Capítulo

4

Estratégias para Lidar com Desbalanceamento de Dados em Aprendizado de Máquina

Hector Batista Ribeiro, Leandro Oliveira da Silva e Ricardo de Andrade Lira Rabêlo

Abstract

Data imbalance is a critical issue in machine learning, impacting model performance due to uneven class representation. This chapter explores the causes and effects of imbalance, discussing preprocessing techniques such as resampling and synthetic data generation. It also covers modeling methods and weight adjustments to enhance performance. Furthermore, the chapter reviews specific evaluation metrics and advanced methodologies, including ensemble methods and reinforcement learning, providing practical insights for addressing imbalance and developing more balanced, accurate models.

Resumo

O desbalanceamento de dados é um desafio crucial em aprendizado de máquina, onde a desigualdade na representação das classes afeta a eficácia dos modelos preditivos. Este capítulo explora as causas e consequências do desbalanceamento, destacando técnicas de pré-processamento como a reamostragem e geração de dados sintéticos. Também aborda métodos de modelagem e ajustes de pesos para melhorar o desempenho. Além disso, examina métricas de avaliação específicas e metodologias avançadas, como ensemble methods e aprendizado por reforço, oferecendo uma visão prática para lidar com o problema e desenvolver modelos mais equilibrados e precisos.

4.1. Introdução

O desbalanceamento de dados é um desafio intrínseco e amplamente reconhecido na prática de aprendizado de máquina, caracterizado pela disparidade na distribuição das classes em um

conjunto de dados [Chaudhary 2023]. Em muitos cenários do mundo real, os dados disponíveis para treinamento de modelos preditivos não são distribuídos uniformemente entre as diferentes categorias ou classes, resultando em um desequilíbrio significativo [Azank 2020]. Este fenômeno não é meramente uma peculiaridade estatística, mas uma questão que pode afetar a eficácia e a precisão dos algoritmos de aprendizado de máquina de maneira substancial.

O desbalanceamento de dados refere-se a uma situação em que uma ou mais classes em um conjunto de dados são representadas com muito mais frequência do que outras. Em contextos binários, isso se traduz em uma classe majoritária que é muito mais abundante do que a classe minoritária [Chaudhary 2023]. Em cenários multiclasse, o problema pode se manifestar como uma discrepância significativa na frequência de ocorrência entre as classes. Esse desbalanceamento pode levar a uma série de desafios, como a degradação da performance dos modelos preditivos e a dificuldade em generalizar para dados não vistos.

A relevância do desbalanceamento de dados se manifesta em diversas áreas e aplicações práticas. Por exemplo, na detecção de fraudes financeiras, os eventos fraudulentos são relativamente raros em comparação com transações legítimas. Sem um tratamento adequado, um modelo treinado em um conjunto de dados desbalanceado pode apresentar um viés sistemático em favor da classe majoritária, comprometendo sua capacidade de identificar fraudes de forma eficaz. Similarmente, em diagnósticos médicos, a identificação de doenças raras pode ser prejudicada se a classe de interesse estiver sub-representada no conjunto de dados de treinamento [Azank 2020].

O impacto do desbalanceamento de dados nos modelos de aprendizado de máquina é multifacetado e pode afetar significativamente a performance dos sistemas preditivos [Hasib et al. 2020]. Quando um modelo é treinado em um conjunto de dados desbalanceado, ele tende a aprender a reconhecer predominantemente a classe majoritária, muitas vezes à custa da capacidade de identificar a classe minoritária. Esse viés pode se manifestar em métricas de desempenho, como precisão, revocação e pontuação F1, que podem ser enganosas se avaliadas apenas com base na classe majoritária [L. A. Jeni et al. 2013].

Além disso, o desbalanceamento pode levar a uma avaliação inadequada do modelo. Por exemplo, um modelo que apresenta uma alta taxa de precisão pode ainda ter um desempenho fraco na detecção da classe minoritária, resultando em uma alta taxa de falsos negativos [Hasib et al. 2020]. Em contextos críticos, como o monitoramento de segurança e o diagnóstico de doenças, isso pode ter consequências graves, como a falha em identificar eventos adversos ou condições de saúde que requerem atenção urgente.

Em diversas aplicações reais, o desbalanceamento de dados é a norma e não a exceção. Na área da saúde, por exemplo, conjuntos de dados usados para diagnosticar doenças raras muitas vezes contêm muito mais exemplos de indivíduos saudáveis do que de indivíduos doentes [A. K. Ilavarasi 2020]. Da mesma forma, em sistemas de detecção de fraudes financeiras, as transações legítimas são vastamente mais numerosas do que as fraudulentas [Rubaidi Z. 2022]. Em ambos os casos, a incapacidade de detectar as classes minoritárias pode ter consequências significativas, seja na forma de diagnósticos incorretos ou perdas financeiras substanciais.

A necessidade de tratar o desbalanceamento de dados se torna ainda mais crucial à medida que os sistemas de aprendizado de máquina são cada vez mais integrados em processos críticos de tomada de decisão. Sem técnicas adequadas para lidar com o desbalanceamento, os modelos podem apresentar uma performance ilusória de alta precisão, enquanto na verdade falham em detectar ou prever os casos mais críticos e raros. Portanto, é imperativo que cientistas de dados e engenheiros de aprendizado de máquina estejam equipados com estratégias eficazes para mitigar os efeitos do desbalanceamento de dados.

A abordagem do desbalanceamento envolve uma combinação de técnicas de pré-processamento de dados, ajustes de modelagem e métricas de avaliação específicas. A utilização dessas técnicas não apenas melhora a performance dos modelos, mas também contribui para a criação de sistemas preditivos mais justos e equilibrados [Thabtah F. 2019]. As seções seguintes deste capítulo irão explorar em detalhes as diferentes abordagens para lidar com o desbalanceamento de dados, incluindo técnicas de reamostragem, geração de dados sintéticos, ajustes de pesos e métodos avançados como o aprendizado por reforço.

