., & Seliya, N. (2016). Enhancing Ensemble Learners with Data Sampling on High-Dimensional Imbalanced Tweet Sentiment Data. In *The Twenty-Ninth International Flairs Conference*.

Raj, V., Magg, S., & Wermter, S. (2016). Towards effective classification of imbalanced data with convolutional neural networks. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer.

Ramentol, E., Vluymans, S., Verbiest, N., Caballero, Y., Bello, R., Cornelis, C., et al. (2015). IFROWANN: imbalanced fuzzy-rough ordered weighted average nearest neighbor classification. *Fuzzy Systems, IEEE Transactions on, 23*(5), 1622–1637.

Raposo, L. M., Arruda, M. B., de Brindeiro, R. M., & Nobre, F. F. (2016). Lopinavir Resistance Classification with Imbalanced Data Using Probabilistic Neural Networks. *Journal of medical systems, 40*(3), 1–7.

Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Ren, F., Cao, P., Li, W., Zhao, D., & Zaiane, O. (2016a). Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm. *Computerized Medical Imaging and Graphics*.

Ren, Y., Wang, Y., Wu, X., Yu, G., & Ding, C. (2016b). Influential factors of red-light running at signalized intersection and prediction using a rare events logistic regression model. *Accident Analysis & Prevention, 95*, 266–273.

Richardson, A. M., & Lidbury, B. A. (2013). Infection status outcome, machine learning method and virus type interact to affect the optimised prediction of hepatitis virus immunoassay results from routine pathology laboratory assays in unbalanced data. *BMC bioinformatics, 14*(1), 1.

Rodriguez, D., Herraiz, I., Harrison, R., Dolado, J., & Riquelme, J. C. (2014). Preliminary comparison of techniques for dealing with imbalance in software defect prediction. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. ACM.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics, 23*(19), 2507–2517.

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.

Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications, 40*(15), 5916–5923.

Sanz, J. A., Bernardo, D., Herrera, F., Bustince, H., & Hagras, H. (2015). A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data. *Fuzzy Systems, IEEE Transactions on, 23*(4), 973–990.

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine learning, 37*(3), 297–336.

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

Shao, Y.-H., Chen, W.-J., Zhang, J.-J., Wang, Z., & Deng, N.-Y. (2014). An efficient weighted Lagrangian twin support vector machine for imbalanced data classification. *Pattern Recognition, 47*(9), 3158–3167.

Song, J., Huang, X., Qin, S., & Song, Q. (2016). A bi-directional sampling based on K-means method for imbalance text classification. In *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on*. IEEE.

Song, L., Li, D., Zeng, X., Wu, Y., Guo, L., & Zou, Q. (2014). nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. *BMC bioinformatics, 15*(1), 1.

Su, C.-T., & Hsiao, Y.-H. (2007). An evaluation of the robustness of MTS for imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 19*(10), 1321.

Subudhi, S., & Panigrahi, S. (2015). Quarter-Sphere Support Vector Machine for Fraud Detection in Mobile Telecommunication Networks. *Procedia Computer Science, 48*, 353–359.

Sultana, M. A. H. F. N. S. R. A. J. (2012). Enhancing the performance of decision tree: A research study of dealing with unbalanced data. In *Digital Information Management (ICDIM), 2012 Seventh International Conference on*. Macau.

Sun, L., Mathew, J., Dhiraj, K., & Saraju, P. (2010). Algorithms for rare event analysis in nano-CMOS circuits using statistical blockade. In *SoC Design Conference (ISOCC), 2010 International*. IEEE.

Sun, Y., Kamel, M. S., & Wang, Y. (2006). Boosting for learning multiple classes with imbalanced class distribution. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE.

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

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

Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., & Zhou, Y. (2015). A novel ensemble method for classifying imbalanced data. *Pattern Recognition, 48*(5), 1623–1637.

Tahir, M. A., Kittler, J., Mikolajczyk, K., & Yan, F. (2009). A multiple expert approach to the class imbalance problem using inverse random under sampling. *Multiple Classifier Systems* (pp. 82–91). Springer.

Tajik, M., Movasagh, S., Shoorehdeli, M. A., & Yousefi, I. (2015). Gas turbine shaft unbalance fault detection by using vibration data and neural networks. In *Robotics and Mechatronics (ICROM), 2015 3rd RSI International Conference on*. IEEE.

Tan, M., Tan, L., Dara, S., & Mayeux, C. (2015a). Online defect prediction for imbalanced data. In *Proceedings of the 37th International Conference on Software Engineering-: Volume 2*. IEEE Press.

Tan, S. C., Watada, J., Ibrahim, Z., & Khalid, M. (2015b). Evolutionary fuzzy ARTMAP neural networks for classification of semiconductor defects. *Neural Networks and Learning Systems, IEEE Transactions on, 26*(5), 933–950.

