A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions

HARSURINDER KAUR, HUSANBIR SINGH PANNU, and AVLEEN KAUR MALHI,
Thapar Institute of Engineering and Technology, India

In machine learning, the data imbalance imposes challenges to perform data analytics in almost all areas of real-world research. The raw primary data often suffers from the skewed perspective of data distribution of one class over the other as in the case of computer vision, information security, marketing, and medical science. The goal of this article is to present a comparative analysis of the approaches from the reference of data pre-processing, algorithmic and hybrid paradigms for contemporary imbalance data analysis techniques, and their comparative study in lieu of different data distribution and their application areas.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

Additional Key Words and Phrases: Data imbalance, machine learning, data analysis, sampling

ACM Reference format:
Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. 2019. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *ACM Comput. Surv.* 52, 4, Article 79 (August 2019), 36 pages.
https://doi.org/10.1145/3343440

1 INTRODUCTION AND MOTIVATION

Imbalanced data classification is a problem in which proportional class sizes of a dataset differs relatively by a substantial margin. In this, one class is at least depicted with just a few number of samples (called the minority class) and rest falls into the other class (called the majority class). Basically with this issue, the performance of a classifier leans to be partial towards some classes (majority class) in the imbalanced data set. Performance bias means solutions behave differently on majority and minority classes. In the majority class, solutions lean toward better accuracy. However, solutions results with poor accuracy carried out on the side of minority class. The problem of imbalanced data distributions is a well posed problem in realistic applications such as fault detection, fraud detection, bleeding detection in medical diagnoses, and so on. The most popular methods used for imbalanced data are neighborhood cleaning rule, safe level smote, cost sensitive algorithm and neural networks. An illustration of imbalanced data is depicted in Figure 1.

There are three broad approaches to address imbalance issues in the data classification: pre-processing methods, algorithmic centered approaches, and hybrid approaches. At the

Authors' addresses: H. Kaur, H. S. Pannu, and A. K. Malhi, CSED, Thapar Institute of Engineering and Technology, Patiala, India 147004; emails: hkaur_me16@thapar.edu, {hspannu, avleen}@thapar.edu.

0360-0300/2019/08-ART79 $15.00
https://doi.org/10.1145/3343440

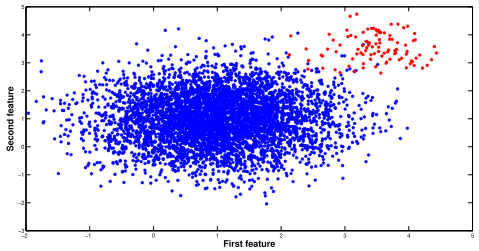

Fig. 1. Example of imbalanced data problem.

pre-processing stage, methods involve a number of re-sampling methods, such as random over-/under-sampling or consolidating the two sampling methods as a motive to obtain approximately equal count of samples in the classes (Abouelenien et al. 2013). However, this technique only deals with training data set and balances it while learning algorithm stays the same (Singh and Purohit 2015). Algorithmic centered approach includes assumptions created to favour the minority class and changing the costs to get the balance classes. In this article, the following key-points are analyzed:

- Prominent issues in imbalanced data classification
- Comparative study of methods used to approach imbalance data distributions

1.1 Motivation for Work

The inspiration for this survey is to thoroughly inspect the problem of imbalance data emerging around all real-world applications and providing an overview for the state-of-art solutions for the imbalance data challenges. This article proposes a cross-wise view of the present trends in the imbalance data techniques and their comparative study using different methods. The distinct motivation for this comprehensive review is as follows:

(1) To study existing and efficient techniques and solutions for imbalanced data in lieu of problems.
(2) To analyze the applicable areas of imbalanced data classification and how performance of real-world applications suffer from imbalance data class population.

Thus, there is need of efficient family of methods to explore this recurrent problem. Therefore, analyzing existing techniques while understanding their focus/assumptions is imperative. It will help to redefine the classification problem with comparable data class proportions to a derive generalized form of the data distribution for optimal solutions pursuit.

1.2 Our Contributions

- The comprehensive study has been conducted to investigate the issues related to imbalanced data classification in machine learning.
- Application areas of imbalanced data classification have been discussed.
- Comparative analysis of all the methods based upon different evaluation metrics has been performed to handle issue of imbalance data.

1.3 Related Search

Imbalanced data issue has gained immense interest due to its complications in many fields, and many researchers have contributed in this area. In this section, we present related contributions of other researchers. Many researchers have reviewed imbalanced data distributions found in many

domains, challenges that arise, and evaluation metrics based on benchmark datasets. They have also discussed comparative analysis of various methods and learning algorithms used to tackle issue of imbalanced data distributions. Haixiang et al. (2017) proposed categorical wise study of all applicable areas of imbalanced data problem. Sun et al. (2009) proposed review on imbalanced data classification in contrast to taxonomy of all application areas and have discussed all open issues. Comparative analysis of methods and suggested that hybrid methods can give better results in Longadge and Dongre (2013). State of the art solutions have been discussed in He and Garcia (2009). Issues related to Multi-class classification for imbalanced data is discussed and methods to tackle and minimize the issue are also proposed in Sahare and Gupta (2012). In Kotsiantis et al. (2006) suggested resampling techniques and hybrid methods can do quite better as compared to other methods. Comparative analysis of various ensemble learning algorithms and concluded that RUSBoost algorithm has good performance and least complex among all (Galar et al. 2012). Anwar et al. (2014) proposed complexity measure related to classifier's performance due to unbalanced data on basis of k-nearest neighbour learning algorithm. Chawla (2009) proposed SMOTE and its combination with other learning algorithms. In Bekkar and Alitouche (2013) relative study of methods in contrast to its advantages and disadvantages is proposed. Table 1 presents comparison of existing surveys on imbalanced data classification.

1.4 Article Organization

Section 1 presents an introduction to the research work and the motivation behind it. Section 2 presents the review technique followed to analyze information sources for existing research, statistics trends, and search criteria. Section 3 describes applications domains of imbalanced data and related literature to applications. Section 4 describes various approaches for imbalanced data. Comparative analysis of learning algorithms is presented in Section 5. Section 6 describes various performance metrics for evaluation of classification model. Section 7 presents comparative analysis of solutions with major open issues in Section 8 and finally our conclusion in Section 9. The outline for the review can be represented as shown in Figure 2.

2 REVIEW TECHNIQUE

The categorical technique proposed in this research article has been taken from the literature proposed by Yanmin et al. (Sun et al. 2009). The several phases used in this survey are to generate a review technique, designing a comprehensive and detailed study, comparison of techniques, comparative result analysis, and investigating open issues. The review technique engaged in this survey is presented in Figure 4.

2.1 Information Sources

The electronic databases used for searching are described in Figure 3.

2.2 Search Criteria

For almost all the searches, the keywords "imbalance" and "sampling" were found in every research paper in the abstract. In this article, we have included the research papers from peer review journals, symposiums, conferences, technical reports, lecture notes, workshops and white papers from industry. The search criteria used is elaborated in Figure 4. The electronic databases used for searching information from Google Scholar, IEEE Explorer, Springer, ACM Digital Library, Wiley Interscience, and so on. At the first stage, there were 385 articles collected, from which 290 were reduced on the basis of title. Further, on the basis of abstract and conclusion, 218 papers were selected. The count was further reduced to 160 on the basis of full text. Finally, this count was explored widely to attain the final count of 152 articles.

Table 1. Related Search

Research Paper	Defines classification problem	Nature of problem	Challenges faced	Learning algo	Evaluation metrics	Domain areas
A review on imbalanced classification (Sun et al. 2009)	✓	✓	✓	✓	✓	✓
A review of class imbalance problem in dtaa mining (Longadge and Dongre 2013)	✓				✓	
A review on handling imbalanced datasets (Kotsiantis et al. 2006)	✓			✓	✓	
A survey on learning from imbalanced data (He and Garcia 2009)		✓		✓	✓	
A review on ensembles for class imbalance (Galar et al. 2012)	✓	✓		✓	✓	
Overview of imbalanced data (Chawla 2009)	✓			✓	✓	
A review on imbalanced data learning approaches (Bekkar and Alitouche 2013)	✓	✓		✓		
A review of multi-class classification for imbalanced data (Sahare and Gupta 2012)	✓			✓		
A review of methods and applications for imbalanced data (Haixiang et al. 2017)	✓		✓	✓		✓
Classification problems with unbalanced data (Anwar et al. 2014)	✓	✓	✓	✓	✓	
Our Survey	✓	✓	✓	✓	✓	✓

2.3 Statistics Trends

The publishing trends demonstrate that interest on imbalanced data classification has been grown from previous years. The publishing trends from year 2006 to 2016 have been elaborated in Figure 5. Figure 6 illustrates the related papers' count in the corresponding journals during the past decade.

3 APPLICATION DOMAINS

The problem of imbalanced data classification emerges as a major issue in many real-world applications, thus reducing the predictive performance of the model.

3.1 Application Areas

This section presents possible applicable areas of Imbalanced data. Imbalance in data distribution arises as a big issue in these application areas, thus reducing the classification accuracy. We have classified application domain into 10 categories. In addition to this, there are other many

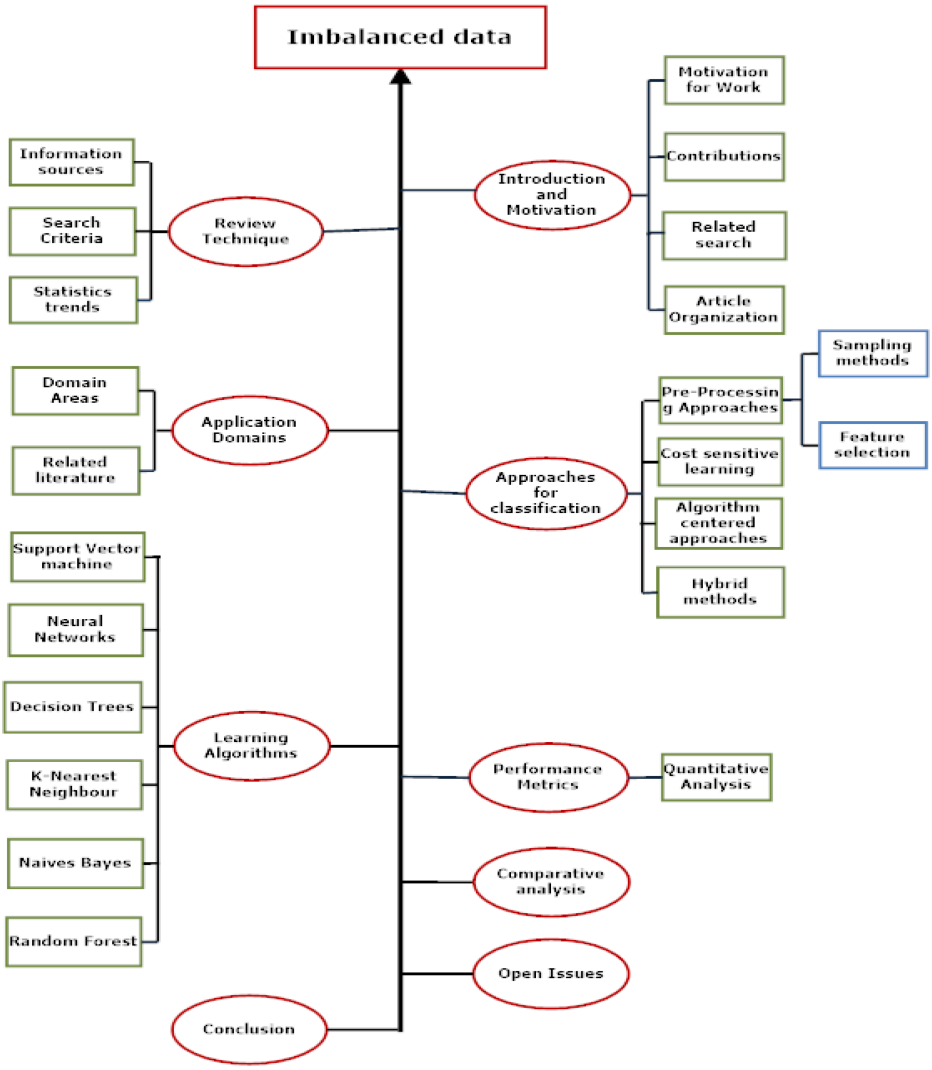

Fig. 2. The outline of the review process.

application areas, like prediction of emergency events, prediction of pollution, and so on. Figure 7 represents the application areas of imbalanced data classification. Table 3 describes various application areas in lieu of detailed applications with related articles.

(1) Computer Vision: System that retrieves some useful information from the images. In this area, one tries to associate the high-dimensional image features to a structured labeling of objects in the image. Imbalance occurs when a large amount of images from the training data does not contain the object of interest, thus known as negative images as compared to positive objects. This might result in misclassification of positive objects. In Gao et al. (2014), Enhanced and Hierarchical Structure (EHS) method is proposed for the imbalance of positive and negative class in a massive video dataset. It outperforms the most common ML algorithms, such as under-sampling and over-sampling.

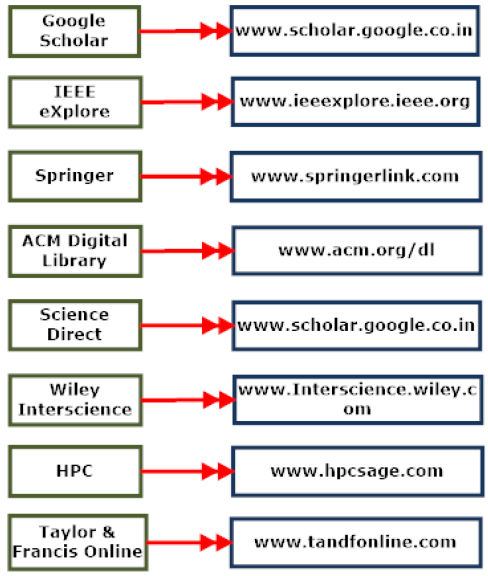

Fig. 3. Electronic database used for information.