Neste capítulo, exploraremos uma variedade de técnicas e abordagens que foram desenvolvidas para enfrentar o problema do desbalanceamento de dados. Estas técnicas podem ser categorizadas em três grandes áreas:

Técnicas de Pré-processamento de Dados: Incluem métodos como reamostragem (oversampling e undersampling), aumento de dados e seleção de características que ajudam a equilibrar as classes antes da modelagem.

Técnicas de Modelagem: Envolvem a modificação dos algoritmos de aprendizado de máquina ou a introdução de novos algoritmos que são mais robustos a dados desbalanceados, incluindo métodos baseados em custo e o uso de métricas de avaliação específicas.

Métodos Avançados: Abrangem abordagens mais recentes e sofisticadas, como aprendizado semi-supervisionado, deep learning e a utilização de arquiteturas especializadas para lidar com o desbalanceamento.

O capítulo será organizado para proporcionar uma compreensão abrangente e prática sobre como lidar com o desbalanceamento de dados. Começaremos com uma discussão sobre a natureza do desbalanceamento, como ele pode ser quantificado e os problemas que ele causa. Em seguida, detalharemos as técnicas de pré-processamento de dados, como reamostragem, gerenciamento de dados e aumento de dados. Exploraremos algoritmos e métodos específicos para dados desbalanceados, bem como as métricas apropriadas para avaliar modelos treinados com esses dados. Analisaremos técnicas avançadas, incluindo aprendizado profundo e semi-supervisionado. Apresentaremos estudos de caso com exemplos reais e lições aprendidas. Introduziremos ferramentas e bibliotecas populares usadas para lidar com desbalanceamento de dados, incluindo exemplos de código. Finalmente, concluiremos com um resumo das principais técnicas discutidas e uma análise dos desafios e tendências futuras na área.

4.2. Tipos de Desbalanceamento

O desbalanceamento de dados é um fenômeno que pode assumir diversas formas, cada uma com suas implicações e desafios específicos. Neste capítulo, exploraremos os diferentes tipos de desbalanceamento, fornecendo uma compreensão mais profunda das suas características e como eles afetam o desempenho dos modelos de aprendizado de máquina. A compreensão desses tipos é fundamental para selecionar as técnicas apropriadas para lidar com o desbalanceamento em diferentes cenários.

4.2.1. Desbalanceamento de Classes

O tipo mais comum de desbalanceamento de dados é o desbalanceamento de classes. Este ocorre quando as diferentes classes em um conjunto de dados têm quantidades significativamente variadas de exemplos [Longadge R. 2013]. Por exemplo, em um conjunto de dados de diagnóstico médico, pode haver milhares de casos de pacientes saudáveis e apenas alguns casos de uma doença rara. Esse desbalanceamento pode levar os modelos a se tornarem altamente enviesados em favor da classe majoritária, resultando em baixa capacidade de detecção da classe minoritária. É crucial reconhecer que o desbalanceamento de classes pode afetar diversas métricas de desempenho do modelo, como a precisão, recall e F1-score, de maneira desproporcional.

4.2.2. Desbalanceamento em Séries Temporais

O desbalanceamento em séries temporais refere-se a situações onde os dados desbalanceados não são apenas desiguais em quantidade, mas também distribuídos de maneira não uniforme ao longo do tempo [H. Cao et al. 2013]. Em cenários de séries temporais, como a detecção de anomalias em dados financeiros ou sistemas de monitoramento industrial, eventos raros podem ocorrer esporadicamente. Por exemplo, uma falha em um sistema de monitoramento pode ser um evento raro que ocorre apenas ocasionalmente, enquanto o comportamento normal é muito mais frequente. O desafio aqui é lidar com a variação temporal e garantir que o modelo possa detectar esses eventos raros sem se perder nos dados normais.

4.2.3. Desbalanceamento em Dados Multiclasse

Em problemas de classificação multiclasse, o desbalanceamento pode ocorrer quando algumas classes estão muito mais representadas do que outras [Mathew R. and Gunasundari R. 2021]. Por exemplo, em um conjunto de dados para classificação de espécies de plantas, algumas espécies podem ter centenas de amostras, enquanto outras podem ter apenas algumas. Esse tipo de desbalanceamento pode tornar a tarefa de classificação mais complexa, pois o modelo pode ter dificuldade em aprender a distinguir as classes minoritárias corretamente. A abordagem para lidar com o desbalanceamento em problemas multiclasse pode exigir técnicas adaptadas, como reamostragem específica para cada classe ou o uso de métricas de avaliação que considerem a importância relativa de cada classe.

4.2.4. Desbalanceamento de Rótulos em Dados Multimodais

No contexto de dados multimodais, onde diferentes tipos de dados (como texto, imagem e áudio) são usados, o desbalanceamento pode ocorrer em diferentes modalidades de forma desigual [S. Pouyanfar et al. 2019]. Por exemplo, em um sistema de reconhecimento de

emoção que utiliza tanto imagens quanto texto, pode haver muito mais dados disponíveis para uma modalidade (como imagens) do que para outra (como texto). Esse desbalanceamento pode complicar a integração e a fusão de informações de diferentes fontes e exigir técnicas específicas para equilibrar a influência de cada modalidade.

Cada tipo de desbalanceamento apresenta seus próprios desafios e requer abordagens específicas para garantir que os modelos de aprendizado de máquina possam realizar previsões precisas e confiáveis. A compreensão desses diferentes tipos de desbalanceamento é essencial para a escolha e implementação das técnicas de tratamento adequadas. No próximo capítulo, abordaremos as técnicas de pré-processamento de dados que podem ser aplicadas para lidar com esses tipos de desbalanceamento, proporcionando estratégias para equilibrar os conjuntos de dados e melhorar o desempenho dos modelos.

4.3. Técnicas de Pré-processamento de Dados

No pré-processamento de dados, a reamostragem é uma técnica essencial para lidar com o desbalanceamento de classes. Esta técnica pode ser dividida em abordagens de oversampling, undersampling e métodos combinados. Além disso, a geração de dados sintéticos é uma estratégia avançada que inclui técnicas como SMOTE e ADASYN. Cada uma dessas abordagens tem suas próprias características e aplicações, que serão discutidas a seguir.

