

**Please cite the Published Version**

Patel, H, Singh Rajput, D, Thippa Reddy, G, Iwendi, C, Kashif Bashir, A and Jo, O (2020) A review on classification of imbalanced data for wireless sensor networks. *International Journal of Distributed Sensor Networks*, 16 (4). ISSN 1550-1329

**DOI:** <https://doi.org/10.1177/1550147720916404>

**Publisher:** Sage

**Version:** Published Version

**Downloaded from:** <https://e-space.mmu.ac.uk/625902/>


**Usage rights:**  [Creative Commons: Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)

**Additional Information:** This is an Author Accepted Manuscript of a paper accepted for publication in *International Journal of Distributed Sensor Networks*, published by and copyright Sage.

**Enquiries:**

If you have questions about this document, contact [rsl@mmu.ac.uk](mailto:rsl@mmu.ac.uk). Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

# A review on classification of imbalanced data for wireless sensor networks

International Journal of Distributed  
Sensor Networks  
2020, Vol. 16(4)  
© The Author(s) 2020  
DOI: 10.1177/1550147720916404  
journals.sagepub.com/home/dsn  


Harshita Patel<sup>1</sup>, Dharmendra Singh Rajput<sup>1</sup>, G Thippa Reddy<sup>1</sup> ,  
Celestine Iwendi<sup>2</sup> , Ali Kashif Bashir<sup>3</sup>  and Ohyun Jo<sup>4</sup> 

## Abstract

Classification of imbalanced data is a vastly explored issue of the last and present decade and still keeps the same importance because data are an essential term today and it becomes crucial when data are distributed into several classes. The term imbalance refers to uneven distribution of data into classes that severely affects the performance of traditional classifiers, that is, classifiers become biased toward the class having larger amount of data. The data generated from wireless sensor networks will have several imbalances. This review article is a decent analysis of imbalance issue for wireless sensor networks and other application domains, which will help the community to understand WHAT, WHY, and WHEN of imbalance in data and its remedies.

## Keywords

Wireless sensor networks, data mining, imbalanced data, data balancing, algorithm modification, ensemble techniques

Date received: 6 January 2020; accepted: 4 March 2020

Handling Editor: Mohamed Elhoseny

## Introduction

One of the important challenges in data mining is handling of imbalanced data in classification.<sup>1-4</sup> We know that classification is an important technique of data mining, in which unknown class samples are assigned to some class based on previous knowledge from training samples.<sup>5,6</sup> Imbalance appears when data are unequally distributed into classes; some classes may have large quantity of data called as majority classes and some may have just few instances of data called minority classes. This uneven distribution causes biased performance of traditional classifiers because they consider the error rate not the distribution of data, and due to having little quantity of data instances, minority classes get ignored in overall classification result. This issue appears in many real-world applications,<sup>7</sup> such as healthcare sector,<sup>8,9</sup> detection of oil spill,<sup>10</sup> fraud detection in usage of credit cards,<sup>11</sup> modeling of cultures,<sup>12</sup>

intrusion detection in networks, categorization of texts, and so on. Figure 1 is a representational picture of imbalanced data.

<sup>1</sup>School of Information Technology & Engineering, Vellore Institute of Technology, Vellore, India

<sup>2</sup>Department of Electronics, BCC of Central South University of Forestry and Technology, Changsha, China

<sup>3</sup>Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, UK

<sup>4</sup>Department of Computer Science, Chungbuk National University, Cheongju-si, South Korea

### Corresponding authors:

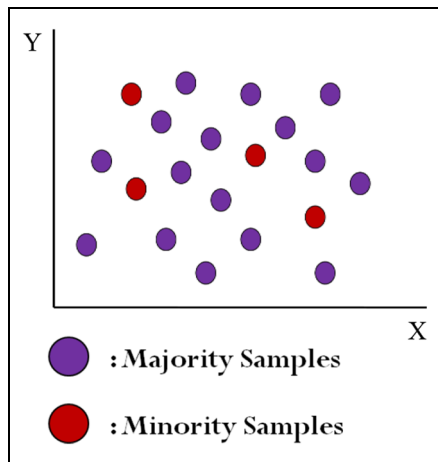
Ohyun Jo, Chungbuk National University, Cheongju-si 28644, South Korea.

Email: ohyunjo@chungbuk.ac.kr

Dharmendra Singh Rajput, School of Information Technology & Engineering, Vellore Institute of Technology, Vellore 632014, India.

Email: dharmendrasingh@vit.ac.in





**Figure 1.** Imbalance in a binary dataset.

Many solutions are proposed to solve this issue in previous years in different ways. In this regard, high quality of review papers have been published in the last decade on imbalanced data and their related aspects, containing information starting from the definition of imbalance in data space including the characteristics, types, and effect on classification performance to all possible ways to deal with the issue. These review and research articles are the valuable sources of knowledge to understand the problem of imbalance comprehensively. Thorough literature survey is bringing out the enormous study and research on imbalance of data. Following are some examples of some popular research areas dealing with this natural case of data distribution.

### Wide ranges of applications

Learning from imbalanced data was mainly driven by numerous applications in real life where we have to deal with the problem of incompatible data representation. The minority class is usually the most important in such cases, and therefore, we need methods to improve their recognition rates. This is closely linked with major issues like preventing malicious attacks, wireless sensor networking, detecting life-threatening diseases,

managing atypical behavior in social networks, or handling rare cases in monitoring systems.

A list of selected, recent real-life applications presenting data imbalance.

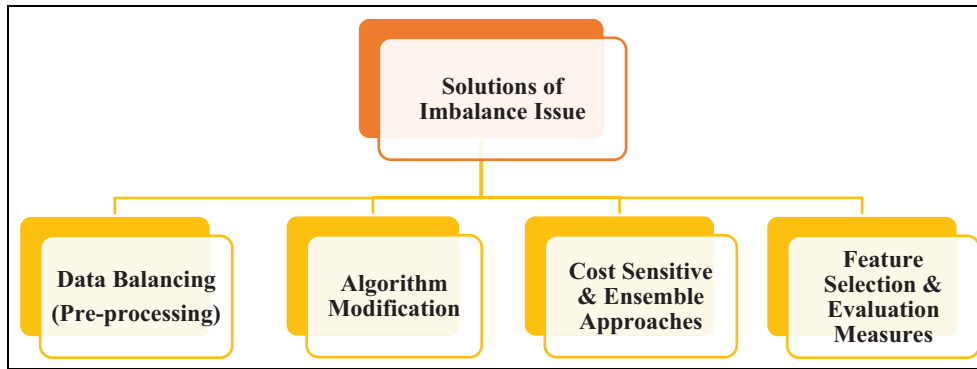
This article analyzes review or survey papers followed by original research articles that are finding potential solutions to the problem in specific manners. Especially, nearest neighbor and their fuzzy versions are considered for discussion. This work suggests several techniques that can be used to balance the imbalances generated by sensors in wireless sensor network (WSN).

Following sections of the article are organized to explain the literature from basics of imbalance to solutions offered to the problem recently, and some wide range of applications in this area has discussed. Section “Literature review on handling imbalanced data at a glance” provides the nuggets of imbalance learning literature, from some of the important base papers. Section “Review of solutions to the problem” concentrates on the discussion of various solutions to the imbalance problem, that is, rebalancing of data, algorithm modification, and so on in its subsections. Section “Imbalanced data in wireless sensor networks” discusses about the imbalanced data and their effect on analyzing WSN data. Section “Lessons learned and approaches suggested for handling imbalanced data in WSN” discusses briefly about the suggestions to improve imbalanced data. Finally, the conclusion and future direction are discussed in section “Conclusion and future direction.”

### Literature review on handling imbalanced data at a glance

Chawla et al.<sup>21</sup> provide an overview of the imbalance in an editorial issue that consists of the information about imbalance available to date with existing solutions and the extracts of important workshops and conferences organized that time by knowing the criticalness of the problem. They reviewed some of the research articles too. The purpose of this editorial was to ensure the awareness among the data mining community about imbalance present in datasets as its natural in most cases.