Taneja, M., Garg, K., Purwar, A., & Sharma, S. (2015). Prediction of click frauds in mobile advertising. In *Contemporary Computing (IC3), 2015 Eighth International Conference on*. IEEE.

Tian, J., Gu, H., & Liu, W. (2011). Imbalanced classification using support vector machine ensemble. *Neural Computing and Applications, 20*(2), 203–209.

Tomek, I. (1976). A generalization of the k-NN rule. *Systems, Man and Cybernetics, IEEE Transactions on*, (2), 121–126.

Topouzelis, K. N. (2008). Oil spill detection by SAR images: dark formation detection, feature extraction and classification algorithms. *Sensors, 8*(10), 6642–6659.

Trafalis, T. B., Adrianto, I., Richman, M. B., & Lakshmivarahan, S. (2014). Machine-learning classifiers for imbalanced tornado data. *Computational Management Science, 11*(4), 403–418.

Tsai, C. H., Chang, L. C., & Chiang, H. C. (2009). Forecasting of ozone episode days by cost-sensitive neural network methods. *Science of the Total Environment, 407*(6), 2124–2135.

Vajda, S., & Fink, G. A. (2010). Strategies for training robust neural network based digit recognizers on unbalanced data sets. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*. IEEE.

Vani, K. S., & Sravani, T. (2014). Multiclass unbalanced protein data classification using sequence features. In *Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on*. IEEE.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research, 218*(1), 211–229.

Vigneron, V., & Chen, H. (2015). A multi-scale seriation algorithm for clustering sparse imbalanced data: application to spike sorting. *Pattern Analysis and Applications*, 1–19.

Vluymans, S., Tarragó, D. S., Saeys, Y., Cornelis, C., & Herrera, F. (2015). Fuzzy rough classifiers for class imbalanced multi-instance data. *Pattern Recognition*.

Vo, N. H., & Won, Y. (2007). Classification of unbalanced medical data with weighted regularized least squares. In *Frontiers in the Convergence of Bioscience and Information Technologies, 2007. FBIT 2007*. IEEE.

Voigt, T., Fried, R., Backes, M., & Rhode, W. (2014). Threshold optimization for classification in imbalanced data in a problem of gamma-ray astronomy. *Advances in Data Analysis and Classification, 8*(2), 195–216.

Vong, C.-M., Ip, W.-F., Chiu, C.-C., & Wong, P.-K. (2015). Imbalanced Learning for Air Pollution by Meta-Cognitive Online Sequential Extreme Learning Machine. *Cognitive Computation, 7*(3), 381–391.

Vorobeva, A. A. (2016). Examining the performance of classification algorithms for imbalanced data sets in web author identification. In *Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT), 2016 18th Conference of*. FRUCT.

Wan, X., Liu, J., Cheung, W. K., & Tong, T. (2014). Learning to improve medical decision making from imbalanced data without a priori cost. *BMC medical informatics and decision making, 14*(1), 1.

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

Wang, J., Zhao, P., & Hoi, S. C. (2014a). Cost-sensitive online classification. *IEEE Transactions on Knowledge and Data Engineering, 26*(10), 2425–2438.

Wang, S., Chen, H., & Yao, X. (2010). Negative correlation learning for classification ensembles. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*. IEEE.

Wang, S., Minku, L. L., & Yao, X. (2013). A learning framework for online class imbalance learning. In *Computational Intelligence and Ensemble Learning (CIEL), 2013 IEEE Symposium on*. IEEE.

Wang, S., Minku, L. L., & Yao, X. (2014b). A multi-objective ensemble method for online class imbalance learning. In *Neural Networks (IJCNN), 2014 International Joint Conference on*. IEEE.

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

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

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

Wang, S., & Yao, X. (2013). Using class imbalance learning for software defect prediction. *Reliability, IEEE Transactions on, 62*(2), 434–443.

Wang, Y., Li, X., & Ding, X. (2016a). Probabilistic framework of visual anomaly detection for unbalanced data. *Neurocomputing*.

Wang, Y., Tian, Y., Su, L., Fang, X., Xia, Z., & Huang, T. (2015b). Detecting Rare Actions and Events from Surveillance Big Data with Bag of Dynamic Trajectories. In *Multimedia Big Data (BigMM), 2015 IEEE International Conference on*. IEEE.

Wang, Z., Xin, J., Tian, S., & Yu, G. (2016b). Distributed Weighted Extreme Learning Machine for Big Imbalanced Data Learning. In *Proceedings of ELM-2015: Volume 1* (pp. 319–332). Springer.