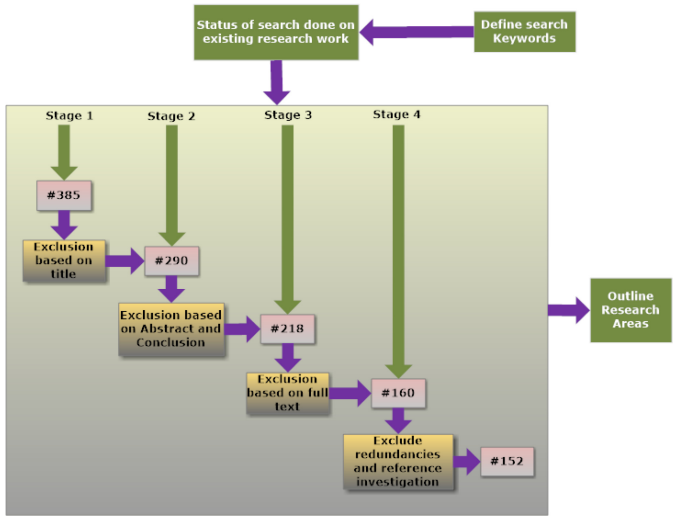

Fig. 4. Search criteria used for information.

(2) Information Security: Over the past few years, machine-learning algorithms have been successfully used in protecting the useful information. Imbalance of data emerges while deciding the decision boundary for anomaly classification. In Nepal and Pathan (2014), a comparative study of the techniques involved in security, trust, and privacy issues in cloud systems has been proposed.
(3) Fraud Detection: Frauds such as credit card fraud, cheque fraud, and so on, might be an expensive issue to an individual or an organization. Every year billions in loss of money is introduced, due to limitations of machine-learning algorithms while dealing with the massive imbalanced class distributions and other reasons. So, this poses a threat in

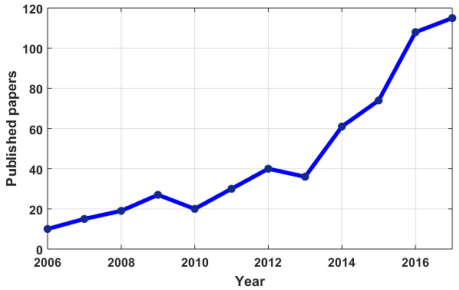

Fig. 5. Imbalanced classification publishing trend.

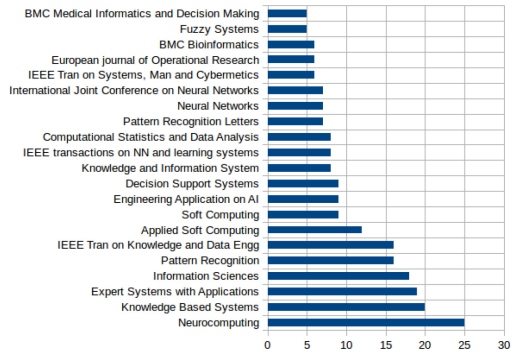

Fig. 6. Top journal publications count in the area of imbalanced data learning during the past decade.

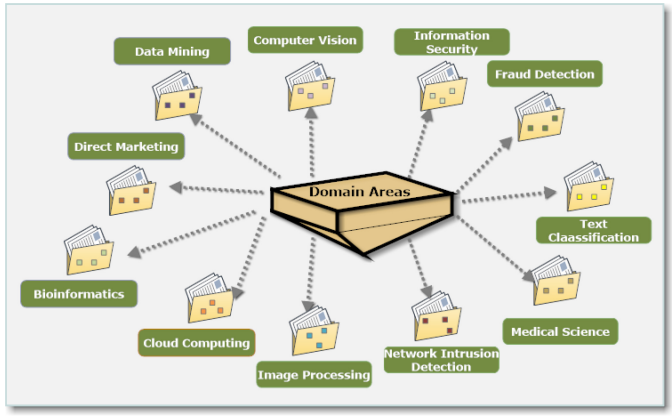

Fig. 7. Imbalanced data application domain areas.

investigating between the fraudulent masterminds and genuine users. However, the number of legitimate users is far more than fraudulent. ContrastMiner algorithm effectively differentiates between genuine and fraud users in online banking (Wei et al. 2013a).

(4) Medical Science: In hospitals, a huge amount of information is stored in large databases about the patients and their medical history. And in medical diagnoses, it is very critical to differentiate between the positive (unhealthy) and negative (healthy) patients, because usually patients with diseases are reasonably rare as compared to the normal or healthy population. So, this poses a critical threat for the classification between the healthy and normal population.

(5) Network Intrusion detection: With the growing demand, network-based computer systems are playing a significant role to reduce the human efforts. However, the attacks on networks and computer machines have also been growing simultaneously. For example, Elbasiony et al. (2013) and Thomas (2013) used different techniques to tackle the problem of imbalance in network traffic.

(6) Image Processing: Image processing is an method of conducting operations on images or extracting useful information from them. In this, issue of imbalance of data occurs

while distinguishing the class distributions between the novel and mis-classified features (Hodge and Austin 2004).

(7) Cloud Computing: The open and distributed cloud environment has been mostly liked by the intruders. Detecting the anomalies in the cloud environment seems as a difficult task, due to imbalance classification. The same issue arises while detecting the failures also.

(8) Data Mining: In Longadge and Dongre (2013), a comparative study of techniques to tackle imbalanced class distribution in data mining has been discussed.

(9) Text Classification: Text classification is an efficient method of supervised machine learning that enables various applications such as spam filtering, sentiment analysis, and so on. For instance, in Sarker and Gonzalez (2015), NLP and ML techniques have been used to automatic detect adverse drug reaction using multi-corpus training using social media text including medical information.

(10) Direct Marketing: Direct marketing in today's digital world has taken over all of the traditional methods. But, while dealing with imbalanced classes, the predictive models of the consumer's behaviour and responses often gets affected. Thus, study of advantage of using SVM algorithm is to overcome the issue of imbalanced class in the consumers responses is proposed in Kim et al. (2013).

(11) Bioinformatics: Imbalanced data has a major application in the area of bioinformatics, which includes protein sub-cellular prediction proposed in Wan et al. (2017), gene prediction in Wang et al. (2015) and Batuwita and Palade (2009). Protein classification includes Song et al. (2014) and Zhao et al. (2008), promoter prediction in Zeng et al. (2009), and so on. There exists sensitivity to imbalance data in the case of classification algorithms, which produces sub-optimal classification results. The goal of employing imbalanced learning methods is to raise sensitivity and lower specificity as much as possible.

3.2 Related Literature

This section presents the literature on different domains areas of imbalanced data. Du et al. (2017) have proposed an innovative classification technique especially to handle imbalanced data. The technique is known as *post-boosting* of classification boundary for *imbalanced* data (PBI) to enhance ANN classification performance considerably over various datasets with diverse range of imbalance ratios, dimensions and sizes consistently. First steps involves imbalanced ANN classification followed by PBI adjustments through geometric mean containing both majority and minority classes. A new metric that is the ratio of majority loss over minority advance ratio (MMR) to handle imbalance ratio as bad as 0.001. PBI is well suited for massive datasets with higher imbalance ratio and consumes 1/10 of the original time required for training.

Vong et al. (2018) have proposed an extended GM-based postboosting (PBG) for multiclass imbalanced learning for online sequential datasets using ANN. Three adjustments are the reason for its effectiveness: GM post adjustment for classification boundary; new rule for updation with higher GM along with information retention; dynamic adjustments for data diversity/scarcity. Below are brief acronyms used in this research.

- Online sequential multiclass imbalance learning (OS-MIL)
- dynamic changing data diversity/scarcity (DCDD, DCDS)
- imbalance class distribution (ICD)

Table 2 shows the brief comparison of PBG against existing OS-MIL methods. Du and Vong (2018) have proposed another robust learning approach to overcome the data imbalanced challenge involving data distribution changes dynamically along with labels.

Table 2. Comparison of PBG Against OSMIL

	ICD Aspect	OSMIL	PBG
General	Data Scarcity	✓	✓
	Data majority	×	✓
Online	DCDS	✓	✓
	DCDD	×	✓

Wan et al. (2017) have explored an efficient prediction algorithm for sub-cellular localization of proteins. It performs multi-label classifier using an ensemble called HPSLPred and is an ideal choice to handle imbalanced data for multi-label classification. It outperforms other state-of-art algorithms while yielding 75.89% average precision value (APValue). Major contributions include designing a 350D comprehensive feature model, self dependable dimension selection and HPSL-Pred ensemble model for the optimal performance.

Song et al. (2014) have classified DNA-binding proteins for imbalanced data using an ensemble approach nDNA-Prot. To define the structure of protein, 188 dimensional features have been extracted to feed into ensemble called imDC for DNA-binding proteins classification. Selected features possess maximum relevance and minimum redundancy to yield accuracy of 95.8% and 0.986 for AUC for cross validation. Testing accuracy for the proposed nDNA-Prot was found to be 86% outperforming both DNA-prot (68%) and iDNA-Prot (76%). Batuwita and Palade (2009) is another proposed technique by Want et al. using ensemble learning with micro RNA data and UCI datasets. The proposed method has been compared against LibID, BalanceCascade, AsymBoost, UnderSampl, HSampl, and AdaBoost. MiRNAs has 193:8494 for positive and negative examples and 30 pos with 1000 neg were used for testing. Sensitivity and specificity over ImDC, LibID, and Triplet-SVM were found to be {0.86,0.83,0.93} and {0.93,0.92,0.88}.

For the application area Computer Vision, in Shyu et al. (2008) and Gao et al. (2014) depicts imbalance of class in semantic analysis in massive video dataset. Celebi et al. (2007) proposed issue of imbalanced data while classifying the dermoscopy images. In Fraud detection, Zhang et al. (2008) and Wei et al. (2013b) discussed detection of online banking fraud where imbalance data arise as a major issue. For credit card fraud detection also, where identifying the anomalous person responsible for fraud is important task, imbalance data diminishes the accuracy of classifiers for detection (Fu et al. 2016; Kulkarni and Ade 2016). In Zakaryazad and Duman (2016) (Mardani and Shahriari 2013) and Sahin et al. (2013), proposed various approaches for detection of credit card fraud. In the case of financial application where predicting one's eligibility to pay full loan back is a most-valuable task for banks, private organization, imbalance data poses threat (Abeysinghe et al. 2016; Sanz et al. 2015). Fuzzy rules have been employed to achieve higher understanding of the anomaly prediction in the imbalanced data classification model in Sanz et al. (2015). It does not uses any pre-processing or sampling and thus avoids any noise involvement. For testing, 11 real-world financial datasets were analyzed while outperforming SMOTE-based oversampling, C4.5 decision tree and type-1 fuzzy models. In Abeysinghe et al. (2016) proposed techniques for detecting insurance fraud while dealing with imbalanced data. Improving the performance of detection of fraud in retail surveillance is also proposed by Pan et al. (2011). Imbalanced data involved in insurance fraud has been explored in Hassan and Abraham (2016), which introduces the ensemble of ANN, SVM and Decision Trees used with and without replacement techniques for under-sampled class. It chooses among the undersampling partitionings to choose the best one as a claim of its originality. DT is the winner in the empirical analysis.

In medical science and diabetes, Yu and Ni (2014) proposes an improved random subspace method and bagging ensemble to derive feature subspace, which keeps balance in diversity and

accuracy of the classifiers. SVM is the base classifier and the proposed ensemble approach has been analyzed on all types of performance metrics. In Vo and Won (2007) there is a discussion of the classification bias induced due to imbalanced data, thus an extended regularized least square method is proposed. Errors are penalized according to various weights, which are determined using the specified rules for each sample. For cancer-related science, in Krawczyk et al. (2016) the ensemble of EUSBoost is studied, which involves the boosting idea for under-sample evolution for each of the base classifier. Level set active contours method yields an effective features extraction for better classification of the breast cancer symptoms for clinical decision support system. In Yang et al. (2016), tumor tissues analysis has been performed in various gene samples. One versus all multi-classification model has been used, which divides the problem into multiple binary classification problem. To deal with unbalanced data, it uses balanced sampling and feature selection. Gentle-Boost ensemble (Can-CSC-GBE) and cost sensitive classifier has been studied in Ali et al. (2016) to detect the breast cancer using the features of protein amino acid. Comparative empirical analysis is performed using cost sensitive learning and ensemble of AdaBoostM1 and bagging. For hepatitis virus (HBV and HCV) prediction, one feature selection method and three balancing methods have been studied in Richardson and Lidbury (2017) using SVMs for pathological cases. Random forest are used for feature selection and data from ACT Pathology Canberra Australia involving records of 18,645 patients have been experimented. For hepatitis virus (HBV and HCV) prediction, one feature selection method and three balancing methods have been studied in Richardson and Lidbury (2017) using SVMs for pathological cases. Random forest are used for feature selection and data from ACT Pathology Canberra Australia involving records of 18,645 patients have been experimented. In Yap et al. (2014), cardiac surgery classification is discussed using undersampling, oversampling, boosting and bagging techniques for binary classification. CHAID, C5 and CART classifiers have been used and sensitivity and precision have been reported to work well using decision tree. Moreover in this study bagging and boosting have no improvement effects on DT performance. Cytogenetic domain classification has been studied in Lerner et al. (2007), using hierarchical decomposition followed by up-sampling of minority classes and reduction of dimensions. Each hierarchical level tackles a smaller problem of approximately balanced data classes. Multilayer perceptron NN and Naive Bayesian have been employed to analyse the smallness of data being better than its imbalance nature. Another semi-supervised technique is studied in (Herndon and Caragea 2016) by using both labeled and unlabeled data in a domain adaptation environment. Logistic regression-based two classifiers have been used for splice site forecast in gene prediction while achieving precision-recall between 50.83%–82.61%.

In Peiravian and Zhu (2013), malicious Android apps are classified using API calls and permissions. Shabtai et al. (2012) discuss malicious code classification on Opcode patterns for anti-virus software as signatures for more than 30,000 files. Nepal and Pathan (2014) have discussed cloud systems based upon QoS premises while discussing security fundamentals and contemporary technology. Security is involved with detection of anomalies and intrusion detection in the majority dominated negative class. In Song et al. (2010), the skewed and concept drift data streams have been studied in context to cloud security using one-class classifier in an ensemble settings and k-means. Use of genetic programming and incremental ensemble has been used to detect cyber security drifts in Folino et al. (2016).

Taneja et al. (2015) have studied advertisement frauds on the internet mobile service. Feature selections has been performed using recursive feature elimination and Hellinger Distance Decision Tree is used for classification while achieving 64.04% accuracy. A fake escrow website has been studied in Abbasi and Chen (2009) for fraud cues, which are extracted from webpages. SVM, ANN, DT, Naive bayes and PCA were compared for the empirical analysis on a test bed of 90,000 pages through 410 websites. In Zhong et al. (2013) concept adapting very fast decision tree has been used

for P2P application based upon large volume data set communities leaving and joining. NonP2P outnumbers P2P data by big margin and CVFDT along with re-sampling technique to monitor the Internet traffic. Imbalanced data learning on User to Root intrusion has been studied in Engen et al. (2008) using MLO and DT along with evolutionary ANN. Kasai and Oike (2010) discuss image pickup apparatus, which receives larger dataset with more light and few dataset with lesser light to perform image processing of various types. Another visual information retrieval for imbalanced dataset has been studied in Chang et al. (2003) by including active learning, quasi-bagging, class-boundary alignment, adaptive dimensionality reduction, and recursive subspace co-training.

