Exploração de Técnicas de Aprendizado em Bases Desbalanceadas

By Dalisson Figueiredo et Al.

Introdução

. Aprendizagem de máquina requer uma grande quantidade de dados etiquetados.

. Um problema comum e de grande interesse acadêmico e prático é a classificação, isto é, a partir das características de uma amostra identificar a qual classe ela pertence

Problema

- No mundo real, é comum que uma das classes presente na base dados seja muito mais abundante que outra, eg., uma base referente a transações de cartão de crédito possui dezenas se não milhares de transações legítimas para cada transação fraudulenta.

- Em virtude disso o modelo resultante é muito melhor em classificar a classe dominante em detrimento às outras classes, seguindo nosso exemplo, se a base possui 999 transações legítimas e apenas 1 transação fraudulenta basta ao modelo classificar todas como legítimas para atingir 99.9% de acerto e ainda assim ser incapaz de fazer a tarefa para a qual foi concebido.

Técnicas de Solução

.Entre as principais técnicas de solução destacam-se aquelas que alteram a amostragem de dados (reduzindo as amostras da classe majoritária ou gerando amostras das classes minoritárias) e as que alteram a penalização do classificado face ao tipo de erro que este comete.

.A seguir exploramos uma técnica para cada um dos dois métodos o SMOTE, que gera amostras sintéticas da classe minoritária a partir do dataset e uma Random Forest com pesos inversamente proporcionais à abundância da classe.

.Para aferir resultados utilizamos dois datasets acadêmicos e a técnica de validação cruzada, que consiste em dividir os dados em k grupos distintos, em seguida utilizar k-1 grupos para treino avaliando-se o resultado no grupo restante em seguida repetindo o processo k vezes, alternando o grupo que é deixado de fora do treino. Também efetuamos o treino sem correção para efeito de comparação.

Dataset 1 - Resultados

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	0.866576605212968 9
4	0.9639224411951685	0.993006993	0.830657978385251 1
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

Dataset 2 - Resultados

k-fold	Rede sem Smote	Rede com Smote	Random Forest
1	1	1	1
2	0.783783783783784	0.680851	0.585365853658537
3	0.962962962962963	0.928571	0.737704918032787
4	0.695652173913043	0.847458	0.590361445783133
5	1	0.981132	0.842105263157895

Tabela 3 - AUC

k	Rede sem Smote	Rede com Smote	Random Forest
1	0.8408136192792395	0.3172673004642936 3	0.8403714348883484
2	0.8458987397744859	0.3468936546539907	0.8452354631881495
3	0.8439089100154764	0.3515365907583462 6	0.8423612646473579
4	0.8527525978332965	0.3740879946937873	0.8408136192792395
5	0.8606811145510835	0.3845643520566121	0.8513931888544891

Tabela 4 - Acurácia

Avaliando os Resultados, notamos grande diferença entre os três modelos treinados e

Conclusão

Avaliando os Resultados, notamos grande diferença entre os três modelos treinados.

.Apesar de as técnicas aumentarem a taxa de exemplos da classe majoritária classificados erroneamente classificados na classe minoritária o vantagem de classificarem corretamente a classe minoritária com mais frequência, pode, em virtude das regras inerentes ao contexto no qual o problema se insere, sobrepujar o efeito deletério do aumento de erros.