Application area	Problem description	Classes in datasets
Behavior analysis <sup>13</sup>	Recognition of dangerous behavior	Binary
Sentiment analysis <sup>14–16</sup>	Emotion and temper recognition in text	Binary and multi-class
Text mining <sup>17</sup>	Detecting relations in literature	Binary
Video mining <sup>18</sup>	Recognizing objects and actions in video sequences	Binary and multi-class
Cancer malignancy grading <sup>19</sup>	Analyzing the cancer severity	Binary and multi-class
WSN <sup>20</sup>	Analyzing data generated by WSN	Binary and multi-class



**Figure 2.** Strategic solutions for classification of imbalanced data.

Visa and Ralescu<sup>22</sup> attempted in a similar way to analyze workshops and conference held in past on imbalanced datasets and came out with the discussion on imbalanced datasets, research gaps, and future research directions. Guo et al.<sup>23</sup> have presented a review on approaches proposed for imbalanced learning on four different levels. They have seen about the evaluation measures for such datasets and relate the other issues to conclude that other factors are also responsible for degradation in classifiers' performance with imbalances such as small disjuncts.

The review article written by He and Garcia.<sup>2</sup> presented an analysis that becomes a milestone for the researchers of imbalanced datasets for comprehensive knowledge of imbalance issue from elementary definitions of the terms to state-of-the-art solutions and particular evaluation measure. Also, possible future research directions and scope were discussed brilliantly.

Review of the classification of imbalanced data by Fernández et al.<sup>24</sup> was aimed on two important matters: first, the discussion of solutions of the problem with pre-processing and cost-sensitive approaches was provided, and then, a look was given on intrinsic data properties that affect the classifiers' performance for data imbalance such as existence of small disjuncts, small sample size, class overlapping, and so on.

López et al.<sup>25</sup> have discussed the issue of imbalanced learning with focusing on two targets: first, the attention given on pre-processing, cost-sensitive, and ensemble learning approaches for imbalanced datasets with experimental examples and second, the intrinsic characteristics put on the light that play important role in classification, that is, presence of small disjuncts, small sample size or low density, class overlap, noisy data, borderline examples, and data shift.

Prati et al.<sup>26</sup> has designed an experimental setup to evaluate classifiers' performances for various degrees of imbalance and came to the conclusion that degree of imbalance is proportionate to the performance of

classifiers, implying that higher misclassification for higher degree of imbalance and vice versa. They have also introduced a confidence interval method to judge the performance of such classifiers. They found that existing remedies for imbalanced datasets are partially able to resolve the issue.

## Review of solutions to the problem

The issue of imbalance is directly related to distribution of data into classes, and therefore, classifiers have to compromise with the performances. Literature suggests that imbalance in the data can be dealt intrinsically by either balancing the data, also known as re-sampling, or pre-processing of data and then applying traditional classifiers or to modify classifier to find correct classification results from imbalanced data. In the first approach, data pre-processing or re-sampling is applied on data so this could be known as "Data-Level Solution" approach; here, only data are altered, and no changes are performed in classifiers. Other strategy considers natural distribution of data as it is and classifiers are modified for the specific case of imbalance. This approach performs changes in classification algorithms so could be termed as "Algorithmic-Level Solution." Other known strategies are "Cost-Sensitive Approaches" and "Ensemble Techniques." Also, imbalance is dealt with feature selection and evolutionary approaches. Subsequent sections of this article contain some good and notable contributions of researchers in all categories of solutions followed by the review of related articles for this article on nearest neighbor and its variants. Following Figure 2 shows the four solutions of imbalance classification:

### Balancing of data or re-sampling pre-processing

Batista et al.<sup>27</sup> have evaluated the behavior of different methods for dealing with oversampling and undersampling in learning from imbalanced data and suggested that the sampling methods perform well on different

cases of imbalance. Also, they proposed two oversampling approaches, and both are good in extracting results from datasets having a less number of positive examples.

Similarly, García et al.<sup>28</sup> analyzed the imbalance ratio effect and classifier properties on various re-sampling approaches. They concluded through experiments that for a low imbalance ratio, performance of oversampling as well as undersampling approaches is equivalent but for high imbalanced cases, oversampling should be preferred for better classification. Also, they found that the influence of classifiers on efficiency of re-sampling strategies is negligible.

A systematic review on the class imbalance issue is done by Menardi and Torelli.<sup>29</sup> They discussed that how various existing classifiers are failing in learning from imbalanced datasets. They emphasized the need of model estimation and model evaluation with refined measures specifically for such skewed environment. Also one re-sampling method is proposed in this research that is leading to boosting and bagging and improves the accuracy estimation in severe imbalanced situations.

Chawla et al.<sup>30</sup> proposed a novel technique, synthetic minority oversampling technique (SMOTE), which is remarkable research in the area of oversampling used in many applications. They used the feature-based similarity to generate synthetic examples among minority examples. This method makes traditional classifier to enhance the decision boundary close to minority examples.

He et al.<sup>31</sup> proposed ADASYN approach using the weighted distribution of different minority class examples on the basis of their difficulty level of learning. The minority class instances that are harder to learn expect more synthetic instance generation in comparison to easier minority examples, so this approach reduces the biasness by shifting the decision boundary aiming the difficult minority instances.

Zhang and Li.<sup>32</sup> have evaluated the effect of oversampling on three traditional classifiers and concluded that the oversampling significantly influences the performances of classifiers and, thus, is helpful in classification of imbalanced datasets. They have also proposed random walk oversampling approach that takes less time to generate synthetic samples than SMOTE because from standard deviation and RWO, mean is calculated by the use of minority data.

Wang et al.<sup>33</sup> proposed a combination of an oversampling SMOTE, an optimization technique particle swarm optimization (PSO), and classifiers C5, 1-nearest neighbor, and linear regression that improves performance of the classification over imbalanced dataset of breast cancer survival for 5 years. They found that the hybridization of SMOTE, PSO, and C5 brought best results for this case.

An extension of SMOTE named SMOTE-IPF is proposed by Sáez et al.<sup>34</sup> that considers borderline example and noise as important factors with class imbalance. They used the iterative partitioning; IPF noise filter to handle the SMOTE-generated noise occurred in creating synthetic data. This approach makes boundaries of classes clearer.

An inverse random undersampling strategy is proposed by Tahir et al.,<sup>35</sup> and in this method, inverse undersampling is performed to get many training datasets. Then, for an individual training set, a decision boundary is identified that separates the minority class from the majority class. This leads to decide a common and complex boundary in the combination phase; this is very applicable in multi-class classification.

Wong et al.<sup>36</sup> proposed an undersampling approach for large imbalanced datasets, where fuzzy logic is used to select samples from the majority class. Then, an evolutionary computational model of cross-generational elitist selection, heterogeneous recombination, and cataclysmic mutation (CHC) is employed to shrink the majority class by undersampling. Classification of this modified datasets is then performed by support vector machine (SVM).

So data balancing techniques alter the original distribution of data to achieve better classification for imbalanced datasets. Various sampling strategies are used to balance the data, either to undersample large class or oversample the small one or to use the combination of both. Table 1 provides a tabular look of these data balancing techniques with their key features.

### Algorithm modification

Xu et al.<sup>37</sup> extended the I-algorithm to E-algorithm to work with data imbalance. The “extended” and fuzzy rule induction is modified for imbalanced datasets. Comparison is done for I-algorithm and also E-algorithm by applying them to Duke Energy outage. The extended algorithm performs better for the majority and minority classes.

Antonelli et al.<sup>38</sup> performed comparison on three evolutionary fuzzy rule base classification approaches (EFC) for imbalanced datasets. First approach is an embedded feature selection and granularity learning that included the rule base generation method. The second EFC is an algorithm for genetic programming that creates rule base for the hierarchical fuzzy rule base classifier. The third EFC proposed by authors is a multi-objective evolutionary algorithm that is enhanced so that rule base and membership function parameters could be run simultaneously for a set of fuzzy rule base classifiers. The comparison summed up that the third approach is the best strategy for imbalanced classification.

