**Problems and future directions**

# Learning from imbalanced data: open challenges and future directions [2016](https://link.springer.com/article/10.1007/s13748-016-0094-0#article-info) <https://link.springer.com/article/10.1007/s13748-016-0094-0>

problems:

* In imbalanced datasets in machine learning, the number of samples of a given class is under-represented compared to other classes.
* Difficulty in learning from skewed distributions of binary tasks.
* The use of binary decomposition methods, such as one-vs-one and one-vs-all, may have drawbacks like high number of base classifiers or introducing additional artificial imbalance.
* The need for solutions that can aggregate classes according to their similarities or dissimilarities and use a sequential, step-wise approach to determine the final class.
* Bias of machine learning algorithms towards the majority class.
* The need for dedicated combination strategies that are able to reconstruct the original multi-class problem, specifically those that are suitable for cases with skewed distributions.

future directions:

* Developing new methods for decomposition that are able to adjust to individual pairwise problems and select specific data or algorithm-level solutions on the basis of subproblem characteristics.
* Researching and implementing alternative techniques to binary decomposition, such as hierarchical methods that aggregate classes according to their similarities or dissimilarities.
* Designing new fusion approaches suitable for cases with skewed distributions that can compensate for the imbalance both on decomposed class level and on final output combination level.
* Focusing on computationally efficient, adaptive and real-time methods for dealing with imbalanced data.
* Expanding the focus of imbalanced learning to cover other areas such as regression, clustering, data streams and big data analytics.

Strategies for learning in class imbalance

<https://www.researchgate.net/publication/220604068_Strategies_for_Learning_in_Class_Imbalance_Problems>

problems :

* High imbalance in class representation can cause deterioration in performance of standard supervised methods.
* Assigning distinct costs to classification errors can be difficult to implement and may not always be appropriate.
* Resampling the original training set can lead to loss of useful information from the majority class.
* Internally biasing the discrimination-based process can be complex and may not always be effective.
* Imbalanced data can lead to overfitting of models to the majority class.

Future directions:

* preserve useful information from both classes.
* Exploring alternative measures of classifier performance that are suitable for imbalanced data.
* Investigating and developing methods to correct the bias of classifiers towards the majority class.
* Researching the effect of imbalanced data on other types of machine learning algorithms beyond the nearest neighbor classifier.
* Examining the impact of class imbalance on performance in a wide range of real-world applications.

Imbalanced datasets classification and solution: a review 2014

<http://researchmanuscripts.com/July2014/2.pdf>

problems :

* The problem of using data intrinsic distinctiveness in classification problems, which can lead to the development of models that lack density in the training data, the presence of small disjuncts, the identification of noisy data, and the dataset shift between the training and the test distributions.
* The small disjuncts problem, which occurs when concepts are represented within small clusters and can lead to data fragmentation with a few representations of instances.
* The lack of density or lack of information problem, where induction algorithms do not have enough data to make generalizations about the distribution of samples, which becomes more difficult in the presence of high dimensional and imbalanced data.
* The presence of noise, which has a greater impact on the minority classes than on usual cases, and can lead to the learned sub concept being affected.
* The dataset shift problem, where training and test data follow different distributions, which is especially relevant when dealing with imbalanced categorization because in highly imbalanced domains, the minority class is mostly sensitive to singular classification errors.

future directions:

* Creating innovative algorithms or modifying existing ones to take the class-imbalance problem into consideration
* Pre-processing the data to reduce the effect caused by class imbalance
* Focusing on the exact classification of minority class as it is more important than majority class
* Using cost-sensitive learning method to associate cost with misclassifying patterns
* Applying boosting method to improve the performance of weak classifiers.

Learning from class-imbalanced data: Review of methods and applications 2017

<https://www.sciencedirect.com/science/article/abs/pii/S0957417416307175>

Problems:

* Standard classifiers such as logistic regression, Support VectoMachine (SVM) and decision tree are not suitable for imbalanced data and r often provide suboptimal classification results.
* The learning process guided by global performance metrics such as prediction accuracy induces a bias towards the majority class, while the rare episodes remain unknown even if the prediction model produces a high overall precision.
* Rare minority examples may be treated as noise by the learning model.
* The imbalanced nature of the data can greatly affect the learning of various machine learning classifiers, as most are designed to handle balanced datasets and perform poorly when applied to imbalanced data.
* Many existing methods have been developed mainly for binary classification, which is not always applicable for multi-class imbalanced datasets.