Cloud computing infrastructure anomaly detection has been studied in Pannu et al. (2012) and Fu et al. (2012) using adaptive one class SVM and both one and two class SVMs, respectively. Credit card fraud mining has been studied in Bhattacharyya et al. (2011) using SVM, RF, and logistic regression. Longadge and Dongre (2013) have the imbalance data mining from the prospective of feature selection, pre-processing of data and algorithmic approach. Dua and Du (2016) is a cyber-security-oriented data-mining reference for imbalanced data and anomaly detection.

In text classification, the imbalanced data has been studied in Liu et al. (2009), using simple probability-based term weighting scheme and information ratios for minor class. Wang et al. (2013) explained the text sentiment classification using BRC while under sampling the higher density regions and SVM was used while reporting precision and recall. Duman et al. (2012) reviewed the database marketing classification problem of a bank using approaches such as CHIAD, ANN and logistic regression. Zakaryazad and Duman (2016) have discussed direct marketing and fraud detection using profit driven penalty function-based ANN. Spam review detection has been studied in Al Najada and Zhu (2014) for imbalanced data using a bagging-based approach called iSRD. A good survey information on review spam detection has been studied in Crawford et al. (2015). Burez and Van den Poel (2009) have studied customer churn prediction against class imbalance problem using sampling and modeling techniques such as weighted RF and gradient boosting.

The summary of literature survey has been given in the Table 3 based upon the application areas.

4 APPROACHES FOR IMBALANCED DATA CLASSIFICATION

In the past decade, various approaches have been proposed to overcome the issue of imbalanced data classification. In this section, we will propose detailed summary of state-of-the-art machine-learning techniques to address the issue. For this, pre-processing approaches, cost-sensitive learning methods, hybrid methods, and algorithmic approaches have been discussed. The classification of various approaches to handle imbalance data is depicted in Figure 8.

Figure 9 gives the citations of the various imbalanced data classification techniques adopted by the researchers in a hierarchical manner for various application areas.

4.1 Pre-processing Approaches

Pre-processing approaches are those that are performed on training data. Pre-processing techniques are applied to gain better training data. The approaches that act on pre-processing phase are also referred as data centered approaches. These approaches work by directly acting upon data space and tries to reduce the imbalance ratio between classes. Actual classification stage is adapted by pre-processing approaches, thus these are flexible to follow. General idea of basic pre-processing strategies is depicted in Figure 10.

The most popular approach, over-sampling adds artificial samples to the data space, known as SMOTE (Ramyachitra and Manikandan 2014). The following subsections describes the various sampling methods.

Table 3. Application in Different Domains

Application areas	Detailed application	Related Articles
Computer Vision	Video dataset	(Shyu et al. 2008) (Gao et al. 2014)
	Object recognition	(Celebi et al. 2007)
Fraud detection	Online banking	(Zhang et al. 2008) (Wei et al. 2013b)
	Credit card	(Zakaryazad and Duman 2016) (Mardani and Shahriari 2013) (Sahin et al. 2013) (Fu et al. 2016) (Kulkarni and Ade 2016)
	Loan prediction	(Abeysinghe et al. 2016) (Sanz et al. 2015)
	Insurance fraud	(Hassan and Abraham 2016) (Pan et al. 2011)
Medical Science	Diabetes	(Yu and Ni 2014) (Vo and Won 2007)
	Cancer diagnoses	(Krawczyk et al. 2016) (Yang et al. 2016) (Ali et al. 2016)
	Hepatitis Virus	(Richardson and Lidbury 2017)
	Parkinson's disease	(Yap et al. 2014)
	Genetics	(Lerner et al. 2007)
	DNA identification	(Herndon and Caragea 2016)
Information Security	Malware dection	(Peiravian and Zhu 2013) (Shabtai et al. 2012)
	Cloud security	(Nepal and Pathan 2014) (Song et al. 2010) (Folino et al. 2016)
	Online advertisement	(Taneja et al. 2015)
	Fake website	(Abbasi and Chen 2009)
Network Intrusion	Traffic analysis	(Zhong et al. 2013)
Detection	Intrusion detection	(Engen et al. 2008)
Image Processing	Pixel evaluation	(Kasai and Oike 2010)
	Information retrieval	(Chang et al. 2003)
Cloud Computing	Adaptive Failure Detection	(Pannu et al. 2012)
	Anomaly Detection	(Fu et al. 2012)
Data Mining	Real-world applications	(Bhattacharyya et al. 2011) (Longadge and Dongre 2013) (Dua and Du 2016)
Text Classification	Textual data	(Liu et al. 2009)
	Sentiment classification	(Wang et al. 2013)
Direct Marketing	Consumer Behaviour	(Duman et al. 2012) (Zakaryazad and Duman 2016)
	Review Spam detection	(Al Najada and Zhu 2014) (Crawford et al. 2015)
	Churn prediction	(Burez and Van den Poel 2009)
Bioinformatics	Gene prediction	(Batuwita and Palade 2009) (Wang et al. 2015)
	Protein classification	(Song et al. 2014) (Zhao et al. 2008)
	Protein sub-cellular prediction	(Wan et al. 2017)
	Promoter prediction	(Zeng et al. 2009)

4.1.1 Sampling Methods. It is an easy and popular approach to balance the class distributions of the training data. The original data space gets balanced by employing one of the method among over-sampling or under-sampling. The basic idea of re-sampling is to yield balanced classes and process is repeated till balanced classes are obtained. It works by adding or removing the samples from the data space to diminish the biased behaviour of imbalanced data, thus changing the size of the training data space. Cao and Zhai (2015) proposed that sampling methods reduces learning time and faster execution once we get the balanced classes. Cao and Zhai (2015) have proved that sampling methods are effective alternative for supervised learning. Yap et al. (2014) have proved that sampling methods outperforms bagging and boosting. The summary of sampling methods is proposed in Table 3. There are various options to perform the sampling:

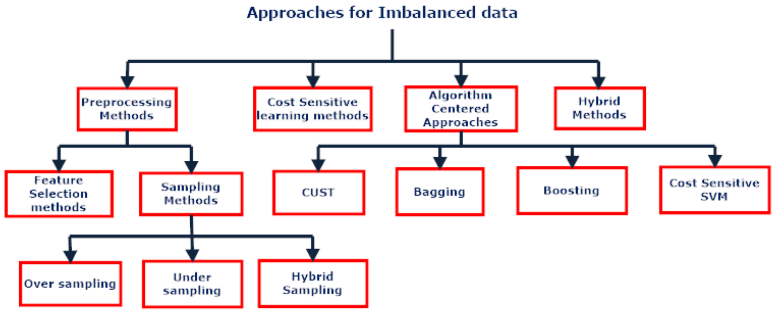

Fig. 8. Classification of approaches for imbalanced data.

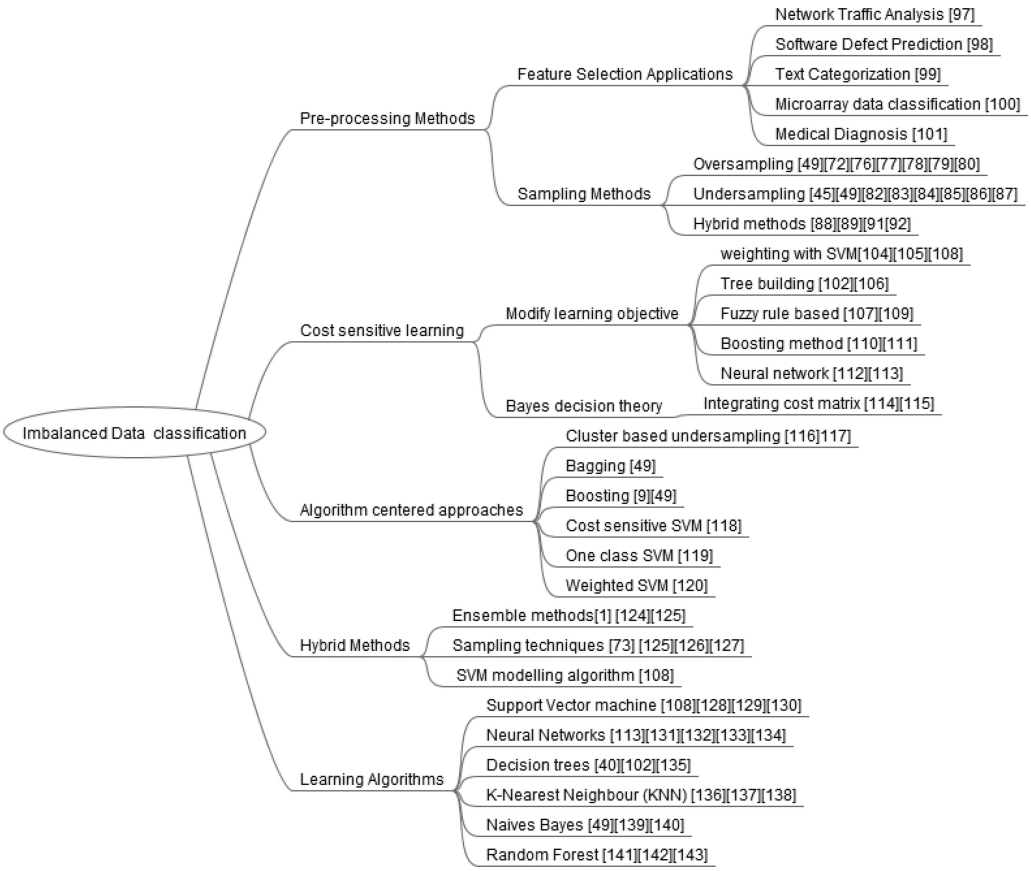

Fig. 9. Overview of the literature for imbalanced data classification.

- Over-sampling: The basic idea of over-sampling is to increase the size of the minority class to obtain balanced classes. Duplication of samples is done in random over-sampling in which samples are randomly selected. Thus, class size increases due to duplication of samples. Likely, over-fitting is the main issue arises in over-sampling (Ganganwar 2012). In Chawla et al. (2002), Chawla proposed Synthetic Minority Over-Sampling technique (SMOTE). In SMOTE, synthetic samples are produced by the help of minority class

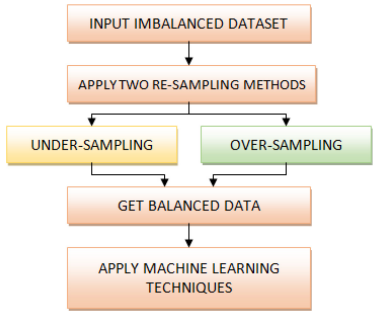

Fig. 10. Flow for data-oriented approaches.

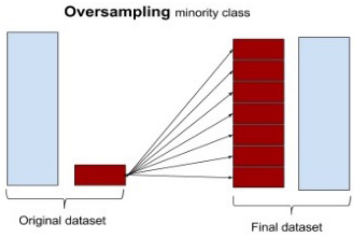

Fig. 11. Example of over-sampling.

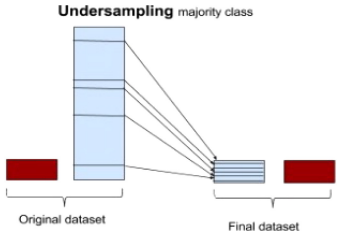

Fig. 12. Example of under-sampling.

samples. It also suggests the idea that using combination of sampling methods can be a good option for improving the classifier performance while dealing with the imbalanced class distributions. In Bach et al. (2017), a comparative study of undersampling and over-sampling against osteoporosis diagnoses where only 7.14% cases are of minority class and their focus was to identify the best outcome from under-sampling and over-sampling with classifiers. They concluded that SMOTE along with RandomForest classifier achieved highest performance. Yap et al. (2014) proposed a comparative study of sampling techniques, bagging and boosting against imbalanced dataset and concluded that sampling techniques are better for imbalanced data using decision tree where bagging and boosting does not improve decision tree performance in regard to their dataset. Zhang and Li (2014) proposed random walk over-sampling approach for problem of imbalance in data distribution by generating samples to increment the number of samples in minority class. Ramyachitra and Manikandan (2014) proposed review for the imbalanced dataset classification and solutions to reduce it. Moreo et al. (2016) proposed oversampling method for imbalanced text classification where to each minority class document in the training class, a probabilistic function is assigned. Zheng et al. (2016) proposed a technique that can overcome the limitations of SMOTE, i.e., SNOCC, which creates new samples and ensures that generated new samples find the new nearest neighbours. This proposed technique outperforms SMOTE and other methods. Ensemble-based method SMOTE along with Boosting for handling imbalanced PubChem BioAssay data, which outperforms combination of Random Forest with SMOTE on the basis of sensitivity and G-mean (Hao et al. 2014). Figure 11 depicts general idea of over-sampling the minority class to gain balance in data.

- Under-sampling: It is a pre-processing method that draws the random set of samples from the majority class to balance the classes and rest of the samples are ignored (Nguyen et al. 2012). The size of the data space is measured to draw desirable class distribution ratio. Thus,

under-sampling helps in gaining the equal number of class samples and makes training phase faster. However, the main issue arises because there lies possibilities of losing informative instances from the majority class while deleting the instances. Galar et al. (2013) proposed random undersampling along with boosting algorithm to handle imbalanced data distribution. Yap et al. (2014) proposed a comparative study of sampling techniques, bagging and boosting against imbalanced dataset and concluded that sampling techniques are better for imbalanced data using decision tree where bagging and boosting does not improve decision tree performance in regard to their dataset. Yu et al. (2013) proposed heuristics under-sampling method, i.e., ACOSampling to address the imbalance data distributions in DNA micro-array dataset. This method while under-sampling does not delete the samples with useful information rather it automatically extract and save them. In Krawczyk et al. (2016), an ensemble of EUSBoost is proposed that involves the boosting idea for under-sample evolution for each of the base classifier. Level set active contours method yields an effective features extraction for better classification of the breast cancer symptoms for clinical decision support system. Another method in Kang et al. (2016) focuses on eliminating the noisy minority class samples, which hinders the performance in imbalanced data classification. Dai and Hua (2016) proposed comparative analysis of various random under-sampling ensemble methods in medical imbalanced dataset where correctly predicting true positive rate is difficult. In Peng et al. (2016), a data gravitation classification model is used, which is efficient for supervised learning methods to handle the issue of imbalanced data by undersample, which shows an effective margin in sensitivity and specificity in results. Tomek Link combined with random under-sampling to deal with imbalanced dataset and comparative analysis of other methods and algorithms has been done and concluded that using Tomek Link with combination of Tomek Link and SMOTE is better idea (Elhassan et al. 2016). Figure 12 depicts general idea of under-sampling the majority class.