García et al.<sup>39</sup> proposed evolutionary generalized instance selection by CHC (EGIS-CHC), an evolutionary based nested example learning approach to classify

**Table 1.** Data balancing algorithms.

S. No.	Authors	Title	Year	Key feature
1	Batista et al. <sup>27</sup>	A study of the behaviour of several methods of balancing machine learning training data	2004	Review article
2	V García et al. <sup>28</sup>	On the effectiveness of preprocessing methods when dealing with different levels of class imbalance	2012	Review article
3	G Menardi and N Torelli <sup>29</sup>	Training and assessing classification rules with imbalanced data	2014	Review article and proposed “ROSE” based on smoothed bootstrap of re-sampled data
4	Chawla et al. <sup>30</sup>	SMOTE: Synthetic minority over-sampling technique	2002	Feature-based similarity is used for generating synthetic examples among minority class instances
5	H He et al. <sup>31</sup>	ADASYN: Adaptive synthetic sampling approach for imbalance learning	2008	Weighted distribution of different minority class examples is used on the basis of their difficulty level of learning
6	H Zhang and M Li <sup>32</sup>	RWO Sampling: A random walk oversampling approach to imbalanced data classification	2014	Mean and standard deviation are used to generate synthetic samples for the minority class
7	Wang et al. <sup>33</sup>	A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients	2014	Hybridization of SMOTE, an optimization technique PSO, and classifier to improve learning from imbalanced data
8	JA Sáez et al. <sup>34</sup>	SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering	2015	Borderline examples and noise are handled using IPF filter generated during SMOTE processing
9	MA Tahir et al. <sup>35</sup>	Inverse random under sampling for class imbalance problem and its application to multi-label classification	2012	Inverse random undersampling is applied to form multiple training sets and then identify decision boundaries for each set to separate minority and majority classes
10	Wong et al. <sup>36</sup>	An under-sampling method based on fuzzy logic for large imbalanced dataset	2014	Fuzzy logic is used for sample selection; then, undersampling is performed by an evolutionary CHC model.

PSO: particle swarm optimization; SMOTE: synthetic minority oversampling technique. ROSE: random over sampling examples; RWO: random walk over; IPF: iterative-partitioning filter; CHC: cross-generational elitist selection, heterogenous recombination, and cataclysmic mutation.

imbalanced data more appropriately. Accumulation of instances in Euclidean n-space is done to perform learning of nested examples in this method.

*Improved K-nearest neighbor algorithms.* AS-KNN approach is proposed by Yang et al.<sup>40</sup> For event tracking, they modified several nearest neighbor methods and then divided the sum of similarity in every each class with the total number of instances of every class from nearest neighbors. Then, the combination of nearest neighbor approaches reduces the performance variation for the event tracking system.

Different K values have been chosen based on adaptive K-nearest neighbor algorithm that was proposed by Baoli et al.<sup>41</sup> This approach is proposed to find correct categories in text data which many have imbalanced text corpus.

Tan<sup>42</sup> proposed a neighbor weighted nearest neighbor strategy to classify imbalanced text data. In this

approach, small weights are allocated to the class having more instances, that is, the majority classes and large weights are assigned to small or minority classes having comparatively less number of data instances. This strategy balances the weightage of instances from both classes.

DragPushing for KNN (DP-KNN) is another method proposed by Tan.<sup>43</sup> In this method, features’ weights are increased or decreased to deal with misclassified data for that training errors are used to improve the KNN by drag and push operations using weights.

Wang et al.<sup>44</sup> presented a K-nearest neighbor method evidence theory. They brought new concepts of global frequency and local frequency estimation, in short GE and LE. This approach deals with class imbalance to a certain level without using re-sampling.

Liu and Chawla<sup>45</sup> suggested an improvement in the weighted K-nearest neighbor strategy for classification of imbalanced datasets. They introduced class

confidence weights to find posterior probabilities for this issue. Class confidence weights were determined using mixture modeling and Bayesian networks.

DCM KNN<sup>46</sup> applies the traditional nearest neighbor approach on training data and decomposes them into misclassified and correctly classified data and finds a suitable nearest neighbor method for these sets. And for test data, this approach checks that to which set test instances will belong, misclassified or correctly classified, and then applies the appropriate nearest neighbor method.

Kriminger et al.<sup>47</sup> used the local geometric structure in data in a class algorithm named class conditional nearest neighbor distribution that diminishes the imbalance effect exist in data. The algorithm can be applied for different degrees of imbalance and also to perform with any number of classes. This approach facilitates to add new instances of training datasets.

Dubey and Pudi<sup>48</sup> proposed a KNN-based approach which considered distribution of classes for neighbors of any test instance. Initially, classification is performed by the KNN algorithm, and then, it is used to calculate the weight for all classes. This weighing strategy improves the classification performance for imbalanced data.

KNN's hubness effect is discussed by Tomašev and Mladenić<sup>49</sup> that minority class instances are the reason for the major misclassification rate in high-dimensional datasets, whereas in small and medium dimensional datasets, misclassification occurs due to majority class instances.

Ryu et al.<sup>50</sup> proposed an HISSN method to predict cross-project defect. In such cases, class imbalance exists in source and target project distributions. In this approach, the KNN algorithm is used to learn local information, and global information is gained by applying naive Bayes approach.

Ando<sup>51</sup> proposed an instance-based learning approach with a model based on mathematics that improves the performance for training data. They designed a class-based weighting strategy to deal with class imbalance, and for these weights, they proposed a convex optimization technique to find out weight parameters.

Patel and Thakur<sup>52</sup> proposed a novel approach which is a hybrid of adaptive nearest neighbor concept and neighbor weighted approach. Large K and small weights are taken for the majority class, and small K and large weights are taken for the minority classes. Table 2 provides the short description of these previously proposed improved nearest neighbor approaches with their key features.

*Fuzzy and weighted variants of K-nearest neighbor.* Prominent work has been done on fuzzy KNN approaches for

imbalanced data. In general, fuzzy concept improves the performance of nearest neighbor classifiers by finding the membership of an instance into a class for normal or balanced data, so for imbalance issue, this could be helpful to use fuzzy membership concept with some strategy to deal with imbalance. These strategies could be alteration in K or some weighing applications. Some notable contributions in this research area fuzzy logic and its variants with K-nearest neighbor are discussed here. Fernández et al.<sup>53</sup> performed an analysis of the fuzzy rule-based classification system for imbalanced environment of datasets which use an adaptive inference system. Genetic algorithms were used to learn parameters of this adaptive inference system. They applied adaptive parametric conjunction operators for varying imbalance ratio and achieved better classification outcomes. This study is the extension work of Fernández et al.,<sup>54</sup> where distinctive setups were concentrated for fuzzy rule-based classification systems keeping in mind the end goal to decide the most suitable model for imbalanced datasets. Moreover, they demonstrated the need to apply a re-sampling method; particularly, they found a decent conduct on account of the Synthetic Minority Over-Sampling Technique.

One fuzzy-rough algorithm is proposed by Han and Mao<sup>55</sup> that considers the existing fuzziness and roughness in data. They proposed a membership function in favor of the minority class to minimize the dominance of the majority class on the minority class and also defined an equivalent relation between instances of unknown classes and their nearest neighbors.

Liu et al.<sup>56</sup> proposed a fuzzy KNN approach to handle unevenly distributed categorical data that have bonds between attributes, classes, and other cases. A fuzzy-rough-based approach is proposed for KNN by Ramentol et al.<sup>57</sup> for imbalance in binary classes with six weight vectors. They have also designed indiscernibility relations to unite these weight vectors. The algorithm is applicable on datasets of different imbalance ratios.

An improved weighted algorithm is proposed by Patel and Thakur<sup>58</sup> in that large weights are assigned to small classes and small weight are assigned to larger classes, and when merging with fuzzy logic, the algorithm provides efficient classification results for imbalanced data.

One step ahead, an optimal fuzzy weighted nearest neighbor concept was proposed by Patel and Thakur,<sup>59</sup> and they have taken into consideration the advantages of both, the optimal weights and embedded fuzzy concept to achieve better classification results of imbalanced data.