Future Directions:

* Develop new classification algorithms specifically for imbalanced data that can overcome the limitations of traditional models.
* Develop new performance metrics that take into account the imbalance of the data and better represent the performance of the classifier.
* Develop new techniques for data preprocessing that can handle minority examples without removing them or considering them as noise.
* Investigate the use of ensemble methods and deep learning techniques for imbalanced data classification.
* Develop new methods for multi-class imbalanced data classification, and investigate the use of transfer learning and domain adaptation techniques.

Handling imbalanced datasets: A review 2006

<https://www.researchgate.net/publication/228084509_Handling_imbalanced_datasets_A_review>

problems:

* The class imbalance problem is pervasive and ubiquitous, causing trouble to a large segment of the data mining community.
* The data collection process can sometimes create "artificial" imbalances.
* The cost of making different errors can vary per case, which makes it difficult to optimize the classifier.
* It is still unclear which sampling method performs best, what sampling rate should be used, and that the proper choice is probably domain specific.
* Rare classes are more error-prone than common classes

future directions:

* Developing methods for identifying a good class distribution in order to generate multiple training sets with the desired class distribution, and using meta-learning to form a composite learner from the resulting classifiers.
* Exploring the use of progressive-sampling algorithms to build larger and larger training sets, where the ratio of positive to negative examples added in each iteration is chosen based on the performance of the various class distributions evaluated in the previous iteration.
* Further research on cost-sensitive learning methods, such as MetaCost and Adacost, which have been shown to produce lower cumulative misclassification costs than traditional methods.
* Developing algorithms that use boosting to address the problems with rare classes, such as Rare-Boost and SMOTEBoost, which scale false-positive examples in proportion to how well they are distinguished from true-positive and true-negative examples.
* Investigating new combination approaches like ensemble methods, mixture of experts and meta-learning to improve the overall performance of classifier with imbalanced data.

# What makes multi-class imbalanced problems difficult? An experimental study 2022

<https://www.sciencedirect.com/science/article/abs/pii/S0020025518304997>

problems:

* Difficulty recognizing instances of the minority class due to bias towards majority classes in standard learning algorithms.
* Overlapping regions between classes, which can make it harder to distinguish between minority and majority classes.
* Rare sub-concepts in minority classes, which can make it harder to identify and classify them.
* Rare minority examples located inside the majority class region, which can make it harder to identify and classify them.
* The combination of various data difficulty factors, such as class imbalance, overlapping regions, and rare sub-concepts, which can make learning from class imbalanced data challenging.

future directions:

* Developing new methods to improve classifiers specifically designed for multi-class imbalanced data.
* Further research on the impact of class overlapping on the performance of classifiers and how to mitigate it.
* Investigating the role of class size configuration in multi-class imbalanced data and how to optimize it for better performance.
* Exploring the use of ensemble methods and other advanced techniques for dealing with multi-class imbalanced data.
* Conducting more experimental studies on real-world applications to better understand the challenges and limitations of multi-class imbalanced data in different fields.

**Solutions**:

# Learning from imbalanced data: open challenges and future directions [2016](https://link.springer.com/article/10.1007/s13748-016-0094-0#article-info)

# In imbalanced datasets in machine learning, the number of samples of a given class is under-represented compared to other classes.

Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning 2017

<https://www.jmlr.org/papers/volume18/16-365/16-365.pdf>  
  
This article deals with the problem of imbalanced datasets in machine learning, where the number of samples for a given class is under-represented compared to other classes. This imbalance can compromise the learning process and affect the performance of standard machine learning algorithms. The authors propose the imbalanced-learn API, a python toolbox that provides a wide range of methods to cope with this problem. These methods can be categorized into four groups: under-sampling, over-sampling, combination of over- and under-sampling, and ensemble learning methods.  
  
Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss 2019

<https://proceedings.neurips.cc/paper/2019/file/621461af90cadfdaf0e8d4cc25129f91-Paper.pdf>

This article aims to tackle the problem of imbalanced datasets in machine learning. Imbalanced datasets are characterized by having a disproportionate number of samples of one class compared to other classes, which can lead to poor performance when training models. To address this issue, the authors propose two new methods for improving model performance on imbalanced datasets.

The first method is the label-distribution-aware margin (LDAM) loss, which is designed to minimize a margin-based generalization bound. This loss can be used in conjunction with other strategies for handling class imbalance, such as re-weighting or re-sampling.