4.3.1. Reamostragem

A reamostragem busca equilibrar o número de exemplos entre classes majoritárias e minoritárias para melhorar a performance do modelo. Existem três abordagens principais para a reamostragem: *oversampling*, *undersampling* e métodos combinados.

4.3.1.1. Oversampling: Aumentando a Quantidade da Classe Minoritária

De acordo com [Thabtah 2019] e [Hasib et al. 2020], *oversampling* é uma técnica que aumenta a quantidade de exemplos da classe minoritária. O objetivo é criar um conjunto de dados mais equilibrado ao adicionar exemplos adicionais que ajudam o modelo a aprender melhor sobre a classe minoritária. Uma abordagem comum de oversampling é a geração de novos exemplos sintéticos para a classe minoritária, o que pode ajudar a melhorar a capacidade do modelo de identificar a classe minoritária sem alterar os exemplos da classe majoritária [Hasib et al. 2020].

4.3.1.2. Undersampling: Reduzindo a Quantidade da Classe Majoritária

De acordo com [Thabtah 2019] e [Hasib et al. 2020], *undersampling* envolve a redução da quantidade de exemplos da classe majoritária para equilibrar o conjunto de dados. Isso é feito removendo exemplos da classe majoritária até que a proporção de exemplos entre as classes seja mais equilibrada. Embora o undersampling possa reduzir o viés em favor da classe majoritária, é importante notar que essa abordagem pode levar à perda de informações valiosas e diminuir a capacidade do modelo de aprender sobre a variabilidade da classe majoritária [Thabtah 2019].

4.3.1.3. Métodos Combinados: Oversampling e Undersampling Combinados

Métodos combinados utilizam uma combinação de oversampling e undersampling para criar um conjunto de dados mais equilibrado, como visto em [Junsomboon N. and Phienthrakul T. 2017] e [H. Shamsudin et al. 2020]. Uma abordagem comum é aplicar oversampling na classe minoritária e undersampling na classe majoritária simultaneamente. Isso pode ajudar a mitigar os problemas de desbalanceamento enquanto mantém uma quantidade razoável de dados para ambas as classes. Métodos combinados podem oferecer uma solução mais equilibrada do que usar apenas uma das técnicas isoladamente.

4.3.1.4. Geração de Dados Sintéticos

A geração de dados sintéticos é uma abordagem avançada para lidar com o desbalanceamento de dados, especialmente quando é difícil ou impossível obter mais dados reais da classe minoritária [Hasib et al. 2020]. As técnicas de geração de dados sintéticos criam novos exemplos a partir dos existentes para aumentar a representação da classe minoritária.

4.3.1.5. SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE (Synthetic Minority Over-sampling Technique) é uma técnica popular para gerar exemplos sintéticos da classe minoritária. Em vez de simplesmente duplicar exemplos existentes, o SMOTE cria novos exemplos interpolando entre exemplos da classe minoritária [Chawla 2002]. Para cada exemplo na classe minoritária, o SMOTE seleciona seus vizinhos mais próximos e cria novos exemplos ao longo das linhas que conectam o exemplo original com seus vizinhos. Isso ajuda a criar uma fronteira de decisão mais suave e generalizável para a classe minoritária.

4.3.1.6. ADASYN (Adaptive Synthetic Sampling)

ADASYN (Adaptive Synthetic Sampling) é uma variação do SMOTE que se concentra em gerar mais exemplos sintéticos em regiões onde a densidade da classe minoritária é baixa. O ADASYN adapta a quantidade de dados sintéticos gerados com base na dificuldade de aprendizado em diferentes regiões do espaço de características [He 2008]. Ao gerar mais exemplos em regiões problemáticas, o ADASYN pode ajudar a melhorar a capacidade do modelo de aprender a distinguir a classe minoritária em áreas onde há maior confusão.

4.3.1.7. Exemplos e Aplicações

A reamostragem e a geração de dados sintéticos são amplamente utilizadas em várias áreas para enfrentar o desbalanceamento de dados.

- **Detecção de Fraude:** No setor financeiro, onde as fraudes são raras comparadas às transações legítimas, técnicas de oversampling como SMOTE são usadas para criar exemplos sintéticos de fraudes e melhorar a capacidade do modelo de detectar atividades fraudulentas, como visto em [Mqadi 2021].
- **Diagnóstico Médico:** Em diagnósticos de doenças raras, onde há poucos casos positivos em relação aos negativos, oversampling e técnicas de geração de dados sintéticos ajudam a equilibrar o conjunto de dados para melhorar a acurácia dos modelos de diagnóstico, como visto em [Kosolwattana T. et al. 2022].

- **Reconhecimento de Anomalias:** Em sistemas de monitoramento industrial, onde anomalias podem ocorrer raramente, undersampling da classe majoritária e oversampling da classe minoritária ajudam a criar um conjunto de dados equilibrado para treinar modelos que detectam falhas com mais eficácia, como visto em [J. -R. Jiang and Y. -T. Chen 2022].

Cada técnica tem suas vantagens e desvantagens, e a escolha da abordagem apropriada pode depender das características específicas do conjunto de dados e dos objetivos do projeto. No próximo capítulo, exploraremos como essas técnicas de pré-processamento se integram com os algoritmos de modelagem para enfrentar o desbalanceamento de dados de forma eficaz.

4.4.Técnicas de modelagem

Após o pré-processamento dos dados desbalanceados, o próximo passo crucial é selecionar e ajustar algoritmos de modelagem que possam lidar eficazmente com o desbalanceamento. Este capítulo explora as principais técnicas de modelagem que podem ser aplicadas para enfrentar o desbalanceamento de dados. Serão discutidos os algoritmos robustecidos para desbalanceamento, ajustes específicos em métodos baseados em árvores e redes neurais, e a importância do ajuste de pesos durante o treinamento.

4.4.1. Algoritmos Robustecidos para Desbalanceamento

Algoritmos robustecidos para desbalanceamento são projetados ou ajustados para operar eficientemente em cenários onde há uma disparidade significativa entre as classes. Esses algoritmos incorporam técnicas internas para mitigar o viés em favor da classe majoritária, tornando-os adequados para problemas de classificação com classes desbalanceadas.