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

Wei, M.-H., Cheng, C.-H., Huang, C.-S., & Chiang, P.-C. (2013a). Discovering medical quality of total hip arthroplasty by rough set classifier with imbalanced class. *Quality & Quantity, 47*(3), 1761–1779.

Wei, W., Li, J., Cao, L., Ou, Y., & Chen, J. (2013b). Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web, 16*(4), 449–475.

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter, 6*(1), 7–19.
Weiss, G. M., & Hirsh, H. (2000). Learning to predict extremely rare events. In *AAAI workshop on learning from imbalanced data sets*.
Wen, H., Ge, S., Chen, S., Wang, H., & Sun, L. (2015). Abnormal event detection via adaptive cascade dictionary learning. In *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE.
Wilk, S., Stefanowski, J., Wojciechowski, S., Farion, K. J., & Michalowski, W. (2016). Application of Preprocessing Methods to Imbalanced Clinical Data: An Experimental Study. *Information Technologies in Medicine* (pp. 503–515). Springer.
Wu, D., Wang, Z., Chen, Y., & Zhao, H. (2016). Mixed-kernel based weighted extreme learning machine for inertial sensor based human activity recognition with imbalanced dataset. *Neurocomputing, 190*, 35–49.
Wu, X., & Meng, S. (2016). E-commerce customer churn prediction based on improved SMOTE and AdaBoost. In *Service Systems and Service Management (ICSSSM), 2016 13th International Conference on*. IEEE.
Xiao, W., Zhang, J., Li, Y., & Yang, W. (2016). Imbalanced Extreme Learning Machine for Classification with Imbalanced Data Distributions. In *Proceedings of ELM-2015: Volume 2* (pp. 503–514). Springer.
Xin, W., Yi-ping, L., Ting, J., Hui, G., Sheng, L., & Xiao-wei, Z. (2011). A new classification method for LIDAR data based on unbalanced support vector machine. In *Image and Data Fusion (ISIDF), 2011 International Symposium on*. IEEE.
Xiong, W., Li, B., He, L., Chen, M., & Chen, J. (2014). Collaborative web service QoS prediction on unbalanced data distribution. In *Web Services (ICWS), 2014 IEEE International Conference on*. IEEE.
Xu, J., Denman, S., Fookes, C., & Sridharan, S. (2016). Detecting rare events using Kullback–Leibler divergence: A weakly supervised approach. *Expert Systems with Applications, 54*, 13–28.
Xu, J., Denman, S., Reddy, V., Fookes, C., & Sridharan, S. (2014). Real-time video event detection in crowded scenes using MPEG derived features: A multiple instance learning approach. *Pattern Recognition Letters, 44*, 113–125.
Xu, L., Chow, M.-Y., & Taylor, L. S. (2007a). Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification e-algorithm. *Power Systems, IEEE Transactions on, 22*(1), 164–171.
Xu, L., Chow, M.-Y., Timmis, J., & Taylor, L. S. (2007b). Power distribution outage cause identification with imbalanced data using artificial immune recognition system (AIRS) algorithm. *Power Systems, IEEE Transactions on, 22*(1), 198–204.
Xu, Y., Yang, Z., Zhang, Y., Pan, X., & Wang, L. (2015). A maximum margin and minimum volume hyper-spheres machine with pinball loss for imbalanced data classification. *Knowledge-Based Systems*.
Qing, Y., B. , W., Peilan, Z., Xiang, C., Meng, Z., & Yang, W. (2015). The prediction method of material consumption for electric power production based on PC-Boost and SVM. In *2015 8th International Congress on Image and Signal Processing (CISP)* (pp. 1256–1260).
Yang, J., Zhou, J., Zhu, Z., Ma, X., & Ji, Z. (2016a). Iterative ensemble feature selection for multiclass classification of imbalanced microarray data. *Journal of Biological Research-Thessaloniki, 23*(1), 13.
Yang, P., Xu, L., Zhou, B. B., Zhang, Z., & Zomaya, A. Y. (2009). A particle swarm based hybrid system for imbalanced medical data sampling. *BMC genomics, 10*(3), 1.
Yang, X., Lo, D., Huang, Q., Xia, X., & Sun, J. (2016b). Automated Identification of High Impact Bug Reports Leveraging Imbalanced Learning Strategies. In *Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual*. IEEE.
Yeh, C.-W., Li, D.-C., Lin, L.-S., & Tsai, T.-I. (2016). A Learning Approach with Under--and Over-Sampling for Imbalanced Data Sets. In *Advanced Applied Informatics (IIAI-AAI), 2016 5th IIAI International Congress on*. IEEE.
Yi, W. (2010). The Cascade Decision-Tree Improvement Algorithm Based on Unbalanced Data Set. In *2010 International Conference on Communications and Mobile Computing*. IEEE.