The second method is a training schedule that defers re-weighting until after the initial stage of training. This allows the model to learn an initial representation before adjusting for class imbalance.

By using these two methods, the authors aim to improve model performance on less frequent classes, which is a common problem in many real-world applications.

* Difficulty in learning from skewed distributions of binary tasks.

# Multi-class imbalanced big data classification on Spark 2021

<https://www.sciencedirect.com/science/article/abs/pii/S0950705120307279>

This paper proposes a compound framework for dealing with multi-class big data problems that addresses both the existence of multiple classes and high volumes of data. The framework analyzes instance-level difficulties in each class to understand what causes learning difficulties, and embeds this information in popular resampling algorithms to allow for informative balancing of multiple classes. The authors propose an efficient implementation of the algorithm on Apache Spark, including a novel version of the Synthetic Minority Over-sampling Technique (SMOTE) that overcomes spatial limitations in distributed environments. Extensive experimental studies show that using instance-level information significantly improves learning from multi-class imbalanced big data.

* The use of binary decomposition methods, such as one-vs-one and one-vs-all, may have drawbacks like high number of base classifiers or introducing additional artificial imbalance.

# Multi-class imbalanced big data classification on Spark 2021

<https://www.sciencedirect.com/science/article/abs/pii/S0950705120307279>

This article proposes a new approach called CODIL for solving the problem of multi-class classification, which is a supervised learning problem where each object is classified into one of multiple classes. The traditional method for solving this problem is either by directly designing a multi-class classifier or decomposing it into a set of binary classification problems. The proposed approach, CODIL, combines the advantages of both direct and indirect strategies.

The article also includes experimental results comparing CODIL to other established multi-class approaches on various benchmark data sets, which show the superiority of the proposed approach.

* The need for solutions that can aggregate classes according to their similarities or dissimilarities and use a sequential, step-wise approach to determine the final class.

# Boosting methods for multi-class imbalanced data classification: an experimental review 2020

<https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00349-y>

This paper reviews the most significant published boosting techniques for multi-class imbalanced datasets. Boosting algorithms are a class of ensemble learning methods that improve the performance of separate base learners by combining them into a composite whole. The paper conducts a thorough empirical comparison to analyze the performance of binary and multi-class boosting algorithms on various multi-class imbalanced datasets. Based on the obtained results for performance evaluation metrics, the selected metrics are compared to determine a suitable performance metric for multi-class imbalanced datasets. The experimental studies show that the CatBoost and LogitBoost algorithms are superior to other boosting algorithms on multi-class imbalanced conventional and big datasets, respectively. Furthermore, the MMCC (Maximum Mean Class-wise Correct Classification Rate) is a better evaluation metric than the MAUC (Macro Area Under the ROC Curve) and G-mean in multi-class imbalanced data domains.

* Bias of machine learning algorithms towards the majority class.

# Dealing with Imbalanced Classes in Machine Learning 2018

<https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2>

This article discusses ways to address class imbalance in machine learning, where one class makes up a larger portion of the data. It suggests using different evaluation metrics and cost-sensitive learning to penalize misclassifying the minority class more heavily. It also recommends using sampling methods such as SMOTE and ADASYN to create synthetic samples of the minority class while considering the majority class distribution. The goal is to improve the performance of the algorithm on the minority class and reduce bias towards the majority class.

* The need for dedicated combination strategies that are able to reconstruct the original multi-class problem, specifically those that are suitable for cases with skewed distributions.

# Machine Learning — Multiclass Classification with Imbalanced Dataset 2018

<https://towardsdatascience.com/machine-learning-multiclass-classification-with-imbalanced-data-set-29f6a177c1a>

This article talks about that multi-class classification with imbalanced dataset presents a different challenge than a binary classification problem. The skewed distribution makes many conventional machine learning algorithms less effective, especially in predicting minority class examples. To overcome this, one can use techniques like cost-sensitive learning, sampling, and using appropriate metrics for evaluation such as recall, precision, and AUROC instead of just accuracy. The confusion matrix is also a useful tool for evaluating the performance of a model in imbalanced datasets, as it allows for the analysis of true positive and false positive rates for each class. Additionally, techniques like ensemble methods and anomaly detection can also be used to improve the performance of the model in imbalanced datasets.

**Strategies for learning in class imbalance**

* High imbalance in class representation can cause deterioration in performance of standard supervised methods.