4.4.1.1. Métodos Baseados em Árvores: XGBoost, Random Forest

Métodos baseados em árvores, como XGBoost e Random Forest, são particularmente populares devido à sua flexibilidade e capacidade de lidar com dados desbalanceados de forma eficaz. Na tabela abaixo é apresentada uma descrição de cada um deles.

Tabela 4.1. Métodos baseados em árvores.

Método	Definição
XGBoost (Extreme Gradient Boosting)	Algoritmo de boosting que combina várias árvores de decisão fracas para formar um modelo poderoso [Pristyanto Y. et al. 2023]. Para lidar com o desbalanceamento, o XGBoost permite o ajuste do parâmetro scale_pos_weight, que pondera o impacto das classes minoritárias no erro de treinamento. Além disso, o XGBoost suporta a amostragem de dados de forma que as classes minoritárias possam ser mais bem representadas durante o treinamento, mitigando o viés em favor da classe majoritária.

Random Forest	Algoritmo de ensemble que combina várias árvores de decisão independentes. Para lidar com o desbalanceamento, uma abordagem comum é ajustar a frequência de amostragem de cada árvore, de modo que as classes minoritárias sejam amostradas com mais frequência [M. Bader-El-Den et al. 2019]. Isso é feito ajustando o parâmetro class_weight durante o treinamento, garantindo que a importância das classes minoritárias seja aumentada, e o modelo seja mais sensível a elas.

4.4.1.2. Métodos Baseados em Redes Neurais: Arquiteturas e Ajustes Específicos

Redes neurais oferecem flexibilidade e poder para modelagem em cenários complexos, mas podem exigir ajustes específicos para lidar com desbalanceamento de dados.

Arquiteturas Profundas: Redes neurais profundas, como CNNs (Redes Neurais Convolucionais) e RNNs (Redes Neurais Recorrentes), podem ser adaptadas para lidar com desbalanceamento de dados [J. Justin and K. Taghi 2019]. Uma estratégia comum é adicionar camadas específicas que ajustam a sensibilidade do modelo para a classe minoritária. Por exemplo, em uma CNN para reconhecimento de imagens, pode-se ajustar os filtros convolucionais ou as camadas de pooling para destacar características mais representativas da classe minoritária.

Ajustes de Hiperparâmetros: Ajustar hiperparâmetros como a taxa de aprendizado e a regularização também pode ajudar a melhorar a performance em dados desbalanceados [Z. Fan et al. 2022]. A regularização L2, por exemplo, pode ser ajustada para evitar que o modelo se torne excessivamente confiante em previsões para a classe majoritária. O dropout também pode ser usado para reduzir o overfitting e melhorar a generalização do modelo em relação à classe minoritária.

4.4.2. Ajuste de Pesos

O ajuste de pesos é uma técnica eficaz para lidar com desbalanceamento ao influenciar diretamente o processo de aprendizado do modelo, garantindo que as classes minoritárias sejam tratadas de forma justa durante o treinamento.

4.4.2.1. Ponderação de Classes: Ajuste de Pesos Durante o Treinamento

A ponderação de classes envolve ajustar os pesos atribuídos a cada classe durante o processo de treinamento, de modo que o erro nas classes minoritárias tenha um impacto maior na atualização dos parâmetros do modelo. Isso pode ser implementado de várias maneiras:

Loss Function (Função de Perda): a função de perda pode ser ajustada para incluir pesos para cada classe [J. Justin and K. Taghi 2019]. Por exemplo, na entropia cruzada ponderada, cada classe recebe um peso inversamente proporcional à sua frequência no conjunto de dados. Isso força o modelo a penalizar mais os erros nas classes minoritárias, promovendo um aprendizado mais equilibrado [M. R. Rezaei-Dastjerdehei et al. 2020].

Custom Loss Functions: em casos mais complexos, funções de perda personalizadas podem ser criadas para incorporar diferentes penalidades para erros em classes minoritárias [J. Justin and K. Taghi 2019]. Por exemplo, em um cenário de detecção de fraudes, a perda pode ser ajustada para penalizar falsos negativos (fraudes não detectadas) muito mais do que falsos positivos, refletindo a gravidade do erro.

4.5. Técnicas de Avaliação

A avaliação de modelos treinados em conjuntos de dados desbalanceados requer a utilização de métricas e métodos que sejam sensíveis à distribuição desigual entre as classes. Este capítulo explora as principais técnicas de avaliação utilizadas para medir a performance de modelos em cenários de desbalanceamento de dados. Discutiremos métricas como a Curva ROC e AUC, a matriz de confusão, precisão, revocação, F1-Score, balanced accuracy, e a importância da validação cruzada estratificada.

4.5.1. Métricas de Avaliação para Dados Desbalanceados

Em um cenário de desbalanceamento de dados, métricas tradicionais como a acurácia podem ser enganosas, pois um modelo que simplesmente prevê a classe majoritária pode obter uma alta acurácia sem ser efetivo em identificar a classe minoritária [L. A. Jeni et al. 2013]. Portanto, é essencial utilizar métricas que ofereçam uma visão mais equilibrada do desempenho do modelo.

4.5.1.1. Curva ROC e AUC

A Curva ROC (*Receiver Operating Characteristic*) é uma ferramenta poderosa para avaliar a capacidade de um modelo de distinguir entre as classes. A curva ROC é um gráfico que plota a taxa de verdadeiros positivos contra a taxa de falsos positivos para diferentes limiares de classificação [Rodrigues 2018]. A principal vantagem da curva ROC é que ela não é afetada pelo desbalanceamento de classes, uma vez que mede a capacidade de discriminação do modelo.

A AUC (*Area Under the ROC Curve*) representa a área sob a curva ROC e varia de 0 a 1 [Rodrigues 2018]. Um AUC de 0,5 indica que o modelo não tem capacidade de discriminação (equivalente a um palpite aleatório), enquanto um AUC próximo de 1 indica excelente discriminação entre as classes. Em cenários de desbalanceamento, a AUC é uma métrica robusta, pois não é influenciada pela distribuição das classes.