Patel and Thakur<sup>60</sup> have proposed an approach which takes an adaptive concept of different K values for different classes to calculate more accurate

**Table 2.** Improved K-nearest neighbor algorithms.

S. No.	Authors	Title	Year	Key feature
1	Y Yang et al. <sup>40</sup>	Improving text categorization methods for event tracking	2000	Sum of similarity of each class is divided by the total number of instances of each class from the nearest neighbors
2	L Baoli et al. <sup>41</sup>	An adaptive k-nearest neighbor text categorization strategy	2004	Large K for large class and small K for small class
3	S Tan <sup>42</sup>	Neighbor-weighted k-nearest neighbor for unbalanced text corpus	2005	Small weights are assigned to large class and large weights are assigned to small class
4	S Tan <sup>43</sup>	An effective refinement strategy for KNN text classifier	2006	Weights of features of classes are increased or decreased using drag and push operation
5	L Wang et al. <sup>44</sup>	An effective evidence theory based k-nearest neighbor (knn) classification	2008	Global frequency estimation of prior probability and local frequency estimation of prior probability
6	W Liu and S Chawla <sup>45</sup>	Class confidence weighted kNN algorithms for imbalanced data sets	2011	Class confidence weights are found using probabilities of attributes to weight prototypes to transform prior probabilities of KNN into posterior probabilities
7	H-S Kang et al. <sup>46</sup>	The decomposed k-nearest neighbor algorithm for imbalanced text classification	2012	After applying traditional KNN on training data, classified data are decomposed into misclassified and correctly classified sets
8	E Kriminger et al. <sup>47</sup>	Nearest neighbor distributions for imbalanced classification	2012	Local geometric structure is used in data to reduce the effect of imbalance
9	H Dubey and V Pudi <sup>48</sup>	Class based weighted k-nearest neighbor over imbalance dataset	2013	Class-based weights are found by considering class distribution for neighbors of any test instance
10	N Tomašev and D Mladenić <sup>49</sup>	Class imbalance and the curse of minority hubs	2013	Hubness effect, minority classes are responsible for imbalance in high-dimensional datasets
11	D Ryu et al. <sup>50</sup>	A hybrid instance selection using nearest-neighbor for cross-project defect prediction	2015	KNN is used to learn local information, and global information is gained by applying naive Bayes approach for software defect prediction
12	S Ando <sup>51</sup>	Classifying imbalanced data in distance-based feature space	2016	Convex optimization technique is proposed to find out weight parameters to calculate class-based weights
13	H Patel and GS Thakur <sup>52</sup>	A Hybrid Weighted Nearest Neighbor Approach to Mine Imbalanced Data	2016	Large K is used with small weights for majority class, and small K is taken with weights for minority classes

KNN: K-nearest neighbor.

membership of data into classes merged with fuzzy nearest neighbor. Their results show the improved classification results on various imbalanced datasets.

Very less number of fuzzy nearest neighbor approaches is applied for imbalance issue until now. It could be seen that modification in K, weights, and fuzzy

concepts all together perform better for imbalanced data. In this research study, some of these combinations are proposed to classify imbalanced data with improved performance of nearest neighbor classifiers. Table 3 gives an instant look on these fuzzy and weighted KNN algorithms for imbalanced data.



**Table 3.** Fuzzy and weighted KNN algorithms for imbalanced data.

S. No.	Authors	Title	Year	Key feature
1	A Fernández et al. <sup>53</sup>	On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets	2009	Parameters of the adaptive inference system are learnt with genetic algorithm and then applied adaptive conjunction operator for different imbalance ratios.
2	H Han and B Mao <sup>55</sup>	Fuzzy-rough k-nearest neighbor algorithm for imbalanced data sets learning	2010	Membership function is designed for fuzzy-rough algorithm, and an equivalence relation is defined between test instances and their nearest neighbors.
3	C Liu et al. <sup>56</sup>	Coupled fuzzy k-nearest neighbors classification of imbalanced non-IID categorical data	2014	Sized membership assignment, similarity calculation, and integration for categorical imbalanced data.
4	E Ramentol et al. <sup>57</sup>	IFROWANN: imbalanced fuzzy-rough ordered weighted average nearest neighbor classification	2015	Fuzzy-rough ordered weighted nearest neighbor with six weight vectors and indiscernibility relations to unite these weight vectors.
5	H Patel and GS Thakur <sup>58</sup>	Classification of Imbalanced Data using a Modified Fuzzy-Neighbor Weighted Approach	2017	Large weights are assigned to small classes and small weights are assigned to larger classes and merging with fuzzy logic provides efficient classification results for imbalanced data.
6	H Patel and GS Thakur <sup>59</sup>	Improved fuzzy-optimally weighted nearest neighbor strategy to classify imbalanced data	2017	Fuzzy-optimal weights were calculated through the covariance method to improve classification of imbalanced data.
7	H Patel and GS Thakur <sup>60</sup>	An improved fuzzy k- nearest neighbor algorithm for imbalanced data using adaptive approach	2019	An adaptive nearest neighbor approach was taken with fuzzy KNN to acquire better classification results.

KNN: K-nearest neighbor.

### Cost-sensitive and ensemble approaches

Ensemble is the concept of merging different approaches intended to achieve the same objective with better accuracy and more reliable results. In the past year, many ensemble approaches have been proposed for classification of imbalanced data. Also, cost matrix plays an important role in classification. This section contains various ensemble techniques, including re-sampling, classifiers, and cost-sensitive approaches to learn from imbalanced datasets.

Zhou and Liu<sup>61</sup> studied the effect of sampling and other factors in training of cost-sensitive neural networks. These factors include undersampling, oversampling, SMOTE threshold-moving, and hard and soft ensembles. They concluded that threshold-moving and soft ensemble performs better for cost-sensitive neural networks. Also, cost-sensitive learning is convenient on binary data in comparison to multi-class data.

Nguyen et al.<sup>62</sup> proposed a feedforward neural network approach for imbalanced datasets. In this method, clustering is used to undersample the majority

class instances with the concept of weighted cluster centers and its desired output.

Köknar-Tezel and Latecki<sup>63</sup> proposed a supervised learning-based oversampling approach that creates and puts synthetic instances into distance space directly. This strategy is very useful for data where general distance measures cannot be used and so SMOTE cannot be applied, for example, time series. Then, they used SVM for classification. This approach performed good on such cases.

Milaré et al.<sup>64</sup> proposed a hybrid evolutionary algorithm to deal with class imbalance issue. In this method, they developed various balanced datasets from minority class instances and a sample from the majority class. The machine learning approach induces rules and these rules are used to select classifier by applying evolutionary algorithm. This approach reduces the overfitting of oversampling and information loss occurred due to undersampling.

Chen et al.<sup>65</sup> proposed a Probabilistic Classification based on Association Rules (PCAR) to classify imbalanced data more correctly. PCAR performs changes in

the pruning method, scoring procedure and rule sorting index of CBA to achieve such purpose.

Another associative classification algorithm for imbalanced learning by improving the scoring based on association (SBA) approach is proposed by Chen et al.<sup>66</sup> This improvement is done by combining the scoring with pruning of association rules in probabilistic classification based on associations (PCBA). Confidence is increased using undersampling and deciding different minimum support and confidence for rules of each class on the basis of distribution to adjust CBA for the forming of PCBA that also removes the pruning rules for the least error rate.

A review article on ensemble methods for classification of imbalanced data is written by Galar et al.<sup>67</sup> They developed taxonomy of ensembles for imbalance learning. They found that the use of ensemble technique performs well on imbalanced data using sampling and single classifier. With more classifiers, it becomes complex but yielding better performance for these unevenly distributed datasets. They also concluded that bagging and boosting approaches provide better classification of imbalanced data.

López et al.<sup>68</sup> conducted an analysis on the performances of data sampling and cost-sensitive approaches for learning from imbalanced data. After experiments, they came to the result that, in general, both strategies yield well and equal results for class imbalance and do determine the best among both; further data intrinsic characteristic analysis is needed.