- Hybrid sampling: Hybrid sampling methods are those that apply both re-sampling techniques to attain balance in the data. Hybrid sampling techniques are proposed in Qian et al. (2014) and Charte et al. (2015). Wang (2014) proposed a technique of combining sampling methods, under-sampling and over-sampling to handle the problem of imbalance data. To get a balanced training data space, under sample to delete the instances without containing useful information. Then over-sampling is done to replicate existing instances. Thus, the proposed method reduces the chances of losing informative instances. Adding huge number of synthetic samples to training space resulted in increasing the classification performance. SMOTE is a hybrid sampling method, because it generates synthetic examples in minority class and it is an alternative to duplication of minority class samples. It is a good method to tackle imbalanced data. It falls into the hybrid category, because it uses combination of under-sampling and over-sampling to overcome the issue of imbalanced data distribution and does not rely on under-sampling only. Figure 13 illustrates the SMOTE method.

Kubat et al. (1997) proposed one side selection method in which Tomek Links are used to reject the noisy and unreliable examples from the majority class. Thus, it under-sample the majority class in an efficient way. And then, CNN is used to delete the samples that are distant to the decision boundary. Thus, it saves informative samples while under-sampling the majority class. In He et al. (2008), Adaptive Synthetic sampling approach (ADASYN) is proposed for assigning different weights to samples according to their level of complication while learning and synthetic data is generated (See Table 4). It proves to be an efficient way of handling imbalanced data in the following ways:

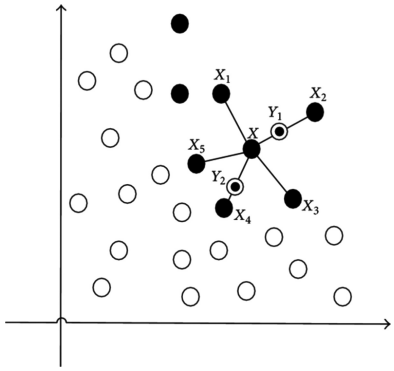

Fig. 13. Illustration of synthetic examples in SMOTE.

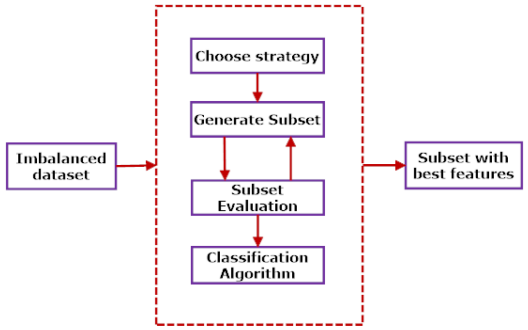

Fig. 14. Example of Feature Selection Method.

Table 4. Summary of Sampling Methods

Sampling	Related Articles	General Idea
Over-sampling	(Yap et al. 2014), (Bach et al. 2017), (Zhang and Li 2014)	increases minority class size
	(Ramyachitra and Manikandan 2014), (Moreo et al. 2016), (Zheng et al. 2016),	flexible,free from classifier use
	(Hao et al. 2014)	over-fitting may occur
Under-sampling	(Galar et al. 2013), (Yap et al. 2014), (Yu et al. 2013), (Krawczyk et al. 2016)	deletes examples from majority class
	(Kang et al. 2016), (Dai and Hua 2016), (Peng et al. 2016), (Elhassan et al. 2016)	domain specific
		goal-oriented
		chances of loosing useful information
		free from classifier use
Hybrid sampling	(He et al. 2008), (Kubat et al. 1997), (Qian et al. 2014), (Charte et al. 2015)	combines sampling methods
		reduces limitations of sampling methods

(1) It tends to reduce the situation of imbalance data where hyperplane always gets biased toward the majority class
(2) Generates the classification hyperplane in a efficient manner that it automatically leans in the direction of instances that are difficult to learn

Over-sampling tends to increase the size of the data space, which results in a situation of over-fitting and takes more time in training phase. However, under-sampling removes some set of samples randomly from the majority class where lies the possibility of losing informative samples.

4.1.2 Feature Selection and Extraction. Selecting subset of relevant features or attributes from high dimensional data sets helps to upgrade the performance of the classifier. This method is proposed in imbalanced data classification (Jamali et al. 2012; Maldonado et al. 2014; Van Hulse et al. 2009). Selecting features is generally gained by three methods filter method, wrapper method and embedded method. The comparative analysis of advantages and disadvantages is discussed in Saeys et al. (2007). The ultimate aim of feature selection is to select the set with best features from the whole dataset to gain better classifier performance. Whereas feature extraction is also a dimensionality reduction technique, it differs from feature selection in the way that it generates new features using the primary ones. Feature extraction methods are Principal Component Analysis (PCA), Singular Value Decomposition (SVD). Feature selection and extraction is found

Table 5. Summary of Cost-sensitive Learning

Strategy	Method	Related Articles
Modifying learning objective to build cost-sensitive classifier	weighting with SVM	(Tang et al. 2009) (Cao et al. 2013) (Dhar and Cherkassky 2015)
	Tree-buliding	(Krawczyk et al. 2014) (Qiu et al. 2017)
	Fuzzy rule-based	(Palacios et al. 2014) (López et al. 2015)
	Boosting method	(Beijbom et al. 2014) (Xia et al. 2017)
	Neural Network	(Arar and Ayan 2015) (Ghazikhani et al. 2014)
Bayes decision theory	Integrating cost matrix	(Chai et al. 2004) (Moepya et al. 2014)

applicable in various real-world applications such as network traffic analysis (Liu et al. 2015), software defect prediction (Khoshgoftaar et al. 2014), text categorization (Yang et al. 2014), microarray data classification (Bolón-Canedo et al. 2015), medical diagnoses (Zhang et al. 2014). The example of selecting subset with best features from the imbalanced dataset is described in Figure 14.

4.2 Cost-Sensitive Learning

Cost-sensitive learning techniques are cost-specific techniques that find cost associated with miss-classified examples. Generally, these takes miss-classification cost into exploration. In comparison to sampling methods, cost-sensitive learning methods are less popular because of the misclassification cost are unidentified from data and difficulty arises while setting costs (Krawczyk et al. 2014). Sampling methods are easy to implement and more popular as compared to these. But cost-sensitive learning is more computationally effective technique. Cost-sensitive neural network for handling imbalanced data is proposed in Khan et al. (2017). Cao et al. (2013) proposed integrating wrapper feature selection method to choose best set of features and misclassification cost for cost-sensitive SVM, showed comparative results with other methods on evaluation basis of AUC and G-mean. Dhar and Cherkassky (2015) proposed Universum Support Vector Machine (U-SVM) for handling highly imbalanced dataset. Qiu et al. (2017) proposed randomly selected decision tree for cost sensitive learning. The proposed technique works by reducing the cost value and improving the classification performance. Palacios et al. (2014) proposed FURIA algorithm for cost sensitive learning of fuzzy rules for handling imbalanced data distributions. Summary of modifying learning objective to build cost-sensitive classifier with strategies is presented in Table 5.

4.2.1 Challenges in Data Centered Approaches. Although data centered approaches are simple and most commonly used in handling skewed data distributions, but there exists various challenges in their application.

(1) The major challenge occurs while completing the objective of resampling the classes, i.e., there are chances of over-fitting in the majority class while oversampling and information loss from the minority class in case of undersampling (Galar et al. 2013).
(2) The foremost challenge is to select the optimal class distribution in a dataset, as the case of selecting optimal class distribution is different in scenario of every dataset, as it affects the effectiveness of the classifier.
(3) The next challenge in skewed data distribution is uncertain method of selecting samples (Sun et al. 2009)
(4) Other challenges include different distributions in various subclasses contained in a single class as it results in incrementing the learning complexity of the dataset. Various solutions have been accounted in data-oriented approaches but may be functional in specific contexts.

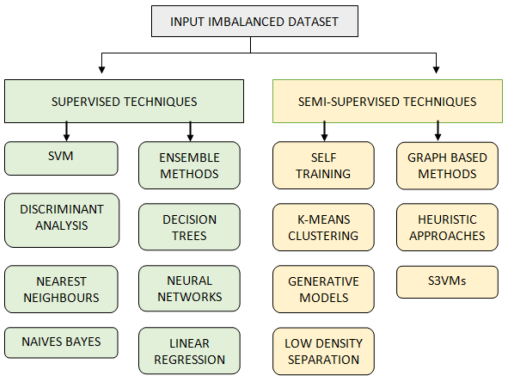

Fig. 15. Classification for supervised and semi-supervised techniques.

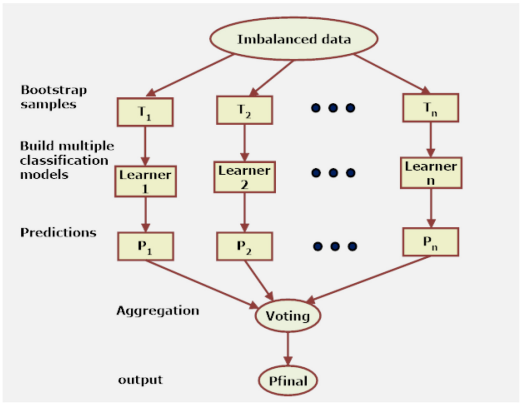

Fig. 16. General idea of bagging.

4.3 Algorithm Centered Approaches

For solving the imbalanced data classification problem, creating modern algorithms or upgrading the existent one is known as algorithmic centered approaches. There are various methods that have been proposed by researchers and literatures for algorithmic centered approaches. Assigning misclassification cost for correcting the class predictions are the most popular algorithm approaches (Singh and Purohit 2015). Algo-centered approaches are presented in Figure 15 and are discussed below:

- Cluster-based under-sampling: As there is possibility of losing the informative instances with under-sampling. So to reduce this effect of under-sampling, Cluster-based under-sampling technique proves to be an alternative to random under-sampling. In cluster-based undersampling training data space is divided into *n* number of clusters. Then, there is selection of right samples from the plagiarised clusters. The primary idea behind this method that training data space splits into n different clusters and every cluster poses distinguishable characteristics (Zhang et al. 2010). Sowah et al. (2016) proposed a new cluster-based under-sampling technique. First, training data space is cleaned by deleting noisy and unreliable samples from majority class. Then, rest of the majority class samples split into n clusters. The experiment was verified on 16 benchmark data sets with class imbalance problem using two machine-learning algorithms C4.5 Decision Tree and OneR, which yielded results better than other existing algorithms.
- Bagging: stands for Bootstrap Aggregation. It is an efficient idea to decrease the variance of the prediction. In this, original data space is divided into multi sets having same size. Thus, each multi set containing same size creates one classifier. Aggregation of particular classifiers contributes a compound classifier. Yap et al. (2014) proposed a comparative study of sampling techniques, bagging and boosting against imbalanced dataset. Figure 16 represents the general idea of bagging algorithm for dealing the imbalanced data classification.
- Boosting: refers to machine-learning ensemble meta-algorithm, which works for eliminating the variance and bias. Effective and accurate prediction rules are generated by integrating various weak and inaccurate ones. It suggests an idea that searching various inaccurate and weak rules seems to be much simplified as compared to highly accurate prediction rule. Comparative analysis of various ensemble learning algorithms and concluded that RUSBoost algorithm yields good performance and least complex among all (Galar et al. 2012).

Yap et al. (2014) proposed a comparative study of sampling techniques, bagging and boosting against imbalanced dataset.

- Cost-Sensitive SVM: the basic goal of SVM is to classify data by building the linear decision boundary (hyper-plane). It splits the data points according to their class. SVM dealing with balanced data, gives best performance. But in the case of highly imbalanced its hyperplane gets partial towards the majority class. So, to handle this problem, a method is proposed in which every class's classification error gets assigned a different penalty cost. It refers to a class-specific error favoring SVM algorithm (Lee et al. 2016). The model formulation is as follows:

$$arg \min_{(w,b)} \frac{1}{2}\|w\|^2 + P \sum_{i:yi>0} \lambda_i + N \sum_{j:yj<0} \lambda_j \tag{1}$$
$$s.t. \quad \left[w^T x_i + b\right].yi \geq 1 - \lambda_i, i \in \mathcal{N},$$

where λ_i is the soft margin slack variable, P, N are penalties associated with positive and negative mis-classifications, (x_i, y_i) is sample and its associated label.

- One Class SVM: In this case, SVM is trained using data space having only one class (normal), thus refers to one class SVM. It was first presented by Scholkopf et al. for evaluation of high-dimensional distribution. In Raskutti and Kowalczyk (2004), the authors proposed imbalance data handling solutions with highly unequal distributions and suggested that one-class learning from class containing positive samples only, yields perfect performance dealing with highly imbalanced data.
- Weighted SVM: Huang and Du (2005) proposed the disadvantages of classification using standard support vector machine. While dealing with imbalanced data classification, the standard support vector machine generates the linear decision boundary(hyper-plane) biased towards the majority class. However, in weighted support vector machines, the classification error of the minority class gets better. Moreover, it assists in decreasing the influence of outliers in binary data distribution as compared to standard SVM. It shows better classification performance. Following equation states the weighted SVM objective function with weights λ_i:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + P \sum_i \lambda_i \xi_i \tag{2}$$
$$\text{s.t. } y_i(w.\phi(x_i) + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0, i \in \mathcal{N}.$$

4.3.1 Challenges in Algorithm Centered Approaches. There is a common problem of dataset shift that exists in all classification problems. Dataset shift is the problem where different distributions is followed by training and testing data.

(1) The minority class is quite sensitive to errors in classification due to less number of examples existing in the case of highly imbalanced domains. If we consider the extreme case of single misclassified example, then there will be significant performance drop. There is need of potential approaches to handle such type of misclassified examples.
(2) The foremost challenge is the natural dataset shift (Cieslak and Chawla 2009), where the data of interest produces relevant degree of shift, which results in drop in performance. It is possible that techniques might be developed for discovering and measuring dataset shift presence but how to adapt it to focus on the minority class is really challenging.