# A hybrid data-level ensemble to enable learning from highly imbalanced dataset 2021

<https://www.sciencedirect.com/science/article/abs/pii/S0020025520311889>

The article discusses a method, called HD-Ensemble, for dealing with cases where one class in a dataset makes up a much larger portion of the data compared to the other class. This is a common problem in machine learning, and it can lead to poor performance when trying to identify instances of the minority class. The HD-Ensemble method combines two techniques, one that filters out unrepresentative majority instances and another that generates diverse minority instances, to balance the data distribution and improve the performance of ensemble learning. This is shown to have better performance than other ensemble solutions when applied to highly imbalanced datasets.

* Assigning distinct costs to classification errors can be difficult to implement and may not always be appropriate.

# Imbalanced Data in Classification: General Solution & Case Study 2020

<https://towardsdatascience.com/imbalanced-data-in-classification-general-solution-case-study-169f2e18b017>

The article discusses the problem of imbalanced datasets in classification problems, where one class makes up a much larger portion of the data compared to the other class. This can lead to poor performance when training a classifier. The article suggests using techniques such as oversampling, undersampling and SMOTE to balance the dataset. Additionally, the article discusses cost-sensitive learning, a technique in which different costs are assigned to misclassifying instances of different classes. However, the article notes that assigning distinct costs to classification errors can be difficult to implement and may not always be appropriate for all situations.

* Resampling the original training set can lead to loss of useful information from the majority class.

# 10 Techniques to deal with Imbalanced Classes in Machine Learning 2020

<https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/>

This article explains that Imbalanced classes in machine learning is a common problem where the number of observations per class is not equally distributed, often resulting in a majority class that has a much larger percentage of the dataset and minor classes that do not have enough examples. This can lead to a decrease in the sensitivity of the model towards minority classes and can result in inaccurate predictions. To combat this problem, there are several techniques that can be used such as Random under-sampling, Random over-sampling, NearMiss, weighted loss, data augmentation techniques, and transfer learning. Additionally, it is important to use appropriate metrics, such as AUC Curves and Average Precision, rather than relying solely on accuracy, when evaluating the performance of a model on imbalanced datasets.

* Internally biasing the discrimination-based process can be complex and may not always be effective.

DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data

<https://arxiv.org/pdf/2105.02340.pdf>

This article talks about DeepSMOTE which is a method that helps deal with class imbalance in machine learning by creating synthetic examples of the minority class. It uses an encoder/decoder framework, which is similar to deep learning models, and a technique called SMOTE to generate new examples. The main advantage of this approach is that it doesn't require a complex process to bias the discrimination and it generates high-quality, artificial examples that are useful for visual inspection. This algorithm can be applied to a wide range of datasets.

* Imbalanced data can lead to overfitting of models to the majority class.

# Dealing with Imbalanced Classes in Machine Learning 2018

<https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2>

**IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW 2014**

* The problem of using data intrinsic distinctiveness in classification problems, which can lead to the development of models that lack density in the training data, the presence of small disjuncts, the identification of noisy data, and the dataset shift between the training and the test distributions.

Training and Prediction Data Discrepancies: Challenges of Text Classification with Noisy, Historical Data 2018

<https://aclanthology.org/W18-6114.pdf>

The article discusses the problem of using data that is not well suited for the task of classification. This can lead to poor performance because the model is not able to learn from the data effectively. The authors demonstrate that using dirty, noisy data can still lead to good performance on cleaner, more relevant data. They also highlight the challenges of using historical data in industry settings, where the data may not be ideal for the task at hand. They explain that this can lead to issues such as models that lack density in the training data, the presence of small disjuncts, the identification of noisy data, and the dataset shift between the training and the test distributions.

* The small disjuncts problem, which occurs when concepts are represented within small clusters and can lead to data fragmentation with a few representations of instances.

# An insight into imbalanced Big Data classification: outcomes and challenges 2017

<https://link.springer.com/article/10.1007/s40747-017-0037-9>

The article talks about how big data can cause problems when trying to classify imbalanced data. It suggests using the MapReduce framework as a solution, but notes that there is not much research on how to adapt traditional techniques to this framework. The article also mentions that imbalanced data can cause issues like small clusters of data, which can make it harder to classify the data correctly. The article discusses the current state of the field and the challenges that need to be addressed in the future.