Chen et al.<sup>69</sup> have carried out a study on the effect of different measures on classification of a dataset based on French bankruptcy and concluded that these measures rigorously affect the classification performance.

Wu et al.<sup>70</sup> proposed a random forest ensemble approach for categorization of imbalanced text data. This strategy contains feature subspaces that are stratified sampled and use SVM to split tree nodes. Stratified sampling is done to find out most important features for both minority and majority classes, and SVM in the learning tree model ensures the better classification of imbalanced text data.

Maldonado and López<sup>71</sup> proposed a cost-sensitive second-order cone programming SVM that is founded on linear programming SVM (LP-SVM) principle. For this, they relaxed the Vapnik–Chervonenkis (VC) bound conditions and maximized the margins directly with two margin variables for both majority and minority classes. By removing conic constraint, this method becomes less complex.

Shao et al.<sup>72</sup> proposed an efficient weighted Lagrangian twin support vector machine (WLTSVM). This approach constructs two proximal hyperplanes by using different training points. WLTSVM first performs graph-based undersampling to maintain proximity information and second Lagrangian twin support

vector machine (TWSVM) is improved by applying weights to diminish biasness. Finally, they proved the convergence of the proposed algorithm.

Peng et al.<sup>73</sup> brought up a data gravitational based classification method for imbalanced data and called it imbalanced data gravitation-based classification (IDGC). Gravitation computing is done by a new amplified gravitation coefficient that consists of information about class imbalance and it causes strengthening and weakening of minority and majority class gravitation fields. In weight optimization, they defined the evaluation function to make it sure for parameters to be improved for class imbalance.

A cost-sensitive decision tree based on ensemble methods is proposed by Krawczyk et al.<sup>74</sup> Classifier is designed based on cost matrix, and to select a classifier, assignment of weights, an evolutionary algorithm is applied. Training of classifier is done in random feature subspace, and parameters of cost matrix are chosen from receiver operating characteristic (ROC) analysis. This optimization technique performed well on imbalanced datasets.

Qian et al.<sup>75</sup> proposed a novel approach based on re-sampling ensemble that performs oversampling for minority classes and undersampling for majority classes. Scale of re-sampling is decided on the ratio of min class and max class instances. Results show that the performance of algorithm correlated to the ratio of the number of records in classes and number of attributes. This performs well on the ratio above than 3.

Krawczyk et al.<sup>19</sup> have proposed an ensemble of three image segmentation approaches to detect malignancy for breast cancer even for early biopsy images. This approach is implemented with boosting and evolutionary undersampling to achieve balanced data.

Table 4 contains short description of cost-sensitive and ensemble algorithms.

### *Feature selection and evaluation measures*

Maldonado et al.<sup>76</sup> proposed a set of approaches in that feature selection is done by successive hold out steps based on the backward elimination method, for which measure of contribution is derived from balanced loss function. The intension of this work is to perform better feature selection and deal with imbalance issue in parallel.

Maratea et al.<sup>77</sup> suggested an enhanced SVM for imbalanced data classification and a modified evaluation measure for evaluation of classifiers for such datasets. To cope up with data imbalance, an asymmetric space is developed in class surroundings by applying the first-step approximation and appropriate kernel transformation. The proposed accuracy measure takes care of imbalance nature of data. The scenario is designed for binary classification.

**Table 4.** Cost-sensitive and ensemble algorithms.

S. No.	Authors	Title	Year	Key feature
1	Z-H Zhou and X-Y Liu <sup>61</sup>	Training cost-sensitive neural networks with methods addressing the class imbalance problem	2006	Study of the effect of sampling and other factors in training of cost-sensitive neural networks
2	GH Nguyen et al. <sup>62</sup>	A supervised learning approach for imbalanced data sets	2008	Feedforward neural network + undersampling using clustering
3	S Köknar-Tezel and LJ Latecki <sup>63</sup>	Improving SVM classification on imbalanced data sets in distance spaces	2009	Oversampling approach that placed in distance space + SVM
4	CR Milaré et al. <sup>64</sup>	A hybrid approach to learn with imbalanced classes using evolutionary algorithms	2010	Oversampling + undersampling + evolutionary algorithm
5	W-C Chen et al. <sup>65</sup>	Adjusting and generalizing CBA algorithm to handling class imbalance	2012	PCAR + CBA
6	W-C Chen et al. <sup>66</sup>	Increasing the effectiveness of associative classification in terms of class imbalance by using a novel pruning algorithm	2012	SBA + PCBA
7	M Galar et al. <sup>67</sup>	A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches	2012	Review on ensemble techniques for imbalanced learning
8	V López et al. <sup>68</sup>	Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics	2012	Review on data sampling and cost-sensitive approaches for imbalanced classification
9	N Chen et al. <sup>69</sup>	Influence of class distribution on cost-sensitive learning: A case study of bankruptcy analysis	2013	Review on the cost ratio, imbalance ratio, and sample size measures
10	Q Wu et al. <sup>70</sup>	ForesTexter: an efficient random forest algorithm for imbalanced text categorization	2014	Random forest + stratified sampling using SVM
11	S Maldonado and J López <sup>71</sup>	Imbalanced data classification using second-order cone programming support vector machines	2014	Cost-sensitive second-order cone programming support vector machine
12	Y-H Shao et al. <sup>72</sup>	An efficient weighted Lagrangian twin support vector machine for imbalanced data classification	2014	Weighted Lagrangian twin support vector machine using graph-based undersampling
13	L Peng et al. <sup>73</sup>	A new approach for imbalanced data classification based on data gravitation	2014	IDGC is developed using oversampling + gravitation computing
14	B Krawczyk et al. <sup>74</sup>	Cost-sensitive decision tree ensembles for effective imbalanced classification	2014	Cost matrix + evolutionary algorithm
15	Y Qian et al. <sup>75</sup>	A resampling ensemble algorithm for classification of imbalance problems	2014	Ensemble of oversampling and undersampling
16	Krawczyk et al. <sup>19</sup>	Machine learning and image processing is combined to find better results for breast cancer prediction	2016	Ensemble of boosting and undersampling with three image segmentation approaches

SVM: support vector machine; PCAR: Probabilistic Classification based on Association Rules; SBA: scoring based on association; PCBA: probabilistic classification based on associations; CBA: classification based on associations; IDGC: imbalanced data gravitation-based classification.

## Imbalanced data in WSNs

WSN has several applications in health care, agriculture, weather forecasting, forest fire detection, Internet of Things, and so on.<sup>78–82</sup> Even when analyzing the routing protocols for WSN, there are many chances of imbalanced data being generated. Consider agriculture

application using WSN.<sup>83–92</sup> The sensors used in this application include soil moisture detection sensors, location sensors, humidity detection sensors, temperature sensors, optical sensors, electrochemical sensors, airflow sensors, and so on. If the sensors are collecting the temperature in a tropical country like African

countries, most of the times, the temperature will be on a higher side. Hence, while analyzing the temperature data generated through these sensors, the balance will be tilted toward high temperatures. Analysis related to lower temperatures in this case is difficult as the number of instances having low temperature is less. In this section, we discuss the work done by several researchers to handle imbalanced data in WSNs.

Yala et al.<sup>93</sup> evaluated WSN data from homes with various re-sampling methods like SMOTE-CSVM, CS-SVM, OS-CSVM along with soft-margin SVM to handle imbalanced data. They proved that SMOTE-CSVM and OS-SVM outperformed other state-of-art approaches. Also, they proved that OS-CSVM is marginally better than SMOTE-SVM when classifying using ubiquitous and binary sensors.

Asur S and Parthasarathy S<sup>94</sup> have used an ensemble classification model to detect rare events by handling of imbalanced data in WSN. They implemented their ensemble model on low-energy adaptive clustering hierarchy (LEACH), a cluster-based WSN architecture. Their approach yielded better accuracy and improved energy utilization.

Zhou H and Yu KM<sup>95</sup> used a model based on KNN and adaptive synthetic sampling (ASS), where KNN is used for imputation of missing values, and ASS is used for treating imbalanced data. Then, they used a feed-forward network for prediction of products which are defective on industrial WSN generated data.