Table 6. Major Challenges in Approaches

Challenge	Data level	Algorithm level
1.	Objective of resampling the classes results in chances of overfitting and loss of significant information	Prior knowledge of influence of imbalance data distribution on classifiers
2.	Selection of optimal class distribution as the affect of this on learning algorithm varies with data	Information of the learning algorithm and application areas
3.	How to effectively resample the data, because resampling is not adequate in all scenarios	Designing of imbalanced classification algorithms that may work under dataset shift conditions
4.	What is the optimal method to select examples? Uncertain method of selecting quality samples as informative re-sampling methods are not sufficient in highly imbalanced data distribution	Methods adopted for handling dataset shift presence that may also focus on minority class issues is really challenging
5.	Unequal distribution of samples in subclasses representing single class that affects the sampling methods and increments cost of data preprocessing	Forcing the classifier toward minority class, which yields error and output compensation does not prove efficient as majority of the samples will originate from majority class only

(3) Furthermore, there exists no articles in the literature that have focused on the designing of imbalanced classification algorithms that may work under dataset shift conditions, which is only possible either by employing some pre-processing technique (Moreno-Torres et al. 2013) or by using ad hoc algorithm (Bickel et al. 2009).
(4) Another problem that exists is induced dataset shift in imbalanced data classification algorithms. The state-of-the-art techniques mostly employ artificial means of inducing potential source of shift in machine-learning process by using stratified cross-validation techniques. Hence, there is requirement of more subtle validation technique so that artificial means of dataset shift can be avoided.
(5) The output adjustment may be done by possibly overdriving the classifier toward minority class, which results in increase in error in majority class. Therefore, output compensation does not prove useful as majority of the objects will originate from majority class only.

Table 6 explains the challenges in the application of data level and algorithmic level approaches. Table 11 lays out the relative comparison among various data and algorithmic techniques along with data sets used, performance metrics and prominent findings.

4.4 Hybrid Methods

Making combination of pre-processing and algorithm-oriented approaches to handle efficiently the problem of imbalanced data classification refers to Hybrid methods. Sun et al. (2015) proposed a novel ensemble method in which imbalanced dataset is divided into creating many balanced dataset. Thus, from data many classifiers are made using appropriate classification algorithm. The aggregate of the output of classifiers generated by a specific ensemble method. Zhang et al. (2015) proposed solution to handle imbalanced data problem by integrating classification and ensemble learning techniques. By using both the pre-processing and algorithmic level approaches, it results in rising classification problem of the imbalanced sentiment dataset. Tang et al. (2009) proposed GSVM-RU, which stands for granular SVM-repetitive under-sampling, an algorithm

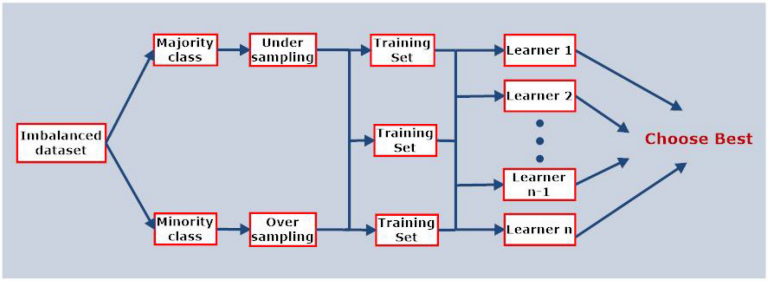

Fig. 17. Possible hybrid combinations for imbalanced data classification.

specifically produced for the highly imbalanced dataset. The main idea behind this is to delete the noisy and unreliable samples and extraction of useful and consistent samples to achieve good classification accuracy. The proposed technique outperforms other methods while comparison. Although, re-sampling techniques achieves success in imbalanced data classification, a hybrid method of combining sampling techniques with Bagging is proposed by Lu et al. (2016). The experiment is performed on 26 benchmark datasets that showed that sampling methods outperforms bagging whereas bagging lacks in outperforming sampling methods. So, a hybrid re-sampling method is introduced in Cao and Zhai (2015) that deals with binary class imbalance data. First, SMOTE is used to rise the count of minority class instances. Then One Side Selection (OSS) rejects the instances without any useful information and achieved feasible results. Abouelenien et al. (2013) introduce cluster-based sampling and ensemble method. To tackle the problem of dealing with large imbalanced dataset, clusters are generated from all training space. Then, from the training space, representative data is elected from generated clusters as training instances, which yields improvement in evaluation measures i.e. accuracy and sensitivity. A SVM modeling Algorithm in which training space is increased by generating synthetic instances with SMOTE is proposed by Tang et al. (2009). Then, building SVM on the sampling technique over-sampling the data set. Majority Weighted Minority Oversampling Technique (MWMOTE) is proposed to tackle imbalanced data by filtering the useful instances from minority class. Then weights are assigned with respect to their euclidean distance (Barua et al. 2014). Figure 17 represents possible combinations of sampling methods with learning algorithms to handle imbalanced dataset. In this scenario, choice can be made among sampling and learning algorithm combination according to classification problem.

5 LEARNING ALGORITHMS

In this section, some of the classifier learning algorithms are discussed. Table 7 describes six learning algorithms with their learning strategy and limitations while dealing with data imbalance and the related literature.

5.1 Support Vector Machine

The traditional SVM algorithms have a limitation of low performance on imbalanced data. Thus, different heuristics were incorporated in SVM modelling, which includes oversampling, undersampling, and cost-sensitive learning by Tang et al. (2009). Another approach was proposed by Hwang et al. (2011) based on lagrangian SVM for classification of imbalanced data. Its feasibility and effectiveness is compared by testing it with other SVM models. Zięba et al. (2014) proposed a hybrid approach by combining the adaboost algorithm with SVM for imbalanced data classification, which minimized the weighed exponential error function that is used to predict the life of lung cancer patients after undertaking an operation. Shao et al. (2014) proposed a weighted

Table 7. Overview of Learning Algorithms

Algorithm	Learning Method	Limitations	Related Articles
Support Vector Machine (SVMs)	Finds optimal hyperplane separation with maximal margin	Sensitive to highly imbalanced data and decision boundary gets inclined towards small class	(Tang et al. 2009), (Hwang et al. 2011), (Zięba et al. 2014), (Shao et al. 2014)
Neural Networks	Iteratively weight adjustment to minimize error	Reduces error for frequent class only	(Liu et al. 2013), (Jeatrakul et al. 2010), (Khoshgoftaar et al. 2010), (Ghazikhani et al. 2014), (Ghazikhani et al. 2013), (Du et al. 2017), (Vong et al. 2018)
Decision Trees	Divide training data recursively and Prunes sub tree if error occurs	So many splits needed to detect imbalance	(Krawczyk et al. 2014), (Sanz et al. 2015), (Park and Ghosh 2014)
K-Nearest Neighbour (KNN)	Determine class label by high rank class among k nearest neighbours	The k nearest neighbour carry higher probabilities of samples form frequent class	(Majid et al. 2014), (Rani et al. 2016), (Patel and Thakur 2016)
Naives Bayes	assumes independencies among features	misclassification of samples of small class	(Tang et al. 2016), (Yap et al. 2014), (Bhandari and Patel 2015)
Random Forest	generates multiple classifiers and combine their output	sensitive with non-linear relationship between dependent and independent variables.	(Wu et al. 2014), (Khabsa et al. 2016), (Du et al. 2015)

Lagrangian twin support vector machine for imbalanced data classification using different training points. It is tested and compared with other models on real as well as synthetic datasets.

5.2 Neural Networks

Neural Networks have got a huge range of applications in machine-learning algorithms. Liu et al. (2013) proposed a technique of data processing for electronic tongue and the results of accuracy were compared with other classification algorithms to detect the samples of orange beverage and chinese vinegar. Arar and Ayan (2015) proposed a technique of training neural network using Artificial Bee Colony (ABC) for finding optimal weights for software defect prediction and this optimization technique proved effective in predicting the defective modules. In Fu et al. (2016), convolutional Neural Network is used for detection of credit card fraud, which is applied on transactions from commercial bank and results show that it outperforms conventional methods of detecting credit card fraud. In Jeatrakul et al. (2010), a combination of techniques by under-sampling using complementary neural network and over-sampling using SMOTE is proposed for re-distribute the data distribution to solve class imbalance issue.

5.3 Decision Trees

Decision trees are the most popular algorithms that can be used for classification. Krawczyk et al. (2014) proposed a cost sensitive decision trees for the imbalanced data classification, since real datasets are mostly imbalanced. Evolutionary algorithm is employed for the classifier selection and weights assignment simultaneously. Another variation for decision tree was given by Sanz

et al. (2015), who adduced an evolutionary fuzzy classification method for the process of modelling and prediction in financial applications of real world, which calculates good prediction accuracies tested on 11 financial application datasets. Park and Ghosh (2014) introduced two ensemble approaches based on decision trees for the problem of imbalanced data classification by utilizing the characteristics of α-divergence. The effectiveness of the proposed ensembles is shown by performing the experimental results on multi-class imbalanced datasets.

5.4 K-Nearest Neighbour

K-Nearest Neighbour is one of the most used classification algorithm in machine learning due to its easy implementation. In Majid et al. (2014), at first stage for balancing the data, samples of the minority class are increased and then two algorithms' results are compared for predicting human breast and colon cancer where SVM outperforms KNN. Patel and Thakur (2016) proposed hybrid weighted nearest neighbour for handling imbalanced data in which weights and number of K are assigned according to class sizes. García-Pedrajas et al. (2017) proposed a technique to set the optimal value of k, as setting the optimal value of k is major issue with KNN.

5.5 Naives Bayes

Tang et al. (2016) proposed classification technique for automatic text categorization using class-specific features. It focuses more on selecting the best features for improving the performance of the classifier. Yap et al. (2014) used oversampling, undersampling, boosting and bagging methods to classify imbalanced data. The precision and sensitivity are used on cardiac surgery dataset to measure the performance, since accuracy will be biased toward majority class. Clustering and classification algorithms were used by Bhandari and Patel (2015) for the early prediction of student failure by using high dimensional imbalanced data.

5.6 Random Forest

Random forest algorithm has got wide range of applications in machine learning. Wu et al. (2014) proposed Random Forest for imbalanced text categorization and applied on three benchmark datasets. The basic idea behind this is to check training space contains useful features in the minority and majority class using sampling methods. In Khabsa et al. (2016), the random forest algorithm is used to accurately predict relevant studies for systematic reviews, which was modeled as imbalanced classification issue. Features such as co-citations and embedded words are used to predict relevant studies. To classify urban building into various categories by integrating VHR image and GHS using semantic information by learning random forest with highly imbalanced data. This study has been done using semantic information, which helps in resolving many environmental and social problems (Du et al. 2015).

Figure 18 gives the detailed analysis of the learning algorithms, which can be used for the various application areas enlisted in Section 3.

6 PERFORMANCE METRICS

This section describes the various performance metrics that can be used in evaluation of the imbalanced data techniques.

6.1 Quantitative Analysis

One of the parameter for evaluating the performance of classification model, i.e., how good a classification model behaves is given by confusion matrix also referred as error matrix. It depicts the relationship between the examples for actual and predicted classifications. Tables 8 and 9 discuss about metrics and confusion matrix used for quantitative analysis of classification performance.

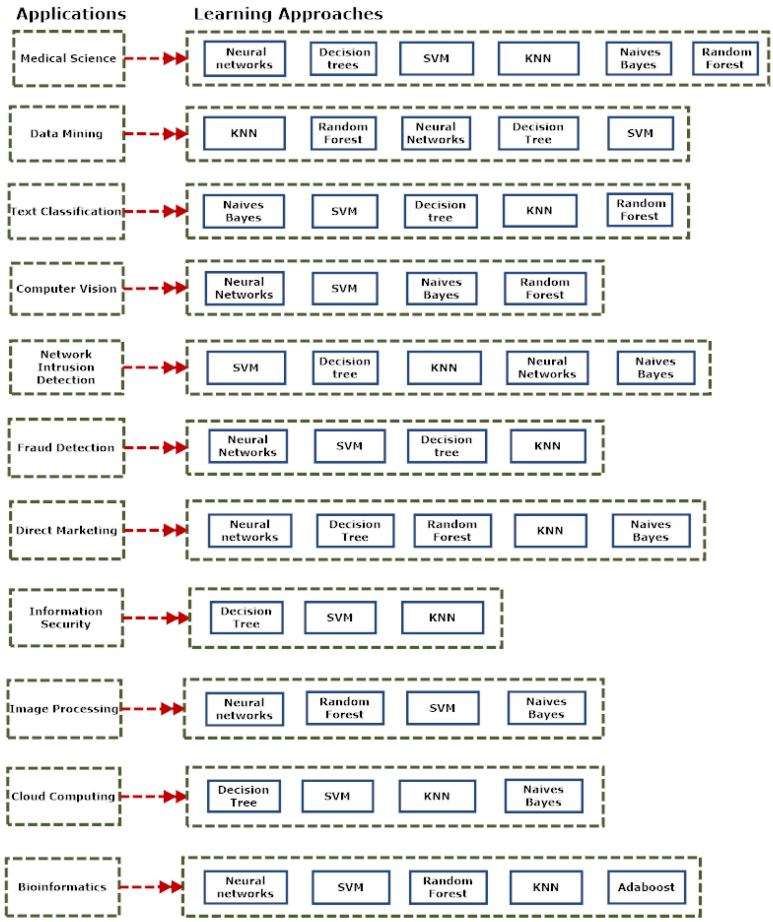

Fig. 18. Possible learning algorithms classification based on application areas.