4.5.1.2. Matriz de Confusão

A matriz de confusão é uma representação tabular que mostra o desempenho do modelo em termos de verdadeiros positivos (TP), falsos positivos (FP), verdadeiros negativos (TN) e falsos negativos (FN) [De Souza 2019]. A matriz de confusão permite calcular diversas métricas importantes para avaliação em cenários de desbalanceamento. A matriz é detalhada na tabela abaixo:

Tabela 2. Matriz de confusão.

Caso	Definição
Verdadeiros Positivos	Casos em que a classe positiva é corretamente identificada pelo modelo.
Falsos Positivos	Casos em que a classe negativa é incorretamente identificada como positiva.
Verdadeiros Negativos	Casos em que a classe negativa é corretamente identificada pelo modelo
Falsos Negativos	Casos em que a classe positiva é incorretamente identificada como negativa.

A matriz de confusão é a base para calcular métricas como precisão, revocação, e F1-Score, que são essenciais para a avaliação em contextos de desbalanceamento.

4.5.1.3. Precisão, Revocação e F1-Score

Precisão (Precision): é a proporção de verdadeiros positivos sobre todos os exemplos que foram classificados como positivos [De Souza 2019]. A precisão é importante em cenários onde o custo de um falso positivo é alto.

Revocação (Recall): também conhecida como sensibilidade ou taxa de verdadeiros positivos, a revocação mede a proporção de verdadeiros positivos sobre todos os exemplos que realmente são positivos [De Souza 2019]. Revocação é crucial quando o custo de um falso negativo é alto.

F1-Score: o F1-Score é a média harmônica entre a precisão e a revocação [De Souza 2019]. É uma métrica útil quando há necessidade de um equilíbrio entre precisão e revocação, especialmente em casos onde as classes estão desbalanceadas.

Essas métricas oferecem uma visão mais completa e equilibrada do desempenho do modelo em relação às classes minoritárias e majoritárias.

4.5.1.4. Balanced Accuracy

A balanced accuracy é uma métrica que corrige o viés presente na acurácia tradicional em cenários de desbalanceamento. Ela é definida como a média das taxas de verdadeiros positivos para cada classe, proporcionando uma avaliação mais justa do modelo em relação às classes minoritárias [Olugbenga 2023].

A balanced accuracy é especialmente útil quando a distribuição das classes é altamente desbalanceada, garantindo que o modelo seja avaliado com base em sua capacidade de predizer corretamente tanto as classes majoritárias quanto as minoritárias.

4.5.2. Validação Cruzada Estratificada

A validação cruzada estratificada é uma técnica que assegura que cada subdivisão (fold) dos dados em um processo de validação cruzada mantém a mesma proporção de classes que o conjunto de dados original [Muralidhar 2021]. Isso é particularmente importante em cenários

de desbalanceamento, pois garante que cada fold usado no treinamento e teste do modelo represente adequadamente a distribuição das classes.

Em um cenário de validação cruzada tradicional, as subdivisões dos dados podem, por acaso, resultar em distribuições desbalanceadas, o que pode levar a uma avaliação enganosa da performance do modelo. A validação cruzada estratificada previne essa situação ao preservar as proporções de classes em cada fold, resultando em uma avaliação mais confiável e representativa.

4.5.3. Importância das Técnicas de Avaliação

As técnicas de avaliação discutidas neste capítulo são fundamentais para medir a eficácia de modelos treinados em conjuntos de dados desbalanceados. O uso de métricas apropriadas, como a curva ROC e AUC, matriz de confusão, precisão, revocação, F1-Score, e balanced accuracy, combinado com métodos robustos como a validação cruzada estratificada, garante uma avaliação mais justa e precisa. Essas práticas permitem que os desenvolvedores e pesquisadores compreendam melhor o desempenho de seus modelos em cenários desafiadores, levando a soluções mais eficazes e generalizáveis para problemas de classificação desbalanceada.

4.6. Estratégias Avançadas

Nesta seção, serão abordadas estratégias avançadas que vão além das técnicas tradicionais de pré-processamento e modelagem, oferecendo soluções mais sofisticadas para lidar com o desbalanceamento de dados. Serão explorados métodos de ensemble, como bagging e boosting, modelos de votação e voto ponderado, além de técnicas de aprendizado por transferência (transfer learning). Também discutiremos outros métodos avançados, como aprendizado semi-supervisionado e aprendizado por reforço, que têm se mostrado promissores em cenários de desbalanceamento.

4.6.1. Ensemble Methods

Os métodos de ensemble combinam múltiplos modelos para melhorar a precisão e a robustez das previsões [Galar M. 2012]. Essas técnicas são particularmente eficazes em cenários de dados desbalanceados, pois diferentes modelos podem capturar diferentes aspectos das classes minoritárias.

O bagging é uma técnica que cria múltiplos modelos treinados em diferentes subconjuntos dos dados, gerados por amostragem com reposição (bootstrap) [Galar M. 2012]. Um exemplo clássico de bagging é o Random Forest. No contexto de dados desbalanceados, o bagging pode ser ajustado para amostrar a classe minoritária com mais frequência, aumentando sua representação nos modelos individuais.

O boosting é uma técnica de ensemble que cria modelos sequencialmente, onde cada novo modelo tenta corrigir os erros dos modelos anteriores [Galar M. 2012]. XGBoost e AdaBoost são exemplos de algoritmos de boosting. Para lidar com o desbalanceamento, o boosting pode ajustar os pesos das observações com base em seu erro de classificação, dando mais importância às instâncias da classe minoritária.

4.6.1.2. Modelos de Votação

Os modelos de votação combinam as previsões de vários modelos para tomar uma decisão final [A. Dogan and D. Birant 2019]. Em cenários de desbalanceamento, é comum utilizar *weighted voting* (votação ponderada), onde as previsões de modelos que se saem melhor na detecção da classe minoritária recebem maior peso na decisão final.

No modelo de votação mais simples, chamado de *majority voting*, a classe predita pela maioria dos modelos é selecionada [A. Dogan and D. Birant 2019]. Essa abordagem pode ser insuficiente em cenários desbalanceados, pois a classe majoritária tende a dominar.