Yu, H., Mu, C., Sun, C., Yang, W., Yang, X., & Zuo, X. (2015). Support vector machine-based optimized decision threshold adjustment strategy for classifying imbalanced data. *Knowledge-Based Systems, 76*, 67–78.
Yu, H., Ni, J., Dan, Y., & Xu, S. (2012). Mining and integrating reliable decision rules for imbalanced cancer gene expression data sets. *Tsinghua Science and technology, 17*(6), 666–673.
Yu, H., Sun, C., Yang, X., Yang, W., Shen, J., & Qi, Y. (2016). ODOC-ELM: Optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data. *Knowledge-Based Systems, 92*, 55–70.
Yun, J., Ha, J., & Lee, J.-S. (2016). Automatic Determination of Neighborhood Size in SMOTE. In *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*. ACM.
Zakaryazad, A., & Duman, E. (2016). A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing. *Neurocomputing, 175*, 121–131.
Zhai, J., Zhang, S., & Wang, C. (2015). The classification of imbalanced large data sets based on MapReduce and ensemble of ELM classifiers. *International Journal of Machine Learning and Cybernetics*, 1–9.
Zhang, B., Zhou, Y., & Faloutsos, C. (2008). Toward a comprehensive model in internet auction fraud detection. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*. IEEE.
Zhang, C., Gao, W., Song, J., & Jiang, J. (2016a). An imbalanced data classification algorithm of improved autoencoder neural network. In *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*. IEEE.
Zhang, D., Ma, J., Yi, J., Niu, X., & Xu, X. (2015a). An ensemble method for unbalanced sentiment classification. In *Natural Computation (ICNC), 2015 11th International Conference on*. IEEE.
Zhang, K., Li, A., & Song, B. (2009). Fraud Detection in Tax Declaration Using Ensemble ISGNN. In *Computer Science and Information Engineering, 2009 WRI World Congress on*. IEEE.
Zhang, N. (2016). Cost-sensitive spectral clustering for photo-thermal infrared imaging data. In *International Conference on Information Science & Technology*.
Zhang, X., Wang, B., & Chen, X. (2015b). Intelligent fault diagnosis of roller bearings with multivariable ensemble-based incremental support vector machine. *Knowledge-Based Systems, 89*, 56–85.
Zhang, X., Yang, Z., Shangguan, L., Liu, Y., & Chen, L. (2015c). Boosting mobile Apps under imbalanced sensing data. *Mobile Computing, IEEE Transactions on, 14*(6), 1151–1161.
Zhang, X., Y. Zhuang, H. Hu and W. Wang (2015d). "3-D Laser-Based Multiclass and Multiview Object Detection in Cluttered Indoor Scenes."
Zhang, Y., Fu, P., Liu, W., & Chen, G. (2014). Imbalanced data classification based on scaling kernel-based support vector machine. *Neural Computing and Applications, 25*(3-4), 927–935.
Zhang, Y., Zhang, D., Mi, G., Ma, D., Li, G., Guo, Y., et al. (2012). Using ensemble methods to deal with imbalanced data in predicting protein–protein interactions. *Computational Biology and Chemistry, 36*, 36–41.
Zhang, Z., Krawczyk, B., Garcìa, S., Rosales-Pérez, A., & Herrera, F. (2016b). Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data. *Knowledge-Based Systems*.
Zhao, X. M., Li, X., Chen, L., & Aihara, K. (2008). Protein classification with imbalanced data. *Proteins: Structure, function, and bioinformatics, 70*(4), 1125–1132.
Zhao, Z., Zhong, P., & Zhao, Y. (2011). Learning SVM with weighted maximum margin criterion for classification of imbalanced data. *Mathematical and Computer Modelling, 54*(3), 1093–1099.
Zhong, W., Raahemi, B., & Liu, J. (2013). Classifying peer-to-peer applications using imbalanced concept-adapting very fast decision tree on IP data stream. *Peer–to-Peer Networking and Applications, 6*(3), 233–246.
Zhou, L. (2013). Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems, 41*, 16–25.
Zhou, Z.-H. (Ed.). (2016). *Machine Learning*. Tsinghua University press.
Zhou, Z.-H., & Liu, X.-Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *Knowledge and Data Engineering, IEEE Transactions on, 18*(1), 63–77.
Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning, 3*(1), 1–130.
Zięba, M., & Tomczak, J. M. (2015). Boosted SVM with active learning strategy for imbalanced data. *Soft Computing, 19*(12), 3357–3368.
Zięba, M., Tomczak, J. M., Lubicz, M., & Świątek, J. (2014). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing, 14*, 99–108.
Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the Best Classification Threshold in Imbalanced Classification. *Big Data Research*.