Radivojac P et al.<sup>20</sup> used machine learning approaches for detecting intrusion in WSNs. They used two approaches, namely, LEACH and unified network protocol framework (UNPF) for handling imbalance in the data generated by WSN sensors. Their experimentation shows that handling of imbalanced data using machine learning mechanism significantly optimized energy consumption of the WSN.

Yang H et al.<sup>96</sup> used naïve Bayes predictors in the decision tree algorithm at the leaf level for handling imbalanced classes generated by sensors in WSN. They have applied their approach at the training phase, and fine-tuning the prediction accuracy using weighted naïve Bayes predictors at leaf nodes.

Yu J et al.<sup>97</sup> proposed a routing protocol based on clusters in WSN to handle imbalanced node distribution to improve the energy consumption. This approach uses energy-aware distributed clustering (EADC), a routing algorithm based on energy-aware clustering approach for non-uniform distributed nodes in WSN. The experimental results proved that their approach balanced energy utilization of nodes resulting in increased lifetime of network.

Tripathi M and Taneja A<sup>98</sup> applied cross-validation technique based on k-fold approach for handling imbalanced data in WSN. Then, random forest classifier is applied on the resultant data for classification on

balanced data. Their experimentation results proved that classification of balanced data yielded better results when compared to classification of unbalanced data.

Rodda S and Erothi USR<sup>99</sup> studied the presence of imbalanced data in intrusion datasets from benchmark NSL\_KDD. They used four prominent classification approaches to study the impact of imbalanced data on intrusion datasets using WSN.

M'hamed BA and Fergani B<sup>100</sup> proposed an updated version of multi-class weighted SVM model to deal with imbalance data problem in WSN. They gathered the data from three houses with different number of sensors and different layouts.

### Lessons learned and approaches suggested for handling imbalanced data in WSN

In this section, we discuss about the lessons learned about handling of imbalanced data in WSN.

Very few works have concentrated on handling of imbalanced data in WSNs. When the data are generated by wide range of sensors, continuously, there is every chance that the data generated from some of these sensors may be discrete; thereby, generated data from those sensors can be sparse. This makes the dataset generated from these sensors imbalanced. Extraction of patterns from these imbalanced datasets can then become biased. To handle these kinds of situations, the following approaches, which were used by several researchers for handling data imbalance in traditional datasets, can be extended to WSNs.

- (a) K-fold cross-validation: It is an approach used during training the machine learning algorithms. This approach re-samples the dataset during the training phase of machine learning. This approach splits the datasets into k different groups. One of these groups is considered as testing data and remaining groups are considered as training data. This approach gives equal priority to imbalanced or rare data also.
- (b) Ensemble re-sampled datasets: This approach re-samples the dataset in such a way that the data which are scarce or rare will be over-sampled. In this way, the overall data can be balanced, and the results achieved by machine learning approaches will be fair and unbiased.
- (c) Reduce the weight of the attributes with higher presence and increase the weight of attributes with fewer instances: In this approach, every attribute is assigned a weight. To balance the data to get fair results, attributes which have more presence will be given more weights and attributes with less presence will be given less

weight. In this way, imbalances in datasets can be handled efficiently.

- (d) Cost-sensitive learning: This approach incorporates misclassification costs in data mining for minimizing the total cost. This approach avoids pre-selection of hyper-parameters adjusting them dynamically.
- (e) Combined class methods: This approach uses a fusion of several methods to handle imbalanced data effectively. This approach can eliminate the noise in imbalanced datasets. This approach ensures that useful information will not be lost.

## Conclusion and future direction

Imbalance is a very common issue in today's scenario which causes severe deviation in the performances of traditional classifiers. This review article presents a thorough review on imbalance problem. Both, review and research, articles taken here give a deep insight into the imbalance problem and its various solutions. All considered contributions are systematically arranged in the manner so the differences could be easily understood. In this review, special emphasis is given to the improved KNN approaches proposed for classification of imbalanced data. Nearest neighbor approach is chosen for its simplicity. This literature study facilitates a background that assists for further research in this area and for improvement of new nearest neighbor-approaches. Also some open issues have been discussed here. In summary, the research community should consider the following directions when further developing solutions to imbalanced learning problems:

- Introduce solution for imbalanced learning for multi-class problem which will take account of different class relationships.
- Reflect on the structure and existence of situations of minority classes to gain a better understanding of the cause of learning problems.
- Introduce new approaches based on the specific organized existence of these problems for binary problem and multi-class problems.
- Propose new solutions for multi-instance and multi-label learning that are based on specific structured nature of these problems.
- Proposed efficient classification techniques for WSN in various scenarios.

This article shows that the vast field of imbalanced learning needs the research community's attention and intensive growth. There are still many fields that are untouched where this problem exists and solution is also required.





## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT; No. NRF-2018R1C1B5045013)

## ORCID iDs

G Thippa Reddy  <https://orcid.org/0000-0003-0097-801X>  
 Celestine Iwendi  <https://orcid.org/0000-0003-4350-3911>  
 Ali Kashif Bashir  <https://orcid.org/0000-0003-2601-9327>  
 Ohyun Jo  <https://orcid.org/0000-0001-8444-2786>

## References

1. Yang Q and Wu X. 10 challenging problems in data mining research. *Int J Inf Tech Decis Mak* 2006; 5: 597–604.
2. He H and Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng* 2009; 21: 1263–1284.
3. Rajput DS, Thakur RS and Thakur GS. A computational model for knowledge extraction in uncertain textual data using Karnaugh map technique. *Int J Comput Sci Math* 2016; 7: 166–176.
4. Rajput DS. Review on recent developments in frequent itemset based document clustering, its research trends and applications. *Int J Data Anal Tech Strateg* 2019; 11: 176–195.
5. Han J, Pei J and Kamber M. *Data mining: concepts and techniques*. Amsterdam: Elsevier, 2011.
6. Patel H and Rajput D. Data mining applications in present scenario: a review. *Int J Soft Comput* 2011; 6: 136–142.
7. Krawczyk B. Learning from imbalanced data: open challenges and future directions. *Prog Artif Intell* 2016; 5: 221–232.
8. Pavón R, Laza R, Reboiro-Jato M, et al. Assessing the impact of class-imbalanced data for classifying relevant/irrelevant Medline documents. In: *5th international conference on practical applications of computational biology & bioinformatics (PACBB 2011)*, Salamanca, 6–8 April 2011, pp.345–353. Berlin: Springer.
9. Rao RB, Krishnan S and Niculescu RS. Data mining for improved cardiac care. *ACM SIGKDD Explor Newsl* 2006; 8: 3–10.
10. Kubat M, Holte RC and Matwin S. Machine learning for the detection of oil spills in satellite radar images. *Mach Learn* 1998; 30: 195–215.
11. Chan PK and Stolfo SJ. Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. In: *Proceedings of the fourth international conference on knowledge discovery and data mining*, New York, 27–31 August 1998, pp.164–168. Menlo Park, CA: AAAI.

