“A Comprehensive Study on Ensemble-Based Imbalanced Data Classification Methods for Bankruptcy Data”

Aprendizagem de Máquina 2020.1

Eládia Cristina Corrêa (ecmac)
Gabriela Leal Magalhães (glm)

Centro de Informática U·F·P·E
Universidade Federal de Pernambuco
CIn.ufpe.br

Roteiro

- Introdução
- Abordagens para dados desbalanceados
  - SMOTEBAGGING
  - UNDERBAGGING
  - SMOTEBOOST
  - RUSBOOST
- Metodologia
- Resultados
- Conclusão
- Referências

Introdução

- Dados reais são comumente desbalanceados
- Classificadores tendem a ficar enviesados para a classe majoritária
- Predição de falência de empresas é importante para a indústria financeira
- Métodos de aprendizagem de máquina são importantes nessa predição, principalmente por mitigar a falta de dados reais sobre falência

Abordagens para dados desbalanceados

- Balancear a proporção dos dados
- Ensemble
  - Bagging: classificadores em paralelo
  - Boosting: classificadores em sequência
- Oversampling e undersampling
  - SMOTE: Synthetic Minority Oversampling Technique (técnica de sobreamostragem sintética da minoria)
  - RUS: Random Undersampling (subamostragem aleatória)
- Combinar as abordagens

SMOTEBAGGING

- SMOTE + Bagging
- Oversampling com SMOTE a cada iteração do algoritmo de bagging
- Gera novos dados sintéticos
- Garantir a diversidade dos dados
- Binários ou multiclasse

UNDERBAGGING

- RUS + bagging
- Aplicação do RUS a cada iteração do bagging
- Remove instâncias
- Binários ou multiclasse
- Controle da diversidade de dados

SMOTEBOOST

- SMOTE + boosting (AdaBoost)
- Melhorar a acurácia do modelo de forma global
- Foco na classe minoritária, evitando danos à acurácia para a majoritária
- Boosting reduz variância e viés
  - mas classe majoritária inevitavelmente prevalece
- Gera novos dados sintéticos para a classe minoritária a cada iteração, ajustando os pesos até convergirem, garantindo assim balanceamento e acurácia.