No *weighted voting*, cada modelo recebe um peso proporcional à sua performance em detectar a classe minoritária [A. Dogan and D. Birant 2019]. Por exemplo, se um modelo tem alta precisão para a classe minoritária, suas previsões têm maior influência no voto final.

4.6.2. Transfer Learning

O *transfer learning* é uma técnica onde um modelo pré-treinado em um grande conjunto de dados (geralmente equilibrado) é ajustado para uma nova tarefa com dados potencialmente desbalanceados [Al-Stouhi S. and Reddy C. 2015]. Essa técnica tem se mostrado eficaz em cenários onde há pouca disponibilidade de dados para a classe minoritária.

4.6.2.1. Aplicação em Cenários de Dados Desbalanceados

Fine-tuning de modelos pré-treinados: em *transfer learning*, um modelo, como uma rede neural profunda pré-treinada em um grande conjunto de dados, pode ser ajustado (fine-tuned) em um novo conjunto de dados desbalanceado. Esse processo permite que o modelo retenha o conhecimento geral adquirido no treinamento inicial e o aplique para melhorar o desempenho em classes minoritárias.

Feature extraction: outra aplicação de *transfer learning* é a extração de características (features) de modelos pré-treinados. As características extraídas podem ser usadas como entrada para outro modelo, como uma SVM ou um random forest, treinado especificamente para a tarefa com dados desbalanceados.

4.6.3. Outros Métodos Avançados

Além das técnicas mencionadas previamente, outros métodos avançados têm sido desenvolvidos para lidar com o desbalanceamento de dados, aproveitando técnicas mais recentes de aprendizado de máquina.

4.6.3.1. Aprendizado Semi-Supervisionado

O aprendizado semi-supervisionado utiliza uma combinação de dados rotulados e não rotulados para treinar modelos [Gui Q. et al. 2024]. Em cenários de desbalanceamento, onde a classe minoritária pode ser sub-representada nos dados rotulados, o aprendizado semi-supervisionado pode ajudar a expandir o conjunto de treinamento.

Label Propagation: técnicas de propagação de rótulos podem ser usadas para inferir rótulos para o conjunto de dados não rotulado, potencialmente aumentando a representação da classe minoritária [Gui Q. et al. 2024].

Pseudo-Labeling: em *pseudo-labeling*, o modelo inicial faz previsões sobre os dados não rotulados, e essas previsões são então tratadas como rótulos para o próximo ciclo de treinamento, ajudando a reforçar a detecção da classe minoritária [Gui Q. et al. 2024].

4.6.3.2. Aprendizado por Reforço

O aprendizado por reforço envolve treinar um agente para tomar decisões em um ambiente, maximizando uma recompensa ao longo do tempo [Lin E. et al. 2020]. Essa abordagem pode ser adaptada para lidar com desbalanceamento de dados, especialmente em contextos onde as decisões precisam ser adaptativas e dinâmicas.

Reward Shaping: no contexto de dados desbalanceados, a função de recompensa pode ser ajustada para penalizar fortemente os erros na detecção da classe minoritária, incentivando o agente a focar mais nessa classe [Lin E. et al. 2020].

Q-Learning e Deep Q-Networks (DQN): técnicas de aprendizado por reforço, como Q-Learning e DQN, podem ser aplicadas em cenários onde as instâncias da classe minoritária são raras, treinando o agente para reconhecer e priorizar a classe minoritária em suas decisões [Lin E. et al. 2020].

4.7. Desafios e Considerações Finais

As técnicas discutidas ao longo deste material oferecem o conhecimento necessário para lidar com o desbalanceamento de dados. No entanto, essas técnicas vêm com suas próprias limitações e desafios, que devem ser cuidadosamente consideradas ao selecionar e implementar uma estratégia. Nesta última seção serão discutidas as limitações das técnicas existentes, os aspectos a serem considerados ao escolher uma abordagem, e as tendências futuras que podem moldar o campo nos próximos anos.

4.7.1. Limitações das Técnicas

Apesar dos avanços significativos na área de desbalanceamento de dados, as técnicas atuais não são universais e apresentam limitações que podem impactar a eficácia em diferentes cenários.

4.7.1.1. Reamostragem

Técnicas de reamostragem, como oversampling e undersampling, são simples e eficazes, mas podem introduzir novos desafios. O oversampling pode levar a overfitting, especialmente se forem gerados muitos exemplos sintéticos da classe minoritária. Já o undersampling pode resultar na perda de informações valiosas, comprometendo o desempenho geral do modelo.

4.7.1.2. Algoritmos de Modelagem

Embora algoritmos robustecidos, como os métodos baseados em árvores (XGBoost, Random Forest), sejam eficazes em muitos casos, eles podem se tornar complexos e difíceis de

interpretar. Além disso, o ajuste de hiperparâmetros para lidar com desbalanceamento pode ser um processo longo e não garantir uma solução ideal. Modelos de redes neurais, por sua vez, podem ser altamente sensíveis à configuração dos pesos e exigem uma quantidade significativa de dados para evitar underfitting.

4.7.1.3. Métodos Avançados

Técnicas avançadas, como aprendizado semi-supervisionado e por reforço, requerem grandes volumes de dados e processamento, tornando-se impraticáveis em cenários onde os recursos são limitados. Além disso, essas abordagens podem ser difíceis de implementar e interpretar, especialmente para quem não tem familiaridade com os conceitos subjacentes.

4.7.2. Aspectos a Serem Considerados ao Escolher uma Técnica

A escolha da técnica apropriada para lidar com o desbalanceamento de dados depende de uma série de fatores, que devem ser ponderados com cuidado.

4.7.2.1. Natureza dos Dados

A natureza dos dados, incluindo o grau de desbalanceamento, o número de características e a complexidade das relações entre as variáveis, é um fator crítico na escolha da técnica. Para conjuntos de dados com alto grau de desbalanceamento, técnicas que ajustam os pesos ou que combinam oversampling e undersampling podem ser mais eficazes.

4.7.2.2. Objetivo do Modelo

O objetivo do modelo, seja ele focado em maximizar precisão, revocação, ou uma combinação das duas, também orienta a escolha da técnica. Modelos que exigem alta precisão podem se beneficiar de técnicas de reamostragem ou ajuste de pesos, enquanto aqueles que priorizam a revocação podem requerer algoritmos mais complexos ou métodos de ensemble.