Table 8. Major Metrics for Data Classification

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$
Sensitivity	$\frac{TP}{(TP+FN)}$
Specificity	$\frac{TN}{(TN+FP)}$
Precision	$\frac{TP}{(TP+FP)}$
Recall	$\frac{TP}{(TP+FN)}$
F-measure	$\frac{2*Precision*Recall}{Precision+Recall}$
G-MEAN	$\sqrt{sensitivity * specificity}$
AUC	$\frac{1+TPR-FPR}{2}$

Table 9. Confusion Matrix

	Actual Positive	Actual Negative
Observed Pos	TP	FP
Observed Neg	FN	TN

For two class classification, i.e., binary classification, each example that is predicted will belong to one of the four possible outcomes which are described in Table 8:

- TP (True Positive): the actually positive examples that are correctly predicted as positive.
- TN (True Negative): the actually negative examples that are correctly predicted as negative.
- FP (False Positive): the actually negative examples that are incorrectly predicted as positive.
- FN (Fasle Negative): the actually positive examples that are incorrectly predicted as negative.
- Accuracy: It acts as main evaluation parameter while dealing with binary decision problems for any prediction model. For any classification model, it measures the number of correctly predicted examples among all the total possible examples. Simply, ratio of the correctly predicted examples to the total number of examples present. Du et al. [2017] has proposed that considering geometric mean of accuracy of minority and majority class is more effective classification method as compared to common accuracy in case of highly imbalanced data.
- Sensitivity: it is the measure of positive examples that are correctly predicted by a model. It is sometimes also called as True Positive Rate (TPR) and equivalent to another evaluation metric, i.e., Recall. In medical diagnoses, sensitivity is defined as fraction of patients with disease, i.e., positive example and their test result also classified positive.
- Specificity: it is the measure of negative examples that are correctly predicted by a model. It is sometimes also called as True Negative Rate (TNR). In medical diagnoses, specificity is defined as fraction of people without disease, i.e., negative example and their test result also classified negative. Sensitivity and Specificity are defined with the help of confusion matrix. These two terms can be used in combination in some domains for the predictive performance of classification model (Sammut and Webb 2011).
- Precision: it is defined as the ratio of true positives (TP) to the total number of positive examples predicted.
- Recall: it is defined as fraction of true positive examples to all examples that are actually positive.
- F-measure: It is used to evaluate the accuracy of predictions while dealing with binary decision problems. Basically, it is harmonic mean of Precision and Recall.
- G-measure: it is evaluated as square root of multiplication of sensitivity and specificity.
- AUC-ROC: The Receiver Operating Curve (ROC) visually depicts the difference between accuracy on positive examples and error on negative examples where AUC is abbreviated as Area Under Curve. This curve evaluates trade-offs between true positives and false positives in respect to the limit of threshold of a predictive model. A detailed introduction to ROC Curve is disscussed in Fawcett (2006). AUC's performance varies in numbers $[0, 1]$ while worst performance is presented by 0 and, however, best performance is presented by 1. The model is classified as best if it obtains True Positive Rate and False Positive Rate as 1 and 0, respectively. A comparitive study of Precision-Recall and ROC Curves is illustrated in Davis and Goadrich (2006). Example of ROC curves for given classification model is illustrated in Figure 19. Various AUC curves based upon the classification of the data have been depicted in Figure 20.

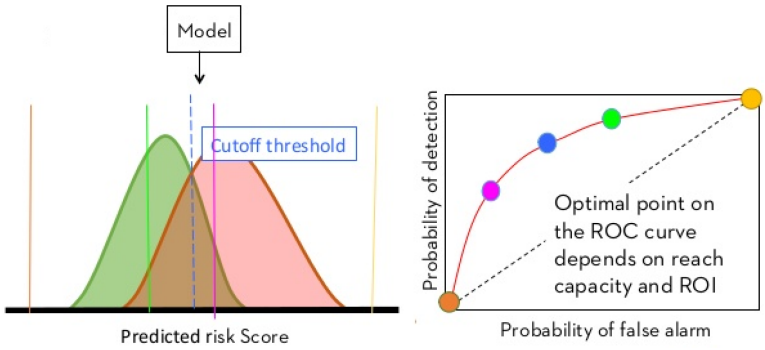

Fig. 19. Showcase of ROC/AUC curves for the given classification model.

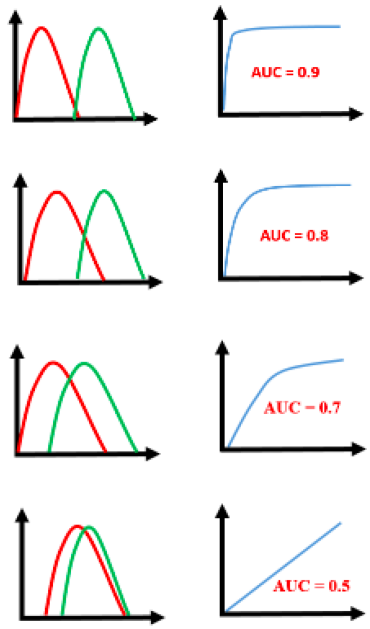

Fig. 20. Better AUC gives the idea of classifier's performance.

Let TPR and FPR be the true positive rate and false positive rates. In contrast to accuracy, the precision recall and F meaure are better for imbalanced data. If the cost is high for FP, then precision is an ideal measure. For example, in detection of spam emails, it is costly to loose an important email as spam. Similarly, if the cost associated with FN is higher, then recall is a good measure. For example, actual positive (patient diagnosis, fraud detection) getting predicted as negative (normal) has a dangerous consequence. F-1 score measures the balance between precision and recall. The formulas have been summarized in Table 8.

Zou et al. (2016) have exposed the limitation of ROC for the comparative analysis by arguing a better threshold rather than 0.5 for the testing set. This AUC limitation was found against detecting protein remote homology and experiments were performed by using an established single benchmark as threshold. The literature discussed an interesting example with precision and recall

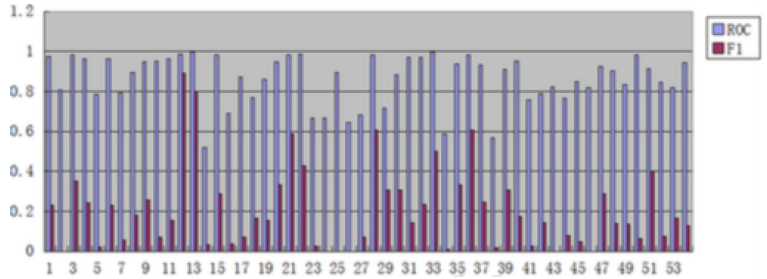

Fig. 21. Positive association between F1 and AUC (Zou et al. 2016).

Table 10. Comparative Analysis of Solutions

Approach	Advantage	Disadvantage
Pre-processing approach	* Relevant to any classification problem	
Sampling methods	* Simple Flexible, no classifier use * Straight-forward * Goal oriented	* Chances of over-fitting while over-sampling the minority class * Chances of information loss while under-sampling the majority class
Cost-Sensitive Weighting	* Simple * Flexible	* Chances of over-fitting in model searching optimal costs, extra learning * Costs are needed while real costs are missing
Algorithm oriented	* Domain specific * Handles imbalance and difference of cost of misclassification	* Algorithm explicit * Many trials are needed to decide appropriate one * Intrinsic Dataset shift * Lack of proper validation technique for datset shift
Cost-Sensitive Learning		* Insufficient when optimal cost values are missing
One-Class Learning		* Inadequate for learning algorithm calling examples from other classes
Hybrid approaches	* Minimizes chances of information loss Faster prediction	

being below 0.5 may yield AUC more than 0.9. Figure 21 shows the positive association between F1 and AUC making the proposed model more credible and efficient.

7 COMPARATIVE ANALYSIS

Table 10 illustrates the advantages and disadvantages of data pre-processing, algorithm and hybrid approaches. Table 11 gives the detailed comparison of the various classification techniques proposed so far for algorithm as well as data-oriented approaches by using various parameters. The comparison parameters used include approach used, learning method used, data type, performance metrics and key findings of the proposed approach. The research articles are divided

Table 11. Comparison of Classification Techniques for Algorithm or Data-oriented Approaches

Article	Approach	learning method	data	metrics used	findings
(Du et al. 2017)	ALGO	ANN	High imbalance ratio	G-mean-98.41±2.81 Sensitivity-96.56,Specificity-97.59	It consumes 1/10 of the original time required for training.
(Tang et al. 2009)	ALGO	GSVM-RU	7 highly imbalanced data set	G-mean-85.2 AUC-ROC-91.4,F-measure-66.5, AUC-PR-65.2	Minimize the chances of info loss in under sampling and faster SVM prediction
(Sowah et al. 2016)	ALGO	CUST	16 datasets	AUC-ROC-0.805, G-mean-0.766	Outperforms CBU, SMOTE,RUS, ROS, OSS
(He et al. 2008)	DATA	ADASYN	5 public datasets	Overall Accuracy-0.9257, Precision-0.8067, Recall-0.9015, F-measure-0.8505, G-mean-0.9168	Dynamically weight adjustment with respect to data distribution.
(Raskutti and Kowalczyk 2004)	ALGO	Extreme relabalcing for SVM	AHR-DATA, Reuters Newswire	AUC-ROC-0.99	Incase of dealing with highly imablanced data, 1-class is better
(Zhang et al. 2015)	DATA	Ensemble method	Songbo Tan's Hotel reviews	F-measure-0.69, G-mean-0.764,Weighted Accuracy(WA)-0.8257	Generates diverse base classifer
(Cao and Zhai 2015)	DATA	Hybrid resampling SVM	5 UCI datasets	AUC-ROC-0.7968	Improvment in AUC
(Chawla et al. 2002)	DATA	SMOTE	9 datasets	AUC-ROC-9560	Combination of over and undersampling enhances performance.
(Lee et al. 2016)	Both	ML algos for imbalance data fault detection	Etching process data and chemical vapor deposition process data	AUC-0.91, G-mean-0.92, F-measure-0.69	Excellent performance for any imbalance factor
(Wang and Yao 2013)	Both	5 learning algorithms	10 SDP dataset from public PROMISE repository	AUC-0.649,G-mean-0.762,Balance-0.711 Recall PF-0.823	Outperforms original AdaBoost.NC overall performance measures
(Gao et al. 2014)	ALGO	Enhanced and Hierachical Structure (EHS)	TRECVID Video Dataset	Features: GCM,Texture,SIFT and MoSIFT	robust and stable while dealing with different features
(Kim et al. 2013)	ALGO	SVM-RFM	4 datasets from Direct marketing education foundation (DMEF)	Accuracy-0.976,specificity-0.974,sensitivity-0.238,Gain value-.52	SVM is efficient for high dimensional datasets
(Arar and Ayan 2015)	ALGO	Neural Network with ABC algorithm	NASA MDP dataset	Accuracy-68.4,probability of false alarm-33.0, balance-71.8,(AUC)-0.79, and Normalized Expected Cost of Misclassification (NECM)	efficient combination of algorithm
(Barua et al. 2014)	ALGO	MWMOTE	20 real-world datasets	G-mean-0.7232,ROC-0.98834	does not perform well in Recall
(Wan et al. 2017)	ALGO	BRkNN, HOMER, MLkNN, $IBLR_{ML}$, DMLkNN	UniProtKB	APValue = 85.89%	Multi-thread tech. for dim. reduction; performs well only on human protein source
(Song et al. 2014)	ALGO	Ensemble of 16 sorting algos	total 44,996 with 9676 positives	accuracy = 86%, validation AUC = 0.986	Predicts DNA-binding protein sequences among all in UniProtKB/Swiss-Prot database; 188d attributes selected using max relv. min redund.
(Wang et al. 2015)	ALGO	Ensemble of DT, RandFor, SVM, NB, K-nn	UCI data and miRNA (193:8494)	Sen = {0.86,.083,0.93}, spc = {0.93,0.92,0.88} for ImDC, LibID and Triplet-SVM	time complexity and parameter tuning not considered

Table 12. Open Issues in Data and Algorithm Centered Approaches

Type	Open Issue	Current state
Algorithm level	Validation technique	dataset shift done artificially
Algorithm level	Overlapping between classes	Distinction of two classes difficult
Algorithm level	Incorporate background knowledge of data into classifier	need to design such classifiers
Algorithm level	Working with different granular levels to cover larger part of problem space	Some classifiers being proposed
Algorithm level	Overdriving classifier towards minority class	Increasing error in majority one
Algorithm level	Undersampling-based ensembles	Excellent classification performance questionable
Data level	Small sample size+imbalanced data	Minority class poorly represented
Data level	Intrinsic dataset shift	No proposed approach exists
Data level	Minority class	performance drop
Data level	Data pre-processing	No authentic solution so far
Data level	Deciding optimal class distribution	No necessarily optimal solution
Data level	Within class concepts	Need to select quality examples

based on the data-oriented or algorithm-oriented approach or any hybrid approach that employs the features of both approaches.

8 OPEN ISSUES

This article has discussed the spectrum of data imbalance in eleven prominent machine-learning application areas. Imbalanced data classification is a skewed data distribution learning of binary target class in this big data era. In addition to algorithm- and data-oriented approaches, hybrid approaches along with ensemble are gaining increasing popularity, which must consider the nature of the empirical dataset. There are often trade offs between real-life solutions, adaptive and computational efficiency. New directions can be found through the track of exploring the data distribution fractures between testing and training data, normalizing the dataset size of each class, incorporating the attribute specific weight adjustment to the under-sampled class, overlapping among the classes and small disjuncts. The current status of the open problems in algorithm and data level approaches is shown in Table 12.

For imbalance data operations, however, scaling issues arise while hampering the traditional approaches to perform well. So, in Fernández et al. (2017), MapReduce-based de facto technique is suggested for big data this year, to recursively divide and solve the distribution. But only few researches have been able to contribute in the big data imbalanced classification due the its adaptation difficulties related to MapReduce programming paradigm. Scarcity of the data and disjunctive nature also highlight the programming solutions for imbalance data classification (Krawczyk 2016; Pannu and Kaur 2017). Below are the some potential measures to develop the solutions against data imbalance in machine learning:

(1) Transformation to higher/different space (kernel space) to analyze the distribution
(2) Integrating active learning and semi-supervised learning with must-link and cannot link example analysis
(3) Minority class data distribution and characteristics

(4) Exploit various regression/clustering techniques to evaluate the anomalous looking minority classes
(5) Explore/ensemble various multi-class learning methods to extract the nature
(6) Capture inter class diverse relationships through multi-class learning

9 CONCLUSION

This article presents a comprehensive analysis about learning challenges due to imbalanced data distribution. In addition, characteristics, problems, and solutions have been discussed. As imbalanced data restricts the performance and accuracy of classifier, various methods and techniques have been proposed to overcome the negative effects of data imbalance. Comparison of modern data pre-processing and algorithmic approaches has been performed and discussed along with their application domains. These approaches have been tuned to get the generalized learning model in various proposed articles with validation of desired results. Although sampling methods are the most popular and simple to implement, real-world applications are involved with skewed data distributions so hybrid algorithmic approaches are also desirable. A detailed comparative study has been given about the various methods proposed in the state-of-the-art discussing various parameters of these studies. The challenges in algorithm as well as data-oriented approaches along with open issues have also been discussed.

REFERENCES

Ahmed Abbasi and Hsinchun Chen. 2009. A comparison of fraud cues and classification methods for fake escrow website detection. *Info. Technol. Manage.* 10, 2–3 (2009), 83–101.

Chirath Abeysinghe, Jianguo Li, and Jing He. 2016. A classifier hub for imbalanced financial data. In *Proceedings of the Australasian Database Conference*. Springer, 476–479.

Mohamed Abouelenien, Xiaohui Yuan, Balathasan Giritharan, Jianguo Liu, and Shoujiang Tang. 2013. Cluster-based sampling and ensemble for bleeding detection in capsule endoscopy videos. *Amer. J. Sci. Eng.* 2, 1 (2013), 24–32.