* The lack of density or lack of information problem, where induction algorithms do not have enough data to make generalizations about the distribution of samples, which becomes more difficult in the presence of high dimensional and imbalanced data.

# An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics 2013

<https://www.sciencedirect.com/science/article/abs/pii/S0020025513005124>

This paper aims to provide an overview of the challenges and current trends in classification with imbalanced data in machine learning. It discusses the characteristics of imbalanced datasets and the specific metrics used to evaluate performance in these scenarios. The paper also reviews various proposed solutions to the problem, including preprocessing techniques, cost-sensitive learning, and ensemble methods. The authors conduct an experimental study to contrast these approaches and provide a thorough discussion on the main issues related to using data intrinsic characteristics in the classification problem. These issues include small disjuncts, lack of density in the training data, overlapping between classes, noisy data, borderline instances, and dataset shift.

* The presence of noise, which has a greater impact on the minority classes than on usual cases, and can lead to the learned sub concept being affected.

# SMOTE-LOF for noise identification in imbalanced data classification 2022

<https://www.sciencedirect.com/science/article/pii/S1319157821000161>

SMOTE is a technique used to handle imbalanced data in machine learning. It uses synthetic data to balance the distribution of the minority and majority classes, by creating new samples of the minority class that are different from the original ones. However, SMOTE can produce synthetic minority data samples considered as noise, which is also part of the majority classes. To overcome this problem, this study proposes a method called SMOTE-LOF, which combines SMOTE with the Local Outlier Factor (LOF) to identify noise from synthetic minority data produced in handling imbalanced data. The experiment was carried out using imbalanced datasets and the results showed that SMOTE-LOF produces better accuracy and f-measure than the SMOTE, but future research needs to be carried out using different datasets with varying number of data samples and the imbalanced ratio.

* The dataset shift problem, where training and test data follow different distributions, which is especially relevant when dealing with imbalanced categorization.

Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution

<https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1385_2011-hais-imbalanced.pdf>

The article talks about a problem in machine learning where the amount of examples for one class is much less than the other class. This is a common problem and can happen in many real-world situations. The article explains two ways to solve this problem: by adjusting the number of examples for each class (sampling) and by changing the importance given to correctly classifying examples from each class (cost-sensitive learning). The article also mentions that there can be specific challenges when dealing with imbalanced data, such as when the training and test data have different distributions. This is particularly important when trying to correctly classify examples from the minority class.

**Learning from class-imbalanced data: Review of methods and applications 2017**

* Standard classifiers such as logistic regression, Support VectoMachine (SVM) and decision tree are not suitable for imbalanced data and r often provide suboptimal classification results.

Generation of Controlled Synthetic Samples and Impact of Hyper-Tuning Parameters to Effectively Classify the Complex Structure of Overlapping Region 2022

<https://www.mdpi.com/2076-3417/12/16/8371>

The classification of imbalanced and overlapping data is a significant problem in data mining, as the number of examples representing one class is much lower than the number of examples representing the other classes. The problem with imbalanced datasets is that standard classification learning algorithms are often biased towards the majority classes, resulting in poor performance for the minority class. To address this problem, researchers have proposed various techniques such as sampling, cost-sensitive learning, algorithm adaptation methods, transformation methods, hybrid methods, and ensemble techniques.

* The learning process guided by global performance metrics such as prediction accuracy induces a bias towards the majority class, while the rare episodes remain unknown even if the prediction model produces a high overall precision.

# Performance Metrics in Machine Learning — Part 1: Classification 2020

<https://towardsdatascience.com/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92>

In this article, the author explains some of the key concepts and performance metrics that are commonly used in classification tasks in machine learning. The author explains that when evaluating the performance of a classification model, it is important to understand the concepts of true value vs predicted value, and true positive, true negative, false positive, and false negative. The author also mentions that there are many different performance metrics that can be used in classification tasks, and choosing the right one for a specific model is key to being able to measure the performance of the model objectively and in the right setting.

* Rare minority examples may be treated as noise by the learning model.

# Types of minority class examples and their influence on learning classifiers from imbalanced data 2016

<https://link.springer.com/article/10.1007/s10844-015-0368-1>

The article suggests that when dealing with imbalanced data, it is important to understand the characteristics of the minority class examples, as they may be treated as noise by the learning model. The authors propose analyzing different types of minority class examples, such as safe, borderline, rare, and outliers, to better understand their influence on classification performance. They suggest using methods such as analyzing the class distribution in a local neighborhood and using tools such as k-nearest examples and kernel functions to differentiate the performance of popular classifiers and pre-processing methods. The article also suggests that by considering the results of this analysis, it is possible to develop new algorithms for learning classifiers and pre-processing methods.