4.7.2.3. Recursos Disponíveis

Os recursos computacionais e o tempo disponível para o desenvolvimento do modelo são considerações importantes. Técnicas mais simples, como reamostragem ou ajuste de pesos, podem ser implementadas rapidamente, mas podem não oferecer a melhor performance. Métodos avançados, como deep learning ou aprendizado por reforço, requerem mais tempo e poder computacional, mas podem oferecer ganhos significativos em precisão e generalização.

4.7.3. Tendências Futuras

O campo de aprendizado de máquina continua a evoluir rapidamente, e novas abordagens para lidar com o desbalanceamento de dados estão emergindo. Nesta seção, discutimos as tendências e tecnologias promissoras que podem definir como tratamos dados desbalanceados futuramente.

4.7.3.1. Aprendizado Federado

O aprendizado federado é uma abordagem emergente que permite que modelos sejam treinados em dados distribuídos entre várias fontes, preservando a privacidade dos dados [C. Nguyen et al. 2021]. Essa técnica tem o potencial de melhorar o desempenho em cenários de desbalanceamento ao permitir que dados minoritários de várias fontes sejam combinados para criar um modelo mais robusto, sem a necessidade de centralizar os dados.

4.7.3.2. Modelos Auto-supervisionados

Modelos auto-supervisionados, que podem aprender representações úteis dos dados sem a necessidade de rótulos extensivos, estão ganhando popularidade [Bergmann 2023]. Essas técnicas podem ajudar a superar o problema do desbalanceamento, permitindo que o modelo aprenda com grandes volumes de dados não rotulados e aplicando esse conhecimento para melhorar a detecção da classe minoritária.

4.8. Referências

[Abdi and Hashemi 2015] Abdi, L., Hashemi, S.: To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE transactions on Knowledge and Data Engineering 28(1), 238–251 (2015).

[Coelho et al. 2022] Coelho, G.M.C., Ramos, A.C., de Sousa, J., Cavaliere, M., de Lima, M.J., Mangeth, A., Frajhof, I.Z., Cury, C., Casanova, M.A.: Text classification in the brazilian legal domain. In: Intern. Conference on Enterprise Information Systems. pp. 355–363 (2022).

[Feng et al. 2019] Feng, W., Dauphin, G., Huang, W., Quan, Y., Bao, W., Wu, M., Li, Q.: Dynamic synthetic minority over-sampling technique-based rotation forest for the classification of imbalanced hyperspectral data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12(7), 2159–2169 (2019).

[Chaudhary 2023] Chaudhary, K.: How to Deal with Imbalanced Data in Classification. Medium. Disponível em:https://medium.com/game-of-bits/how-to-deal-with-imbalanced-data-in-classification-bd03cfc66066. Acesso em: 23 ago. 2024.

[Azank 2020] Azank F.: Dados Desbalanceados — O que são e como lidar com eles. Disponível em: https://medium.com/turing-talks/dados-desbalanceados-o-que-s%C3%A3o-e-como-evit%C3%A1-los-43df4f49732b. Acesso em: 23 ago. 2024.

[Hasib et al. 2020] Hasib K., Iqbal Md., Shah F., Mahmud J., Popel M., Showrov Md. I. H., Ahmed S., Rahman O. (2020). A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem. Journal of Computer Science. 16. 1546 - 1557. 10.3844/jcssp.2020.1546.1557.

[A. K. Ilavarasi 2020] A. K. Ilavarasi (2020). Class imbalance learning for Identity Management in Healthcare. In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2020, pp. 995-1000, doi: 10.1109/I-SMAC49090.2020.9243420.

[Rubaidi Z. 2022] Rubaidi Zainab, Ben Ammar Boulbaba, Ben Aouicha (2022). Fraud Detection Using Large-scale Imbalance Dataset. In: International Journal on Artificial Intelligence Tools. 31. 10.1142/S0218213022500373.

[L. A. Jeni et al. 2013] L. A. Jeni, J. F. Cohn, F. De La Torre (2013). Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2013, pp. 245-251, doi: 10.1109/ACII.2013.47.

[Thabtah F. 2019] Thabtah F., Hammoud S., Kamalov F., Gonsalvesv A. (2019). Data Imbalance in Classification: Experimental Evaluation. In: Information Sciences. 513. 10.1016/j.ins.2019.11.004.

[Longadge R. and Dongre S. 2013] Longadge Rushi, Dongre Snehlata. (2013). Class Imbalance Problem in Data Mining Review. In: Int. J. Comput. Sci. Netw.. 2.

[H. Cao et al. 2013] H. Cao, X. -L. Li, D. Y. -K. Woon and S. -K. Ng (2013). Integrated Oversampling for Imbalanced Time Series Classification. In: IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 12, pp. 2809-2822, Dec. 2013, doi: 10.1109/TKDE.2013.37.

[Mathew R. and Gunasundari R. 2021] R. Mary Mathew, R. Gunasundari (2021). A Review on Handling Multiclass Imbalanced Data Classification In Education Domain. In: 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2021, pp. 752-755, doi: 10.1109/ICACITE51222.2021.9404626.

[S. Pouyanfar et al. 2019] S. Pouyanfar, T. Wang and S. -C. Chen (2019). A Multi-label Multimodal Deep Learning Framework for Imbalanced Data Classification. In: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 2019, pp. 199-204, doi: 10.1109/MIPR.2019.00043.

[Junsomboon N. and Phienthrakul T. 2017] Junsomboon Nutthaporn and Phienthrakul Tanasanee (2017). Combining Over-Sampling and Under-Sampling Techniques for Imbalance Dataset. In: Proceedings of the 9th International Conference on Machine Learning and Computing, 2017, pp. 243-247, doi: 10.1145/3055635.3056643.

[H. Shamsudin et al. 2020] H. Shamsudin, U. K. Yusof, A. Jayalakshmi and M. N. Akmal Khalid (2020). Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset. In: 2020 IEEE 16th International Conference on Control & Automation (ICCA), Singapore, 2020, pp. 803-808, doi: 10.1109.

