EXPLORAÇÃO DE TÉCNICAS PARA APRENDIZADO SUPERVISIONADO EM BASES DE DADOS DESBALANCEADAS

Dalisson Almeida[1], Thiago Villela[2], Rafael Azevedo[3], Armando Schiara[4]

[1]Universidade Federal de Minas Gerais, Escola de Engenharia, dalissonfigueiredo@ufmg.br

[2]Universidade Federal de Minas Gerais, Escola de Engenharia, tvillela@ufmg.br

[3]Universidade Federal de Minas Gerais, Escola de Engenharia, rafaelvieiraz@ufmg.br

[4]Universidade Federal de Minas Gerais, Escola de Engenharia, armandos@ufmg.br

Resumo: Aprendizado de Máquina representa uma das áreas de pesquisa mais importantes da atualidade. Todavia, para que a máquina aprenda, faz-se necessários milhares de dados etiquetados, frequentemente possuem um viés para determinada classe do conjunto de dados. Esse problema é conhecido como o desbalanceamento de classes e representa grande problema na aplicação de técnicas de aprendizado. No presente artigo exploramos técnicas de tratamento de bases desbalanceadas aferindo seus resultados através de múltiplas métricas em duas bases acadêmicas.

Palavras-chave: Aprendizado de Máquina, Desbalanceamento, Algoritmo, Inteligência Artificial.

1. Introdução

Em anos recentes, com o aumento do poder de processamento de computadores conjugado com a explosão da disponibilidade de dados concernentes a diversos problemas do mundo real, técnicas de aprendizagem de máquina e mineração de dados se tornaram tema comum em aplicações de interesse na indústria 4.0. Todavia, o aumento de aplicações que confiam em tais técnicas trouxeram a tona desafios novos sobretudo para problemas onde é mais difícil obter grande quantidade de dados relevantes. Um desses desafios é o que ficou conhecido como problema das classes desbalanceadas, ou, problema de base de dados desbalanceada.

Definimos o problema da base dados desbalanceados como a dificuldade em obter um classificador útil a partir de dados que apresentam quantidades diferentes de instâncias para cada classe que se objetiva prever. Com efeito, o treinamento de de algoritmos de aprendizagem a partir de técnicas tradicionais de performance (e.g.

correção de erro a partir do erro quadrático médio, avaliação de modelo exclusivamente através de acurácia)produz classificadores que se ajustam especialmente bem à classe principal (classe com mais instâncias na base de dados), mas que falham em prever classes minoritárias que em geral são a classe de maior interesse no problema.

Isso ocorre pois, o erro padrão oferece a mesma penalidade para cada instância classificada erroneamente e a desproporcionalidade entre classes gera o incentivo implícito para que o algoritmo aprenda a classificar melhor a classe majoritária, o que ocasiona menos classificações errôneas e menor penalidade.

O presente trabalho tem como objetivo explorar e comparar duas técnicas comuns para a abordagem do problema, sendo elas, Random Forest com pesos de erro proporcionais à classe [1], SMOTE(Synthetic Minority Over-Sampling Technique) [2], comparando resultados obtidos através de métricas AUC e acurácia.

2. Fundamentação científica

2.1 Random Forest

Uma mais comuns de Ensemble Learning que consiste em combinar a capacidade de previsão de mais de um algoritmo para alcançar melhores resultados do que seria esperado através de um único algoritmo [9]. Random Forests são modelos baseados em árvores de decisão, o algoritmo base do método de aprendizagem simbolista [7], consistem em combinar diversas árvores com o objetivo de obter melhora significativa na previsão de classes. Diferentemente de técnicas comuns de bagging (bootstrap agre-gation), random forests melhoram a variância reduzindo a correlação entre as árvores [3].

2.2 SMOTE

Pertencente à categoria de pré-processamento de dados, isto é, técnicas aplicadas anteriormente a qualquer tentativa de aprendizagem às bases de dados, SMOTE na geração de amostras sintéticas da classe minoritária, com o intuito de aumentar a quantidade de amostras disponíveis para a classe [5]. A técnica consiste

em identificar pares de vetores de características, fazer a diferença entre esses vetores, em seguida multiplicar essa diferença por um número entre 0 e 1 e criar assim uma observação sintética nos dados, o processo é repetido entre diferentes pares por determinado número de vezes.

2.3 Base de dados

Lista-se a seguir as bases de dados utilizadas no problema. Foram selecionadas bases de dados com diferentes níveis de desbalanceamento com o objetivo de verificar a aplicabilidade e eficiência das técnicas apresentadas em diferentes graus de dificuldade:

- Cars: Datasheet relacionado a carros. O objetivo é prever a condição do carro com base em seis atributos (buying, maint, doors, persons, lug boot, safety). A condição do carro é uma variável que apresenta quatro valores possíveis (unacc, acc, good, vgood).
- Census: Problema de previsão de renda cujo objetivo é classificar as observações entre renda inferior ou igual a cinquenta mil dólares ou renda superior a cinquenta mil dólares por ano. O problema apresenta treze atributos, são eles age, workclass, education, marital-status, education-num, occupation, relationship, sex, race, capital-gain, capital-loss, hours-per-week, native-country.