* The imbalanced nature of the data can greatly affect the learning of various machine learning classifiers, as most are designed to handle balanced datasets and perform poorly when applied to imbalanced data.

# Classification of Imbalanced Data:Review of Methods and Applications 2021

<https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012077/pdf>

This article discusses the problem of imbalanced datasets in classification tasks and its negative impact on performance. It explains that imbalanced datasets occur when the sample size from one class is much larger or smaller than another class. The article states that standard machine learning algorithms, such as decision trees and neural networks, are not effective when dealing with imbalanced datasets. To combat this problem, the article covers three approaches: data level, combining methods and algorithmic level.

* Many existing methods have been developed mainly for binary classification, which is not always applicable for multi-class imbalanced datasets.

Classification Problem in Imbalanced Datasets 2019

<https://www.intechopen.com/chapters/70393>

This article discusses a problem in classification -imbalanced datasets, where one class has significantly more examples than the other class. It explains that traditional methods used for classification can perform poorly in these cases and mentions various techniques that have been developed to address this problem. However, the article also notes that many of these methods are designed for binary classification, which only has two classes, and may not be appropriate for datasets with multiple classes. The article discusses this limitation and the challenges of applying these techniques to multi-class imbalanced datasets.

**Handling imbalanced datasets: A review 2006**

* The class imbalance problem is pervasive and ubiquitous, causing trouble to a large segment of the data mining community.  
    
  DATA MINING FOR IMBALANCED DATASETS: AN OVERVIEW 2005

<https://www3.nd.edu/~dial/publications/chawla2005data.pdf>

In this paper, the author discusses the issue of imbalanced datasets in machine learning and the challenges it poses in real-world applications. The author mentions that traditional measures of performance such as accuracy may not be appropriate when the data is imbalanced or when the costs of different errors vary greatly. The author then discusses various sampling techniques used for balancing datasets and alternative performance measures that are more suitable for mining imbalanced datasets. The author also mentions different approaches to addressing the problem of imbalanced datasets such as random oversampling and undersampling, focused oversampling and undersampling, and synthetic generation of new samples.

* The data collection process can sometimes create "artificial" imbalances.

# 7 Techniques to Handle Imbalanced Data 2022

<https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html>

The article talks about how the process of collecting data can sometimes lead to imbalanced data, where one class has much more examples than the others. This can happen because of various reasons such as the way the data is collected, the population being studied or the way the data is labeled. This can be a problem for machine learning models as they may not perform well on imbalanced data. To prevent this from happening, the article suggests being aware of these issues during the data collection process and taking steps to correct them. This can include techniques such as oversampling, undersampling and data augmentation.

* The cost of making different errors can vary per case, which makes it difficult to optimize the classifier.

# A Gentle Introduction to Imbalanced Classification 2019

<https://machinelearningmastery.com/what-is-imbalanced-classification/>

The article discusses the problem of imbalanced classification, where the distribution of examples across the known classes is skewed. It highlights that this poses a challenge for predictive modeling, as most machine learning algorithms were designed around the assumption of equal numbers of examples for each class.

To address this problem, the article suggests different techniques such as using appropriate evaluation metrics, resampling the training set, using ensemble techniques, and cost-sensitive learning. However, it also mentions that the cost of making different errors can vary depending on the specific case, which can make it difficult to optimize the classifier for the best performance. This means that different approaches may need to be considered to optimize the classifier for the specific problem at hand.

* It is still unclear which sampling method performs best, what sampling rate should be used, and that the proper choice is probably domain specific.

# Imbalanced Learning: sampling techniques 2020

<https://tungmphung.com/imbalanced-learning-sampling-techniques/>

The article suggests that when dealing with imbalanced classification, it can be difficult to determine which method of resampling the data (such as oversampling or undersampling) is best and at what rate the data should be resampled. The article suggests that a combination of methods may be more effective and that the best method may vary depending on the specific application or domain. To address this problem, the article suggests using an ensemble approach, where multiple classifiers are learned using different sampling strategies and then combined, to improve performance.