12. Li X-C, Mao W-J, Zeng D, et al. Performance evaluation of machine learning methods in cultural modeling. *J Comput Sci Technol* 2009; 24: 1010–1017.
13. Azaria A, Richardson A, Kraus S, et al. Behavioral analysis of insider threat: a survey and bootstrapped prediction in imbalanced data. *IEEE Trans Comput Soc Syst* 2014; 1: 135–155.
14. Xu R, Chen T, Xia Y, et al. Word embedding composition for data imbalances in sentiment and emotion classification. *Cogn Comput* 2015; 7: 226–240.
15. Basha SM and Rajput DS. A roadmap towards implementing parallel aspect level sentiment analysis. *Multimed Tools Appl* 2019; 78: 29463–29492.
16. Basha SM and Rajput DS. A supervised aspect level sentiment model to predict overall sentiment on tweeter documents. *Int J Metadata Semant Ontol* 2018; 13: 33–41.
17. Munkhdalai T, Namsrai O-E and Ryu KH. Self-training in significance space of support vectors for imbalanced biomedical event data. *BMC Bioinform* 2015; 16: S6.
18. Gao Z, Zhang L-F, Chen M-Y, et al. Enhanced and hierarchical structure algorithm for data imbalance problem in semantic extraction under massive video dataset. *Multimed Tools Appl* 2014; 68: 641–657.
19. Krawczyk B, Galar M, Jeleń Ł, et al. Evolutionary under-sampling boosting for imbalanced classification of breast cancer malignancy. *Appl Soft Comput* 2016; 38: 714–726.
20. Radivojac P, Korad U, Sivalingam KM, et al. Learning from class-imbalanced data in wireless sensor networks. In: *2003 IEEE 58th vehicular technology conference: VTC 2003-fall (IEEE Cat. No. 03CH37484)*, Orlando, FL, 6–9 October 2003, pp.3030–3034. New York: IEEE.
21. Chawla NV, Japkowicz N and Kotcz A. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explor Newsl* 2004; 6: 1–6.
22. Visa S and Ralescu A. Issues in mining imbalanced data sets—a review paper. In: *Proceedings of the sixteen Midwest artificial intelligence and cognitive science conference*, Dayton, Ohio, 16–17 April 2005, pp.67–73.
23. Guo X, Yin Y, Dong C, et al. On the class imbalance problem. In: *2008 fourth international conference on natural computation (ICNC'08)*, Jinan, China, 18–20 October 2008, pp.192–201. New York: IEEE.
24. Fernández A, García S and Herrera F. Addressing the classification with imbalanced data: open problems and new challenges on class distribution. In: *International conference on hybrid artificial intelligence systems*, Wroclaw, 23–25 May 2011, pp.1–10. Berlin: Springer.
25. López V, Fernández A, García S, et al. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Inform Sci* 2013; 250: 113–141.
26. Prati RC, Batista GE and Silva DF. Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. *Knowl Inform Syst* 2015; 45: 247–270.
27. Batista GE, Prati RC and Monard MC. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor Newsl* 2004; 6: 20–29.
28. García V, Sánchez JS and Mollineda RA. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowl Based Syst* 2012; 25: 13–21.
29. Menardi G and Torelli N. Training and assessing classification rules with imbalanced data. *Data Min Knowl Disc* 2014; 28: 92–122.
30. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002; 16: 321–357.
31. He H, Bai Y, Garcia EA, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, Hong Kong, China, 1–8 June 2008, pp.1322–1328. New York: IEEE.
32. Zhang H and Li M. RWO-sampling: a random walk over-sampling approach to imbalanced data classification. *Inf Fus* 2014; 20: 99–116.
33. Wang K-J, Makond B, Chen K-H, et al. A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. *Appl Soft Comput* 2014; 20: 15–24.
34. Sáez JA, Luengo J, Stefanowski J, et al. SMOTE-IPF: addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Inf Sci* 2015; 291: 184–203.
35. Tahir MA, Kittler J and Yan F. Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recognit* 2012; 45: 3738–3750.
36. Wong GY, Leung FH and Ling S-H. An under-sampling method based on fuzzy logic for large imbalanced dataset. In: *2014 IEEE international conference on fuzzy systems (FUZZ-IEEE)*, Beijing, China, 6–11 July 2014, pp.1248–1252. New York: IEEE.
37. Xu L, Chow M-Y and Taylor LS. Data mining based fuzzy classification algorithm for imbalanced data. In: *2006 IEEE international conference on fuzzy systems*, Vancouver, BC, Canada, 16–21 July 2006, pp.825–830. New York: IEEE.
38. Antonelli M, Ducange P, Marcelloni F, et al. Evolutionary fuzzy classifiers for imbalanced datasets: an experimental comparison. In: *2013 joint IFSA world congress and NAFIPS annual meeting (IFSA/NAFIPS)*, Edmonton, AB, Canada, 24–28 June 2013, pp.13–18. New York: IEEE.
39. García S, Triguero I, Carmona CJ, et al. Evolutionary-based selection of generalized instances for imbalanced classification. *Knowl Based Syst* 2012; 25: 3–12.
40. Yang Y, Ault T, Pierce T, et al. Improving text categorization methods for event tracking. In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Greece, 24–28 July 2000, pp.65–72. New York: ACM.
41. Baoli L, Qin L and Shiwen Y. An adaptive k-nearest neighbor text categorization strategy. *ACM Trans Asian Lang Inf Process* 2004; 3: 215–226.
42. Tan S. Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Syst Appl* 2005; 28: 667–671.
43. Tan S. An effective refinement strategy for KNN text classifier. *Expert Syst Appl* 2006; 30: 290–298.

44. Wang L, Khan L and Thuraisingham B. An effective evidence theory based k-nearest neighbor (knn) classification. In: *Proceedings of the 2008 IEEE/WIC/ACM international conference on Web intelligence and intelligent agent technology*, vol. 1, Sydney, NSW, Australia, 9–12 December 2008, pp.797–801. New York: IEEE.
45. Liu W and Chawla S. Class confidence weighted KNN algorithms for imbalanced data sets. In: *Pacific-Asia conference on knowledge discovery and data mining*, Shenzhen, China, 24–27 May 2011, pp.345–356. Berlin: Springer.
46. Kang H-S, Nam K and Kim S-I. The decomposed k-nearest neighbor algorithm for imbalanced text classification. In: *International conference on future generation information technology*, Gangneung-si, South Korea, 16–19 December 2012, pp.87–94. Berlin: Springer.
47. Kriminger E, Principe JC and Lakshminarayan C. Nearest neighbor distributions for imbalanced classification. In: *The 2012 international joint conference on neural networks (IJCNN)*, Brisbane, QLD, Australia, 10–15 June 2012, pp.1–5. New York: IEEE.
48. Dubey H and Pudi V. Class based weighted k-nearest neighbor over imbalance dataset. In: *Pacific-Asia conference on knowledge discovery and data mining*, Gold Coast, QLD, Australia, 14–17 April 2013, pp.305–316. Berlin: Springer.
49. Tomašev N and Mladenčić D. Class imbalance and the curse of minority hubs. *Knowl Based Syst* 2013; 53: 157–172.
50. Ryu D, Jang J-I and Baik J. A hybrid instance selection using nearest-neighbor for cross-project defect prediction. *J Comput Sci Technol* 2015; 30: 969–980.
51. Ando S. Classifying imbalanced data in distance-based feature space. *Knowl Inf Syst* 2016; 46: 707–730.
52. Patel H and Thakur G. A hybrid weighted nearest neighbor approach to mine imbalanced data. In: *Proceedings of the international conference on data mining (DMIN)*, Las Vegas, NV, 25–28 July 2016, p.106. CSREA Press.
53. Fernández A, del Jesus MJ and Herrera F. On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets. *Expert Syst Appl* 2009; 36: 9805–9812.
54. Fernández A, García S, del Jesus MJ, et al. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Set Syst* 2008; 159: 2378–2398.
55. Han H and Mao B. Fuzzy-rough k-nearest neighbor algorithm for imbalanced data sets learning. In: *2010 seventh international conference on fuzzy systems and knowledge discovery*, Yantai, China, 10–12 August 2010, pp.1286–1290. New York: IEEE.
56. Liu C, Cao L and Philip SY. Coupled fuzzy k-nearest neighbors classification of imbalanced non-IID categorical data. In: *2014 international joint conference on neural networks (IJCNN)*, Beijing, China, 6–11 July 2014, pp.1122–1129. New York: IEEE.
57. Ramentol E, Vluymans S, Verbiest N, et al. IFRO-WANN: imbalanced fuzzy-rough ordered weighted average nearest neighbor classification. *IEEE Trans Fuzzy Syst* 2015; 23: 1622–1637.
58. Patel H and Thakur G. Classification of imbalanced data using a modified fuzzy-neighbor weighted approach. *Int J Intell Eng Syst* 2017; 10: 56–64.
59. Patel H and Thakur GS. Improved fuzzy-optimally weighted nearest neighbor strategy to classify imbalanced data. *Int J Intell Eng Syst* 2017; 10: 156–162.
60. Patel H and Thakur GS. An improved fuzzy k-nearest neighbor algorithm for imbalanced data using adaptive approach. *IETE J Res* 2019; 65: 780–789.
61. Zhou Z-H and Liu X-Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Trans Knowl Data Eng* 2006; 18: 63–77.
62. Nguyen GH, Bouzerdoum A and Phung SL. A supervised learning approach for imbalanced data sets. In: *2008 19th international conference on pattern recognition*, Tampa, FL, 8–11 December 2008, pp.1–4. New York: IEEE.
63. Kökner-Tezel S and Latecki LJ. Improving SVM classification on imbalanced data sets in distance spaces. In: *2009 ninth IEEE international conference on data mining*, Miami, FL, 6–9 December 2009, pp.259–267. New York: IEEE.
64. Milaré CR, Batista GE and Carvalho AC. A hybrid approach to learn with imbalanced classes using evolutionary algorithms. *Log J IGPL* 2011; 19: 293–303.
65. Chen W-C, Hsu C-C and Hsu J-N. Adjusting and generalizing CBA algorithm to handling class imbalance. *Expert Syst Appl* 2012; 39: 5907–5919.
66. Chen W-C, Hsu C-C and Chu Y-C. Increasing the effectiveness of associative classification in terms of class imbalance by using a novel pruning algorithm. *Expert Syst Appl* 2012; 39: 12841–12850.
67. Galar M, Fernandez A, Barrenechea E, et al. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Trans Syst Man Cybern Part C Appl Rev* 2012; 42: 463–484.
68. López V, Fernández A, Moreno-Torres JG, et al. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification: open problems on intrinsic data characteristics. *Expert Syst Appl* 2012; 39: 6585–6608.
69. Chen N, Chen A and Ribeiro B. Influence of class distribution on cost-sensitive learning: a case study of bankruptcy analysis. *Intell Data Anal* 2013; 17: 423–437.
70. Wu Q, Ye Y, Zhang H, et al. ForesTexter: an efficient random forest algorithm for imbalanced text categorization. *Knowl Based Syst* 2014; 67: 105–116.
71. Maldonado S and López J. Imbalanced data classification using second-order cone programming support vector machines. *Pattern Recognit* 2014; 47: 2070–2079.
72. Shao Y-H, Chen W-J, Zhang J-J, et al. An efficient weighted Lagrangian twin support vector machine for imbalanced data classification. *Pattern Recognit* 2014; 47: 3158–3167.
73. Peng L, Zhang H, Yang B, et al. A new approach for imbalanced data classification based on data gravitation. *Inf Sci* 2014; 288: 347–373.
74. Krawczyk B, Woźniak M and Schaefer G. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Appl Soft Comput* 2014; 14: 554–562.