Hamzah Al Najada and Xingquan Zhu. 2014. iSRD: Spam review detection with imbalanced data distributions. In *Proceedings of the IEEE 15th International Conference on Information Reuse and Integration (IRI'14)*. IEEE, 553–560.

Safdar Ali, Abdul Majid, Syed Gibran Javed, and Mohsin Sattar. 2016. Can-CSC-GBE: Developing cost-sensitive classifier with gentleboost ensemble for breast cancer classification using protein amino acids and imbalanced data. *Comput. Biol. Med.* 73 (2016), 38–46.

Nafees Anwar, Geoff Jones, and Siva Ganesh. 2014. Measurement of data complexity for classification problems with unbalanced data. *Stat. Anal.ysis and Data Min.: ASA Data Sci. J.* 7, 3 (2014), 194–211.

Ömer Faruk Arar and Kürşat Ayan. 2015. Software defect prediction using cost-sensitive neural network. *Appl. Soft Comput.* 33 (2015), 263–277.

Malgorzata Bach, Aleksandra Werner, J. Żywiec, and W. Pluskiewicz. 2017. The study of under-and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. *Info. Sci.* 384 (2017), 174–190.

Sukarna Barua, Md Monirul Islam, Xin Yao, and Kazuyuki Murase. 2014. MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Trans. Knowl. Data Eng.* 26, 2 (2014), 405–425.

Rukshan Batuwita and Vasile Palade. 2009. microPred: Effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics* 25, 8 (2009), 989–995.

Oscar Beijbom, Mohammad Saberian, David Kriegman, and Nuno Vasconcelos. 2014. Guess-averse loss functions for cost-sensitive multiclass boosting. In *Proceedings of the International Conference on Machine Learning*. 586–594.

Mohamed Bekkar and Taklit Akrouf Alitouche. 2013. Imbalanced data learning approaches review. *Int. J. Data Min. Knowl. Manage. Process* 3, 4 (2013), 15.

Sanket M. Bhandari and Krunal Patel. 2015. A review on using clustering and classification techniques to predict student failure with high dimensional and imbalanced data.

Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, and J. Christopher Westland. 2011. Data mining for credit card fraud: A comparative study. *Decis. Supp. Syst.* 50, 3 (2011), 602–613.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. 2009. Discriminative learning under covariate shift. *J. Mach. Learn. Res.* 10 (Sep.2009), 2137–2155.

Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. 2015. Distributed feature selection: An application to microarray data classification. *Appl. Soft Comput.* 30 (2015), 136–150.

Jonathan Burez and Dirk Van den Poel. 2009. Handling class imbalance in customer churn prediction. *Expert Syst. Appl.* 36, 3 (2009), 4626–4636.

Lu Cao and Yikui Zhai. 2015. Imbalanced data classification based on a hybrid resampling SVM method. In *Proceedings of the Ubiquitous Intelligence and Computing and IEEE 12th International Conference on Autonomic and Trusted Computing and IEEE 15th International Conference on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom'15)*. IEEE, 1533–1536.

Peng Cao, Dazhe Zhao, and Osmar Zaiane. 2013. An optimized cost-sensitive SVM for imbalanced data learning. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 280–292.

M. Emre Celebi, Hassan A. Kingravi, Bakhtiyar Uddin, Hitoshi Iyatomi, Y. Alp Aslandogan, William V. Stoecker, and Randy H. Moss. 2007. A methodological approach to the classification of dermoscopy images. *Comput. Med. Imag. Graph.* 31, 6 (2007), 362–373.

Xiaoyong Chai, Lin Deng, Qiang Yang, and Charles X. Ling. 2004. Test-cost sensitive naive bayes classification. In *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM'04)*. IEEE, 51–58.

Edward Y. Chang, Beitao Li, Gang Wu, and Kingshy Goh. 2003. Statistical learning for effective visual information retrieval. In *Proceedings of the International Conference on Image Processing (ICIP'03)*, vol. 3. IEEE, III–609.

Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. 2015. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing* 163 (2015), 3–16.

Nitesh V. Chawla. 2009. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*. Springer, 875–886.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *J. Artific. Intell. Res.* 16 (2002), 321–357.

David A. Cieslak and Nitesh V. Chawla. 2009. A framework for monitoring classifiers' performance: When and why failure occurs? *Knowl. Info. Syst.* 18, 1 (2009), 83–108.

Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. 2015. Survey of review spam detection using machine-learning techniques. *J. Big Data* 2, 1 (2015), 23.

Dong Dai and Shaowen Hua. 2016. Random under-sampling ensemble methods for highly imbalanced rare disease classification. In *Proceedings of the International Conference on Data Mining (DMIN'16)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 54.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 233–240.

Sauptik Dhar and Vladimir Cherkassky. 2015. Development and evaluation of cost-sensitive universum-SVM. *IEEE Trans. Cybernet.* 45, 4 (2015), 806–818.

Jie Du and C. M. Vong. 2018. Online multi-label learning under dynamic changes in data distribution with labels. *Accepted and in Press IEEE Trans. Cybernet.* (2018).

Jie Du, Chi-Man Vong, Chi-Man Pun, Pak-Kin Wong, and Weng-Fai Ip. 2017. Post-boosting of classification boundary for imbalanced data using geometric mean. *Neural Netw.* 96 (2017), 101–114.

Shihong Du, Fangli Zhang, and Xiuyuan Zhang. 2015. Semantic classification of urban buildings combining VHR image and GIS data: An improved random forest approach. *ISPRS J. Photogram. Remote Sens.* 105 (2015), 107–119.

Sumeet Dua and Xian Du. 2016. *Data Mining and Machine Learning in Cybersecurity*. CRC Press.

Ekrem Duman, Yeliz Ekinci, and Aydın Tanrıverdi. 2012. Comparing alternative classifiers for database marketing: The case of imbalanced datasets. *Expert Syst. Appl.* 39, 1 (2012), 48–53.

Reda M. Elbasiony, Elsayed A. Sallam, Tarek E. Eltobely, and Mahmoud M. Fahmy. 2013. A hybrid network intrusion detection framework based on random forests and weighted k-means. *Ain Shams Eng. J.* 4, 4 (2013), 753–762.

T. Elhassan, M. Aljurf, F. Al-Mohanna, and M. Shoukri. 2016. Classification of imbalance data using tomek link (T-link) combined with random under-sampling (RUS) as a data reduction method. *J. Info. Data Min.* (2016).

Vegard Engen, Jonathan Vincent, and Keith Phalp. 2008. Enhancing network-based intrusion detection for imbalanced data. *Int. J. Knowl.-Based Intell. Eng. Syst.* 12, 5–6 (2008), 357–367.

Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recogn. Lett.* 27, 8 (2006), 861–874.

Alberto Fernández, Sara del Río, Nitesh V. Chawla, and Francisco Herrera. 2017. An insight into imbalanced big data classification: Outcomes and challenges. *Complex Intell. Syst.* (2017), 1–16.

Gianluigi Folino, Francesco Sergio Pisani, and Pietro Sabatino. 2016. An incremental ensemble evolved by using genetic programming to efficiently detect drifts in cyber security datasets. In *Proceedings of the Conference on Genetic and Evolutionary Computation Conference Companion*. ACM, 1103–1110.

Kang Fu, Dawei Cheng, Yi Tu, and Liqing Zhang. 2016. Credit card fraud detection using convolutional neural networks. In *Proceedings of the International Conference on Neural Information Processing*. Springer, 483–490.

Song Fu, Jianguo Liu, and Husanbir Pannu. 2012. A hybrid anomaly detection framework in cloud computing using one-class and two-class support vector machines. In *Proceedings of the International Conference on Advanced Data Mining and Applications*. Springer, 726–738.

Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2012. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst., Man, Cybernet., Part C (Appl. Rev.)* 42, 4 (2012), 463–484.

Mikel Galar, Alberto Fernández, Edurne Barrenechea, and Francisco Herrera. 2013. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recogn.* 46, 12 (2013), 3460–3471.

Vaishali Ganganwar. 2012. An overview of classification algorithms for imbalanced datasets. *Int. J. Emerg. Technol. Adv. Eng.* 2, 4 (2012), 42–47.

Zan Gao, Longfei Zhang, Ming yu Chen, Alexander G. Hauptmann, Hua Zhang 0003, and An-Ni Cai. 2014. Enhanced and hierarchical structure algorithm for data imbalance problem in semantic extraction under massive video dataset. *Multimedia Tools Appl.* 68, 3 (2014), 641–657.

Nicolás García-Pedrajas, Juan A. Romero del Castillo, and Gonzalo Cerruela-García. 2017. A proposal for local k values for k-nearest neighbor rule. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 2 (2017), 470–475.

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

Adel Ghazikhani, Reza Monsefi, and Hadi Sadoghi Yazdi. 2014. Online neural network model for non-stationary and imbalanced data stream classification. *Int. J. Mach. Learn. Cybernet.* 5, 1 (2014), 51–62.

Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. 2017. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* 73 (2017), 220–239.

Ming Hao, Yanli Wang, and Stephen H. Bryant. 2014. An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data. *Analyt. Chim. Acta* 806 (2014), 117–127.

Amira Kamil Ibrahim Hassan and Ajith Abraham. 2016. Modeling insurance fraud detection using imbalanced data classification. In *Advances in Nature and Biologically Inspired Computing*. Springer, 117–127.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'08) (IEEE World Congress on Computational Intelligence)*. IEEE, 1322–1328.

Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 21, 9 (2009), 1263–1284.

Nic Herndon and Doina Caragea. 2016. A study of domain adaptation classifiers derived from logistic regression for the task of splice site prediction. *IEEE Trans. Nanobiosci.* 15, 2 (2016), 75–83.

Victoria Hodge and Jim Austin. 2004. A survey of outlier detection methodologies. *Artificial Intell. Rev.* 22, 2 (2004), 85–126.

Yi-Min Huang and Shu-Xin Du. 2005. Weighted support vector machine for classification with uneven training class sizes. In *Proceedings of 2005 International Conference on Machine Learning and Cybernetics*, Vol. 7. IEEE, 4365–4369.

Jae Pil Hwang, Seongkeun Park, and Euntai Kim. 2011. A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function. *Expert Syst. Appl.* 38, 7 (2011), 8580–8585.

Ilnaz Jamali, Mohammad Bazmara, and Shahram Jafari. 2012. Feature selection in imbalance data sets. *Int. J. Comput. Sci. Iss.* 9, 3 (2012), 42–45.

Piyasak Jeatrakul, Kok Wai Wong, and Chun Che Fung. 2010. Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. In *Proceedings of the International Conference on Neural Information Processing*. Springer, 152–159.

Qi Kang, XiaoShuang Chen, SiSi Li, and MengChu Zhou. 2016. A noise-filtered under-sampling scheme for imbalanced classification. *IEEE Trans. Cybernet.* (2016).

Masanori Kasai and Yusuke Oike. 2010. Image pickup apparatus, image processing method, and computer program capable of obtaining high-quality image data by controlling imbalance among sensitivities of light-receiving devices. U.S. Patent 7,839,437.

Madian Khabsa, Ahmed Elmagarmid, Ihab Ilyas, Hossam Hammady, and Mourad Ouzzani. 2016. Learning to identify relevant studies for systematic reviews using random forest and external information. *Mach. Learn.* 102, 3 (2016), 465–482.

Salman H. Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A. Sohel, and Roberto Togneri. 2017. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Trans. Neural Netw. Learn. Syst.* (2017).

Taghi M. Khoshgoftaar, Kehan Gao, Amri Napolitano, and Randall Wald. 2014. A comparative study of iterative and non-iterative feature selection techniques for software defect prediction. *Info. Syst. Front.* 16, 5 (2014), 801–822.

Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2010. Supervised neural network modeling: An empirical investigation into learning from imbalanced data with labeling errors. *IEEE Trans. Neural Netw.* 21, 5 (2010), 813–830.

Gitae Kim, Bongsug Kevin Chae, and David L. Olson. 2013. A support vector machine (SVM) approach to imbalanced datasets of customer responses: Comparison with other customer response models. *Service Bus.* 7, 1 (2013), 167–182.

Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. 2006. Handling imbalanced datasets: A review. *GESTS Int. Trans. Comput. Sci. Eng.* 30, 1 (2006), 25–36.

Bartosz Krawczyk. 2016. Learning from imbalanced data: Open challenges and future directions. *Progr. Artific. Intell.* 5, 4 (2016), 221–232.

Bartosz Krawczyk, Mikel Galar, Łukasz Jeleń, and Francisco Herrera. 2016. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Appl. Soft Comput.* 38 (2016), 714–726.

Bartosz Krawczyk, Michał Woźniak, and Gerald Schaefer. 2014. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Appl. Soft Comput.* 14 (2014), 554–562.

Miroslav Kubat, Stan Matwin, et al. 1997. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the International Conference on machine Learning (ICML'97)*, Vol. 97. 179–186.

Pallavi Kulkarni and Roshani Ade. 2016. Logistic regression learning model for handling concept drift with unbalanced data in credit card fraud detection system. In *Proceedings of the 2nd International Conference on Computer and Communication Technologies*. Springer, 681–689.

Taehyung Lee, Ki Bum Lee, and Chang Ouk Kim. 2016. Performance of machine-learning algorithms for class-imbalanced process fault detection problems. *IEEE Trans. Semicond. Manufact.* 29, 4 (2016), 436–445.

Boaz Lerner, Josepha Yeshaya, and Lev Koushnir. 2007. On the classification of a small imbalanced cytogenetic image database. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 4, 2 (2007).

Miao Liu, Mingjun Wang, Jun Wang, and Duo Li. 2013. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. *Sens. Actuat. B: Chem.* 177 (2013), 970–980.

Ying Liu, Han Tong Loh, and Aixin Sun. 2009. Imbalanced text classification: A term weighting approach. *Expert Syst. Appl.* 36, 1 (2009), 690–701.

Zhen Liu, Ruoyu Wang, Ming Tao, and Xianfa Cai. 2015. A class-oriented feature selection approach for multi-class imbalanced network traffic datasets based on local and global metrics fusion. *Neurocomputing* 168 (2015), 365–381.

Rushi Longadge and Snehalata Dongre. 2013. Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707* (2013).

Victoria López, Sara del Río, José Manuel Benítez, and Francisco Herrera. 2015. Cost-sensitive linguistic fuzzy rule-based classification systems under the MapReduce framework for imbalanced big data. *Fuzzy Sets Syst.* 258 (2015), 5–38.

Yang Lu, Yiu-ming Cheung, and Yuan Yan Tang. 2016. Hybrid sampling with bagging for class imbalance learning. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 14–26.

Abdul Majid, Safdar Ali, Mubashar Iqbal, and Nabeela Kausar. 2014. Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines. *Comput. Methods Programs Biomed.* 113, 3 (2014), 792–808.