* Rare classes are more error-prone than common classes

# Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE 2018

<https://www.sciencedirect.com/science/article/abs/pii/S0020025518304997>

The article discusses the problem of imbalanced data in classification tasks, where the number of examples for certain classes is much lower than for other classes. This can lead to poor performance for the minority class, as most machine learning algorithms are designed with the assumption of equal class distribution. The article suggests that an oversampling method based on k-means clustering and SMOTE can be used to address this problem by balancing the class distribution and avoiding the generation of noise. Additionally, the article highlights the issue of "small disjuncts" where classification rules covering only a small number of training examples can be more susceptible to error, particularly for rare classes. The proposed method aims to reduce the error rate for rare classes by effectively generating synthetic minority data, which improves the performance of the classifier.

# **What makes multi-class imbalanced problems difficult? An experimental study 2022**

* Difficulty recognizing instances of the minority class due to bias towards majority classes in standard learning algorithms.

# BRACID: a comprehensive approach to learning rules from imbalanced data 2011

<https://link.springer.com/article/10.1007/s10844-011-0193-0>

In this paper, the authors propose a novel algorithm called BRACID for induction of rule-based classifiers from imbalanced data. They argue that standard techniques for induction of rule-based classifiers are biased towards majority classes and have difficulties with correct recognition of the minority class. BRACID addresses this issue by incorporating a hybrid representation of rules and single examples, learning of rules, and a local classification strategy using nearest rules. The algorithm is evaluated on several imbalanced datasets and is shown to significantly outperform other well-known rule-based classifiers and approaches specialized for imbalanced data. Additionally, BRACID improves the support of minority class rules, leading to better recognition of minority class instances.

1. ML | Handling Imbalanced Data with SMOTE and Near Miss Algorithm in Python 2023

<https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/>

In Machine Learning and Data Science, imbalanced data distribution occurs when observations in one of the classes are much higher or lower than the other classes. This problem is prevalent in examples such as Fraud Detection, Anomaly Detection, and Facial recognition. Standard ML techniques tend to have a bias towards the majority class and ignore the minority class, leading to major misclassification of the minority class. To handle imbalanced data distribution, two commonly used algorithms are SMOTE and Near Miss Algorithm. SMOTE (Synthetic Minority Oversampling Technique) balances class distribution by randomly increasing minority class examples by replicating them. Near Miss Algorithm, on the other hand, balances class distribution by randomly eliminating majority class examples. Both techniques aim to improve the recognition of the minority class by balancing the class distribution.

* Overlapping regions between classes, which can make it harder to distinguish between minority and majority classes.

# On the class overlap problem in imbalanced data classification 2021

<https://www.sciencedirect.com/science/article/abs/pii/S0950705120307607#>

The article discusses the negative impact of class overlap on the performance of machine learning algorithms in imbalanced data classification. It presents a thorough experimental comparison of class overlap and class imbalance and provides a critical review of existing approaches to handle imbalanced datasets. The article emphasizes the need for further research towards handling class overlap in imbalanced datasets to effectively improve the performance of learning algorithms.

* Rare sub-concepts in minority classes, which can make it harder to identify and classify them.

The impact of data difficulty factors on classification of imbalanced and concept drifting data streams [2021](https://link.springer.com/article/10.1007/s10115-021-01560-w#article-info)  
  
<https://link.springer.com/article/10.1007/s10115-021-01560-w>

# 

The article talks about the problem of rare minority examples in classification tasks. It states that these examples are often harder to identify and classify because they are represented by a small number of instances. This is known as a "small disjuncts" problem. The authors propose a new approach to handle this problem by creating synthetic minority data, which helps to improve the performance of the classifier and reduce the error rate for rare classes. Additionally, the authors conduct experiments to study the influence of these challenges on predictions of representative online classifiers and conclude that existing classifiers are only partially capable of coping with these problems and new approaches are needed to address challenges posed by imbalanced data streams with rare sub-concepts.

* Rare minority examples located inside the majority class region, which can make it harder to identify and classify them.

# Learning class-imbalanced data with region-impurity synthetic minority oversampling technique 2022

<https://www.sciencedirect.com/science/article/abs/pii/S0020025522006612>

In "Learning class-imbalanced data with region-impurity synthetic minority oversampling technique" the authors propose a new algorithm, Region-Impurity Synthetic Minority Oversampling Technique (RIOT), that addresses the issue of identifying rare minority examples located inside the majority class region.  
RIOT performs better than other oversampling methods in terms of several model performance indicators.