[Chawla 2002] Chawla Nitesh, Bowyer Kevin, Hall Lawrence, Kegelmeyer W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In: J. Artif. Intell. Res. (JAIR). 16. 321-357. 10.1613/jair.953.

[He 2008] Haibo He, Yang Bai, E. A. Garcia, Shutao Li (2008). 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, 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

[Mqadi 2021] Mqadi Nhlakanipho, Naicker Nalindren, Adeliyi Timothy (2021). A SMOTe based Oversampling Data-Point Approach to Solving the Credit Card Data Imbalance Problem in Financial Fraud Detection. In: International Journal of Computing and Digital Systems. 10. 277 - 286. 10.12785/ijcds/100128.

[Kosolwattana T. et al. 2022] Kosolwattana Tanapol, Liu Chenang, Hu Renjie, Han Shizhong, Chen Hua, Lin Ying (2022). A Self-inspected Adaptive SMOTE Algorithm (SASMOTE) for Highly Imbalanced Data Classification in Healthcare. 10.21203/rs.3.rs-1647776/v1.

[J. -R. Jiang and Y. -T. Chen 2022] J. -R. Jiang and Y. -T. Chen (2022). Industrial Control System Anomaly Detection and Classification Based on Network Traffic. In IEEE Access, vol. 10, pp. 41874-41888, 2022, doi: 10.1109/

[Zhang P. 2022] Zhang Ping, Jia Yiqiao, Shang Youlin (2022). Research and application of XGBoost in imbalanced data. In: International Journal of Distributed Sensor Networks. 18. 155013292211069. 10.1177/15501329221106935.

[Pristyanto Y. et al. 2023] Pristyanto Yoga, Mukarabiman Zulfikar, Nugraha Anggit (2023). Extreme Gradient Boosting Algorithm to Improve Machine Learning Model Performance on Multiclass Imbalanced Dataset. In: JOIV: International Journal on Informatics Visualization. 7. 710-715. 10.30630/joiv.7.3.1102.

[M. Bader-El-Den et al. 2019] M. Bader-El-Den, E. Teitei and T. Perry (2019). Biased Random Forest For Dealing With the Class Imbalance Problem. In: IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 7, pp. 2163-2172, July 2019, doi: 10.1109/TNNLS.2018.2878400.

[J. Justin and K. Taghi 2019] Johnson Justin and Khoshgoftaar Taghi (2019). Survey on deep learning with class imbalance. In: Journal of Big Data. 6. 27. 10.1186/s40537-019-0192-5.

[Z. Fan et al. 2022] Zhang Fan, Petersen Melissa, Johnson Leigh, Hall James and O'Bryant Sid (2022). Hyperparameter Tuning with High Performance Computing Machine Learning for Imbalanced Alzheimer's Disease Data. In: Applied Sciences. 12. 6670. 10.3390/app12136670.

[M. R. Rezaei-Dastjerdehei et al. 2020] M. R. Rezaei-Dastjerdehei, A. Mijani and E. Fatemizadeh (2020). Addressing Imbalance in Multi-Label Classification Using Weighted Cross Entropy Loss Function. In: 2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME), Tehran, Iran, 2020, pp. 333-338, doi: 10.1109/ICBME51989.2020.9319440.

[Rodrigues 2018] Rodrigues V.: Entenda o que é AUC e ROC nos modelos de Machine Learning. Disponível em: https://medium.com/bio-data-blog/entenda-o-que-%C3%A9-auc-e-roc-nos-modelos-de-machine-learning-8191fb4df772. Acesso em: 26 ago 2024.

[De Souza 2019] De Souza Emanuel (2019). Entendendo o que é Matriz de Confusão com Python. Disponível em: https://medium.com/data-hackers/entendendo-o-que-%C3%A9-matriz-de-confus%C3%A3o-com-python-114e683ec509. Acesso em 26 ago 2024.

[Olugbenga 2023] Olugbenga Motunrayo (2023). Balanced Accuracy: When Should You Use It? Disponível em: https://neptune.ai/blog/balanced-accuracy. Acesso em 26 ago. 2024.

[Muralidhar 2021] Muralidhar KSV. What is Stratified Cross-Validation in Machine Learning? Disponível em: https://towardsdatascience.com/what-is-stratified-cross-validation-in-machine-learning-8844f3e7ae8e. Acesso em: 26 ago. 2024.

[Galar M. 2012] Galar Mikel, Fernández Alberto, Barrenechea Edurne, Sola Humberto and Herrera, Francisco (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. In: Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. 42. 463 - 484. 10.1109/TSMCC.2011.2161285.

[A. Dogan and D. Birant 2019] A. Dogan and D. Birant (2019). A Weighted Majority Voting Ensemble Approach for Classification. In: 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, 2019, pp. 1-6, doi: 10.1109/UBMK.2019.8907028.

[Al-Stouhi S. and Reddy C. 2015] Al-Stouhi Samir and Reddy Chandan (2015). Transfer Learning for Class Imbalance Problems with Inadequate Data. Knowledge and Information Systems. 48. 10.1007/s10115-015-0870-3.

[Gui Q. et al. 2024] Gui Q., Zhou H., Guo N. *et al.* (2024). A survey of class-imbalanced semi-supervised learning. In: Mach Learn **113**, 5057–5086 (2024). https://doi.org/10.1007/s10994-023-06344-7.

[Lin E. et al. 2020] Lin E., Chen Q. and Qi X (2020). Deep reinforcement learning for imbalanced classification. In: Appl Intell **50**, 2488–2502 (2020). https://doi.org/10.1007/s10489-020-01637-z.

[C. Nguyen et al. 2021] C. Nguyen Dinh, Ding Ming, Pathirana Pubudu, Seneviratne Aruna, Li Jun and Poor H. Vincent. (2021). In: Federated Learning for Internet of Things: A Comprehensive Survey.

[Bergmann 2023] Bergmann Dave (2023). What is self-supervised learning? Disponível em: https://www.ibm.com/topics/self-supervised-learning. Acesso em 26 ago. 2024.