Sebastián Maldonado, Richard Weber, and Fazel Famili. 2014. Feature selection for high-dimensional class-imbalanced data sets using support vector machines. *Info. Sci.* 286 (2014), 228–246.

Shahla Mardani and Hamid Reza Shahriari. 2013. A new method for occupational fraud detection in process aware information systems. In *Proceedings of the 10th International ISC Conference on Information Security and Cryptology (ISCISC'13)*. IEEE, 1–5.

Stephen O. Moepya, Sharat S. Akhoury, and Fulufhelo V. Nelwamondo. 2014. Applying cost-sensitive classification for financial fraud detection under high class-imbalance. In *Proceedings of the IEEE International Conference on Data Mining Workshop (ICDMW'14)*. IEEE, 183–192.

Jose G. Moreno-Torres, Xavier Llorà, David E. Goldberg, and Rohit Bhargava. 2013. Repairing fractures between data using genetic programming-based feature extraction: A case study in cancer diagnosis. *Info. Sci.* 222 (2013), 805–823.

Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. 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, 805–808.

Surya Nepal and Mukaddim Pathan. 2014. *Security, Privacy and Trust in Cloud Systems*. Springer.

Hien M. Nguyen, Eric W. Cooper, and Katsuari Kamei. 2012. A comparative study on sampling techniques for handling class imbalance in streaming data. In *Proceedings of the Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS'12), 13th International Symposium on Advanced Intelligent Systems (ISIS'12)*. IEEE, 1762–1767.

Ana Palacios, Krzysztof Trawiński, Oscar Cordón, and Luciano Sánchez. 2014. Cost-sensitive learning of fuzzy rules for imbalanced classification problems using FURIA. *Int. J. Uncertain. Fuzz. Knowl.-based Syst.* 22, 05 (2014), 643–675.

Jiyan Pan, Quanfu Fan, Sharath Pankanti, Hoang Trinh, Prasad Gabbur, and Sachiko Miyazawa. 2011. Soft margin keyframe comparison: Enhancing precision of fraud detection in retail surveillance. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV'11)*. IEEE, 549–556.

Husanbir Singh Pannu and Harsurinder Kaur. 2017. Anomaly detection survey for information security. In *Proceedings of the 10th International Conference on Security of Information and Networks*. ACM, 251–258.

Husanbir S. Pannu, Jianguo Liu, Qiang Guan, and Song Fu. 2012. AFD: Adaptive failure detection system for cloud computing infrastructures. In *Proceedings of the IEEE 31st International Performance Computing and Communications Conference (IPCCC'12)*. IEEE, 71–80.

Yubin Park and Joydeep Ghosh. 2014. Ensembles of alpha-trees for imbalanced classification problems. *IEEE Trans. Knowl. Data Eng.* 26, 1 (2014), 131–143.

Harshita Patel and G. S. Thakur. 2016. A hybrid weighted nearest neighbor approach to mine imbalanced data. In *Proceedings of the International Conference on Data Mining (DMIN'16)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 106.

Naser Peiravian and Xingquan Zhu. 2013. Machine learning for Android malware detection using permission and API calls. In *Proceedings of the IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI'13)*. IEEE, 300–305.

Lizhi Peng, Bo Yang, Yuehui Chen, and Xiaoqing Zhou. 2016. An under-sampling imbalanced learning of data gravitation-based classification. In *Proceedings of the 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD'16)*. IEEE, 419–425.

Yun Qian, Yanchun Liang, Mu Li, Guoxiang Feng, and Xiaohu Shi. 2014. A resampling ensemble algorithm for classification of imbalance problems. *Neurocomputing* 143 (2014), 57–67.

Chen Qiu, Liangxiao Jiang, and Chaoqun Li. 2017. Randomly selected decision tree for test-cost sensitive learning. *Appl. Soft Comput.* 53 (2017), 27–33.

D. Ramyachitra and P. Manikandan. 2014. Imbalanced dataset classification and solutions: A review. *Int. J. Comput.ing and Bus. Res.* 5, 4 (2014).

K. Usha Rani, G. Naga Ramadevi, and D. Lavanya. 2016. Performance of synthetic minority oversampling technique on imbalanced breast cancer data. In *Proceedings of the 3rd International Conference on Computing for Sustainable Global Development (INDIACom'16)*. IEEE, 1623–1627.

Bhavani Raskutti and Adam Kowalczyk. 2004. Extreme re-balancing for SVMs: A case study. *ACM SIGKDD Explor. Newslett.* 6, 1 (2004), 60–69.

Alice M. Richardson and Brett A. Lidbury. 2017. Enhancement of hepatitis virus immunoassay outcome predictions in imbalanced routine pathology data by data balancing and feature selection before the application of support vector machines. *BMC Med. Info. Decis. Mak.* 17, 1 (2017), 121.

Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 19 (2007), 2507–2517.

Mahendra Sahare and Hitesh Gupta. 2012. A review of multi-class classification for imbalanced data. *Int. J. Adv. Comput. Res.* 2, 3 (2012), 160–164.

Yusuf Sahin, Serol Bulkan, and Ekrem Duman. 2013. A cost-sensitive decision tree approach for fraud detection. *Expert Syst. Appl.* 40, 15 (2013), 5916–5923.

Claude Sammut and Geoffrey I. Webb. 2011. *Encyclopedia of Machine Learning*. Springer Science & Business Media.

José Antonio Sanz, Dario Bernardo, Francisco Herrera, Humberto Bustince, and Hani Hagras. 2015. A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data. *IEEE Trans. Fuzzy Syst.* 23, 4 (2015), 973–990.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J. Biomed. Info.* 53 (2015), 196–207.

Asaf Shabtai, Robert Moskovitch, Clint Feher, Shlomi Dolev, and Yuval Elovici. 2012. Detecting unknown malicious code by applying classification techniques on opcode patterns. *Secur. Info.* 1, 1 (2012), 1.

Yuan-Hai Shao, Wei-Jie Chen, Jing-Jing Zhang, Zhen Wang, and Nai-Yang Deng. 2014. An efficient weighted Lagrangian twin support vector machine for imbalanced data classification. *Pattern Recogn.* 47, 9 (2014), 3158–3167.

Mei-Ling Shyu, Zongxing Xie, Min Chen, and Shu-Ching Chen. 2008. Video semantic event/concept detection using a subspace-based multimedia data mining framework. *IEEE Trans. Multimedia* 10, 2 (2008), 252–259.

Arpit Singh and Anuradha Purohit. 2015. A survey on methods for solving data imbalance problem for classification. *Work* 127, 15 (2015).

Li Song, Dapeng Li, Xiangxiang Zeng, Yunfeng Wu, Li Guo, and Quan Zou. 2014. nDNA-prot: Identification of DNA-binding proteins based on unbalanced classification. *BMC Bioinformat.ics* 15, 1 (2014), 298.

Qun Song, Jun Zhang, and Qian Chi. 2010. Assistant detection of skewed data streams classification in cloud security. In *Proceedings of the IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS'10)*, Vol. 1. IEEE, 60–64.

Robert A. Sowah, Moses A. Agebure, Godfrey A. Mills, Koudjo M. Koumadi, and Seth Y. Fiawoo. 2016. New cluster under-sampling technique for class imbalance learning. *Int. J. Mach. Learn. Comput.* 6, 3 (2016), 205.

Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. 2009. Classification of imbalanced data: A review. *Int. J. Pattern Recogn. Artific. Intell.* 23, 04 (2009), 687–719.

Zhongbin Sun, Qinbao Song, Xiaoyan Zhu, Heli Sun, Baowen Xu, and Yuming Zhou. 2015. A novel ensemble method for classifying imbalanced data. *Pattern Recogn.* 48, 5 (2015), 1623–1637.

Mayank Taneja, Kavyanshi Garg, Archana Purwar, and Samarth Sharma. 2015. Prediction of click frauds in mobile advertising. In *Proceedings of the 8th International Conference on Contemporary Computing (IC3'15)*. IEEE, 162–166.

Bo Tang, Haibo He, Paul M. Baggenstoss, and Steven Kay. 2016. A Bayesian classification approach using class-specific features for text categorization. *IEEE Trans. Knowl. Data Eng.* 28, 6 (2016), 1602–1606.

Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, and Sven Krasser. 2009. SVMs modeling for highly imbalanced classification. *IEEE Trans. Syst., Man, Cybernet., Part B (Cybernet.)* 39, 1 (2009), 281–288.

Ciza Thomas. 2013. Improving intrusion detection for imbalanced network traffic. *Secur. Commun. Netw.* 6, 3 (2013), 309–324.

Jason Van Hulse, Taghi M. Khoshgoftaar, Amri Napolitano, and Randall Wald. 2009. Feature selection with high-dimensional imbalanced data. In *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW'09)*. IEEE, 507–514.

Nguyen Ha Vo and Yonggwan Won. 2007. Classification of unbalanced medical data with weighted regularized least squares. In *Proceedings of the Conference on Frontiers in the Convergence of Bioscience and Information Technologies (FBIT'07)*. IEEE, 347–352.

Chi-Man Vong, Jie Du, Chi-Man Wong, and Jiu-Wen Cao. 2018. Postboosting using extended G-Mean for online sequential multiclass imbalance learning. *IEEE Trans. Neural Netw. Learn. Syst.* (2018).

Shixiang Wan, Yucong Duan, and Quan Zou. 2017. HPSLPred: An ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. *Proteomics* 17, 17–18 (2017), 1700262.

C. Wang, L. Hu, M. Guo, X. Liu, and Q. Zou. 2015. imDC: An ensemble learning method for imbalanced classification with miRNA data. *Genet. Mol. Res.* 14, 1 (2015), 123–133.

Qiang Wang. 2014. A hybrid sampling SVM approach to imbalanced data classification. In *Abstract and Applied Analysis*, Vol. 2014. Hindawi Publishing Corporation.

Suge Wang, Deyu Li, Lidong Zhao, and Jiahao Zhang. 2013. Sample cutting method for imbalanced text sentiment classification based on BRC. *Knowl.-Based Syst.* 37 (2013), 451–461.

Shuo Wang and Xin Yao. 2013. Using class imbalance learning for software defect prediction. *IEEE Trans. Reliabil.* 62, 2 (2013), 434–443.

Wei Wei, Jinjiu Li, Longbing Cao, Yuming Ou, and Jiahang Chen. 2013a. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web* 16, 4 (2013), 449–475.

Wei Wei, Jinjiu Li, Longbing Cao, Yuming Ou, and Jiahang Chen. 2013b. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web* 16, 4 (2013), 449–475.

Qingyao Wu, Yunming Ye, Haijun Zhang, Michael K. Ng, and Shen-Shyang Ho. 2014. ForesTexter: An efficient random forest algorithm for imbalanced text categorization. *Knowl.-Based Syst.* 67 (2014), 105–116.

Yufei Xia, Chuanzhe Liu, and Nana Liu. 2017. Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electron. Comm. Res. Appl.* 24 (2017), 30–49.

Jieming Yang, Zhaoyang Qu, and Zhiying Liu. 2014. Improved feature-selection method considering the imbalance problem in text categorization. *Sci. World J.* 2014 (2014).

Junshan Yang, Jiarui Zhou, Zexuan Zhu, Xiaoliang Ma, and Zhen Ji. 2016. Iterative ensemble feature selection for multiclass classification of imbalanced microarray data. *J. Biol. Res. Thessaloniki* 23, 1 (2016), 13.

Bee Wah Yap, Khatijahhusna Abd Rani, Hezlin Aryani Abd Rahman, Simon Fong, Zuraida Khairudin, and Nik Nik Abdullah. 2014. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In *Proceedings of the 1st International Conference on Advanced Data and Information Engineering (DaEng'13)*. Springer, 13–22.

Hualong Yu and Jun Ni. 2014. An improved ensemble learning method for classifying high-dimensional and imbalanced biomedicine data. *IEEE/ACM Trans. Comput. Biol. Bioinformat.* 11, 4 (2014), 657–666.

Hualong Yu, Jun Ni, and Jing Zhao. 2013. ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. *Neurocomputing* 101 (2013), 309–318.

Ashkan Zakaryazad and Ekrem Duman. 2016. A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing. *Neurocomputing* 175 (2016), 121–131.

Jia Zeng, Shanfeng Zhu, and Hong Yan. 2009. Towards accurate human promoter recognition: A review of currently used sequence features and classification methods. *Brief. Bioinformat.* 10, 5 (2009), 498–508.

Bin Zhang, Yi Zhou, and Christos Faloutsos. 2008. Toward a comprehensive model in internet auction fraud detection. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*. IEEE, 79–79.

Dongmei Zhang, Jun Ma, Jing Yi, Xiaofei Niu, and Xiaojing Xu. 2015. An ensemble method for unbalanced sentiment classification. In *Proceedings of the 11th International Conference on Natural Computation (ICNC'15)*. IEEE, 440–445.

Huaxiang Zhang and Mingfang Li. 2014. RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. *Info. Fusion* 20 (2014), 99–116.

Yan-Ping Zhang, Li-Na Zhang, and Yong-Cheng Wang. 2010. Cluster-based majority under-sampling approaches for class imbalance learning. In *Proceedings of the 2nd IEEE International Conference on Information and Financial Engineering (ICIFE'10)*. IEEE, 400–404.

Zhancheng Zhang, Jun Dong, Xiaoqing Luo, Kup-Sze Choi, and Xiaojun Wu. 2014. Heartbeat classification using disease-specific feature selection. *Comput. Biol. Med.* 46 (2014), 79–89.

Xing-Ming Zhao, Xin Li, Luonan Chen, and Kazuyuki Aihara. 2008. Protein classification with imbalanced data. *Proteins: Struct. Funct. Bioinformat.* 70, 4 (2008), 1125–1132.

Zhuoyuan Zheng, Yunpeng Cai, and Ye Li. 2016. Oversampling method for imbalanced classification. *Comput. Info.* 34, 5 (2016), 1017–1037.

Weicai Zhong, Bijan Raahemi, and Jing Liu. 2013. Classifying peer-to-peer applications using imbalanced concept-adapting very fast decision tree on IP data stream. *Peer-to-Peer Netw. Appl.* 6, 3 (2013), 233–246.

Maciej Zięba, Jakub M. Tomczak, Marek Lubicz, and Jerzy Świątek. 2014. Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Appl. Soft Comput.* 14 (2014), 99–108.

Quan Zou, Sifa Xie, Ziyu Lin, Meihong Wu, and Ying Ju. 2016. Finding the best classification threshold in imbalanced classification. *Big Data Res.* 5 (2016), 2–8.

Received March 2018; revised October 2018; accepted May 2019