3. Metodologia

Foram executados k testes sobre as bases de dados seguindo a técnica de divisão em k diferentes conjuntos de teste e treinamento executando-se k-fold-cross-validation com k igual a cinco. Seguindo-se a metodologia leave one out, a cada iteração do programa um subconjunto diferente é deixado de fora do treinamento para validação de resultados do modelo. Cada teste foi executado utilizando três classificadores diferentes, sendo o primeiro uma rede neural padrão com minimização do erro quadrado médio e método de otimização do gradiente descendente, este será considerado o classificador base sem otimização, contra o qual os outros dois modelos serão comparados. O segundo modelo consistiu em

uma rede neural de três camadas com mesma configuração da rede inicial, porém com classes balanceadas a partir de dados gerados pela técnica de SMOTE. Por último, foi implementado o modelo baseado em Random Forest com pesos de erro proporcionais a disponibilidade de classes na base dados cem árvores de decisão.

Como mencionado na seção introdutória serão apresentados os resultados obtidos para a acurácia, precisão, bem como análise das curvas rocs obtidas pelos classificadores. A seguir são apresentados os resultados obtidos para as bases **cars** e **census.**

4. Análise e Interpretação dos Dados

Para a base de dados cars foram obtidos os seguintes resultados para as métricas mencionadas na introdução.

k-fold	Rede sem smote	Rede com Smote	Random Forest
1	0.9917355371900827	0.99634456452	0.6346153846153846
2	0.9028925619834711	0.90909090909	0.8210425937698665
3	0.9992053401144311	0.99952320406	0.8665766052129689
4	0.9639224411951685	0.993006993	0.8306579783852511
5	1	1.0	0.9243483788938335

Tabela 1: AUC

k-fold	Rede sem smote	Rede com Smote	Random Forest
1	0.8439306358381503	0.884393063583815	0.7803468208092486
2	0.8208092485549133	0.7976878612716763	0.7803468208092486
3	0.9884393063583815	0.976878612716763	0.8670520231213873
4	0.8554913294797688	0.9364161849710982	0.7861271676300579
5	1	0.9942196531791907	0.9248554913294798

Tabela 2: Acurácia

Ademais, seguem os resultados obtidos para a base census.

k-fold	Rede sem Smote	Rede com Smote	Random Forest
1	1	1	1
2	0.783783783783784	0.680851	0.585365853658537
3	0.962962962963	0.928571	0.737704918032787
4	0.695652173913043	0.847458	0.590361445783133
5	1	0.981132	0.842105263157895

Tabela 4: AUC

k	Rede sem Smote	Rede com Smote	Random Forest
1	0.8408136192792395	0.31726730046429363	0.8403714348883484
2	0.8458987397744859	0.3468936546539907	0.8452354631881495
3	0.8439089100154764	0.35153659075834626	0.8423612646473579
4	0.8527525978332965	0.3740879946937873	0.8408136192792395
5	0.8606811145510835	0.3845643520566121	0.8513931888544891

Tabela 5: Acurácia

A comparação entre a acurácia obtida para os modelos releva expressiva queda em relação ao modelo que utiliza SMOTE, o mesmo é verdade para a precisão ao longo de todos os k-folds, também a pequena queda em relação à AUC. Destaca-se que a queda de acurácia do modelo treinado com SMOTE acontece devido ao fato do aumento substancial de falsos positivos, ao passo que o modelo é capaz de classificar corretamente todos os verdadeiros falhando em classificar corretamente a classe de interesse em apenas duas instâncias ao longo de todos os dez testes.

5. Conclusão

O problema de classes desbalanceadas se mostra um problema desafiador para o aprendizado de máquina, mesmo para as técnicas utilizadas. Houveram resultados variados em virtude dos diferentes níveis de complexidade de problema como em virtude dos diferentes níveis de desbalanceamento enfrentados. Todavia, para a todos dos casos o uso da técnica de SMOTE foi consistente em aumentar o erro do Tipo I e diminuir consideravelmente erros do Tipo II. Embora, os benefícios em trocar diferentes tipos de erro devam ser considerados no contexto do problema e regras de negócios [10] e uma discussão detalhada sobre o assunto da aplicação em negócio fuja ao escopo deste trabalho, foi possível demonstrar os efeitos da aplicação das referidas técnicas de correção de desbalanceamento sobre um problema real.

6. Referências

[1] PASAPITCH, Chujai; KITTIPONG, Chomboon; PONGSAKORN, Teeraras-samee. Ensemble learning for imbalanced data classification problem. 2015.

[2] NITESH, V Chawla; KEVIN, W Bowyer; HALL, Lawrence O; KEGELMEYER, W Philip. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.

[3] BREIMAN, Leo. Random forests. Machine Learning, volume 45.

[4] DE CASTRO, Cristiano L; BRAGA, Antônio P. Aprendizado supervisionado com conjuntos de dados desbalanceados. Rev. Controle Autom, 22(5):441–466, 2011.

[5] NITESH, V Chawla; BOWYER, W Kevin; HALL, Lawrence O; KEGELMEYER, Philip Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.

[6] PASAPITCH, Chujai; CHOMBOON, Kittipong; TEERARAS-SAMEE, Pongsakorn; KERDPRASOP, Nittaya; KERDPRASOP, Kittisak. Ensemble learning for imbalanced data classification problem. 2015.

[7] DOMINGOS, Pedro. The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Basic Books, Inc., New York, NY, USA, 2018.

[8] HODGSON, Robert T. How expert are “expert” wine judges? Journal of Wine Economics, 4(2):233–241, 2009.

[9] POLIKAR, Robi. Ensemble learning. 2009.

[10] PROVOST, Foster; FAWCETT, Tom. Data Science Para Negócios. ELSEVIER/ALTA BOOKS.