75. Qian Y, Liang Y, Li M, et al. A resampling ensemble algorithm for classification of imbalance problems. *Neurocomputing* 2014; 143: 57–67.
76. Maldonado S, Weber R and Famili F. Feature selection for high-dimensional class-imbalanced data sets using support vector machines. *Inf Sci* 2014; 286: 228–246.
77. Maratea A, Petrosino A and Manzo M. Adjusted F-measure and kernel scaling for imbalanced data learning. *Inf Sci* 2014; 257: 331–341.
78. Cai Z-W and Huang L-H. Finite-time synchronization by switching state-feedback control for discontinuous Cohen–Grossberg neural networks with mixed delays. *Int J Mach Learn Cybern* 2018; 9: 1683–1695.
79. Yin X, Zhang K, Li B, et al. A task allocation strategy for complex applications in heterogeneous cluster-based wireless sensor networks. *Int J Distrib Sens Netw* 2018; 14: 1–15.
80. Wang J, Gao Y, Yin X, et al. An enhanced PEGASIS algorithm with mobile sink support for wireless sensor networks. *Wirel Commun Mob Comput* 2018; 2018: 9472075.
81. Hosen AS, Singh S, Mariappan V, et al. A secure and privacy preserving partial deterministic RWP model to reduce overlapping in IoT sensing environment. *IEEE Access* 2019; 7: 39702–39716.
82. Hosen AS, Singh S, Sharma PK, et al. A QoS-aware data collection protocol for LLNs in fog-enabled internet of things. *IEEE Trans Netw Serv Manag* 2020; 17: 430–444.
83. Palve A and Patel H. Towards securing real time data in IoMT environment. In: *2018 8th international conference on communication systems and network technologies (CSNT)*, Bhopal, India, 24–26 November 2018, pp.113–119. New York: IEEE.
84. Mittal M and Iwendi C. A survey on energy-aware wireless sensor routing protocols. *EAI Endors Trans Energy Web* 2019; 6: 1–16.
85. Iwendi C, Allen A and Offor K. Smart security implementation for wireless sensor network nodes. *J Wirel Sensor Netw* 2015; 1: 1–13.
86. Iwendi C and Offor K. Alternative protocol implementation for wireless sensor network nodes. *J Telecommun Syst Manage* 2013; 2: 106.
87. Iwendi C, Alqarni MA, Anajemba JH, et al. Robust navigational control of a two-wheeled self-balancing robot in a sensed environment. *IEEE Access* 2019; 7: 82337–82348.
88. Bashir AK, Lim S-J, Hussain CS, et al. Energy efficient in-network RFID data filtering scheme in wireless sensor networks. *Sensors* 2011; 11: 7004–7021.
89. Shafiq M, Yu X, Bashir AK, et al. A machine learning approach for feature selection traffic classification using security analysis. *J Supercomput* 2018; 74: 4867–4892.
90. Yaseen S, Abbas SMA, Anjum A, et al. Improved generalization for secure data publishing. *IEEE Access* 2018; 6: 27156–27165.
91. Qureshi NMF, Siddiqui IF, Unar MA, et al. An aggregate MapReduce data block placement strategy for wireless IoT edge nodes in smart grid. *Wirel Pers Commun* 2019; 106: 2225–2236.
92. Chauhdary SH, Hassan A, Alqarni MA, et al. A twofold sink-based data collection in wireless sensor network for sustainable cities. *Sustain Cities Soc* 2019; 45: 1–7.
93. Yala N, Fergani B and Clavier L. Soft margin SVM modeling for handling imbalanced human activity datasets in multiple homes. In: *2014 international conference on multimedia computing and systems (ICMCS)*, Marrakech, Morocco, 14–16 April 2014, pp.421–426. New York: IEEE.
94. Asur S and Parthasarathy S. Correlation-based feature partitioning for rare event detection in wireless sensor networks. In: *Proceedings of the 1st international workshop on knowledge discovery from sensor data (Sensor-KDD)*, San Jose, CA, 12–15 August 2007. New York: ACM.
95. Zhou H and Yu K-M. Imbalanced data classification for defective product prediction based on industrial wireless sensor network. In: *2017 sixth international conference on future generation communication technologies (FGCT)*, Dublin, 21–23 August 2017, pp.1–6. New York: IEEE.
96. Yang H, Fong S, Wong R, et al. Optimizing classification decision trees by using weighted naïve Bayes predictors to reduce the imbalanced class problem in wireless sensor network. *Int J Dist Sensor Netw* 2013; 9: 460641.
97. Yu J, Qi Y, Wang G, et al. A cluster-based routing protocol for wireless sensor networks with nonuniform node distribution. *AEU Int J Electron Commun* 2012; 66: 54–61.
98. Tripathi M and Taneja A. K-fold cross-validation machine learning approach on data imbalance for wireless sensor network, 2019, [https://ijsret.com/wp-content/uploads/2019/09/IJSRET\\_V5\\_issue5\\_441.pdf](https://ijsret.com/wp-content/uploads/2019/09/IJSRET_V5_issue5_441.pdf)
99. Rodda S and Erothi USR. Class imbalance problem in the network intrusion detection systems. In: *2016 international conference on electrical, electronics, and optimization techniques (ICEEOT)*, Chennai, India, 3–5 March 2016, pp.2685–2688. New York: IEEE.
100. M’hamed BA and Fergani B. A new multi-class WSVM classification to imbalanced human activity dataset. *J Comput* 2014; 9: 1560–1565