RUSBOOST

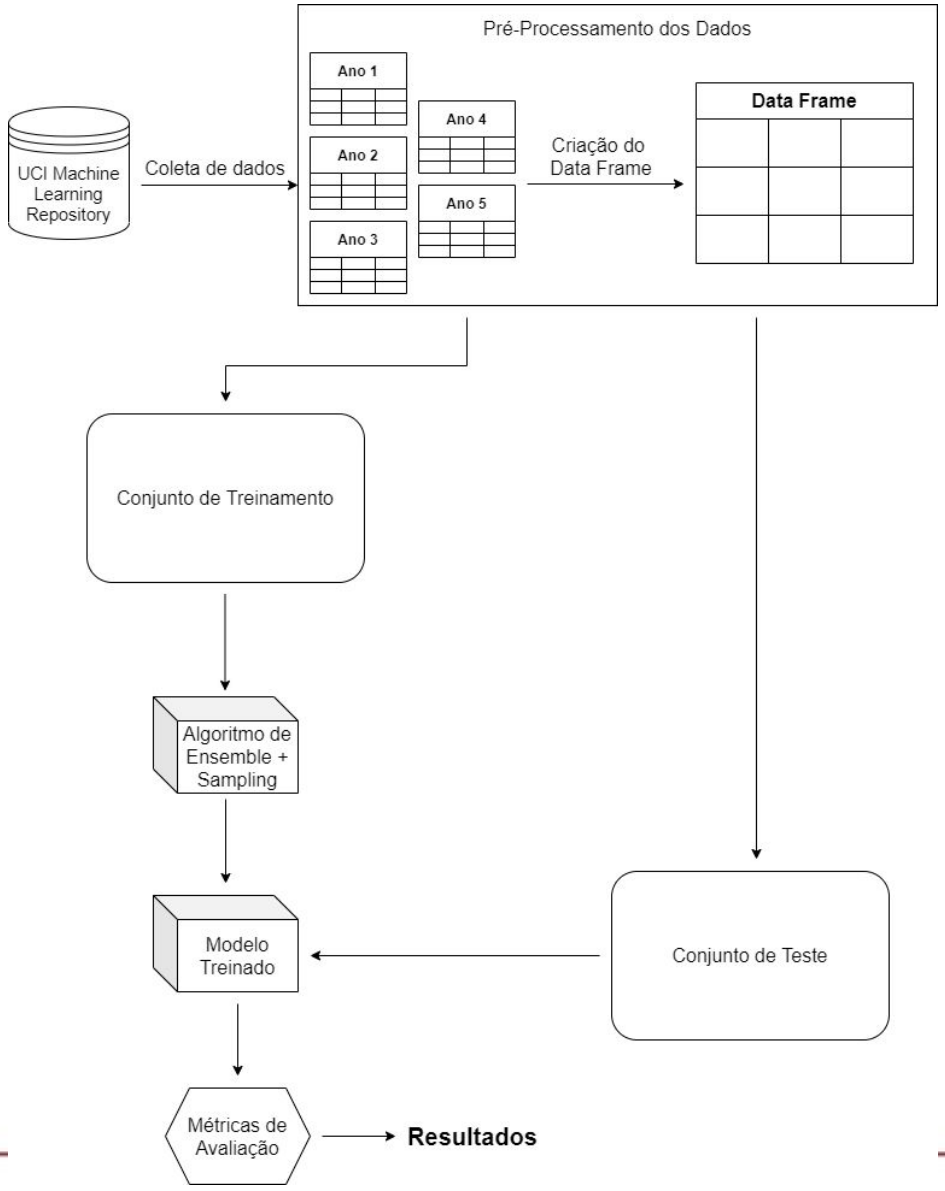
- RUS + boosting (AdaBoost)
- Também propõe melhorar acurácia global
- Visa atenuar a complexidade de tempo do SMOTE
  - Tempo de construção do modelo diminui, pois a base torna-se mais concisa
- Perda de informações do RUS não causa um impacto muito negativo, por também utilizar boosting
- A cada iteração do AdaBoost, dados da classe majoritária são removidos e pesos ajustados

Metodologia

- Python no Google Colab
  - Pandas
  - imbalanced-learn
  - scikit-learn
  - SMOTEBOOST do DIAL Lab

- Dados sobre falência de empresas polonesas
  - remoção de dados faltantes
  - normalização no intervalo (-1, 1)
  - remoção de linhas duplicadas
  - 19449 registros: 19027 não faliram, e 422 faliram
  - Taxa de desbalanceamento: 0.022179008777

Metodologia



Metodologia: Experimentos

- Hold-out: treino 70%, teste 30%
- SMOTEBAGGING
  - BalancedBaggingClassifier
  - hiperparâmetro “sampler”: SMOTE
- UNDERBAGGING
  - BalancedBaggingClassifier
  - hiperparâmetro “sampler”: RUS
- SMOTEBOOST
  - Implementação do DIAL Lab
- RUSBOOST
  - RUSBoostClassifier

Metodologia: Avaliação

- Métricas
  - Recall
  - Precisão
  - Acurácia
  - F-measure
  - ROC
- Médias: função “weighted”
  - Conjunto de teste permaneceu desbalanceado

Resultados: Métricas

Algoritmo	*Recall*	Acurácia	Precisão	*F-Measure*
SMOTEBAGGING	0.973	0.973	0.968	0.970
UNDERBAGGING	0.865	0.865	0.973	0.909
SMOTEBOOST	0.972	0.972	0.968	0.970
RUSBOOST	0.829	0.829	0.971	0.888

Tabela 2. Resultados das métricas de avaliação.

- SMOTE teve o melhor resultado geral em ambos métodos
- RUS, apesar de ter uma precisão levemente maior, ficou bem abaixo de SMOTE nas outras métricas
  - Baixo False Positive, porém False Negative mais alto
- Boosting e bagging apresentaram resultados similares.
  - Com RUS, bagging foi um pouco superior

Resultados: ROC

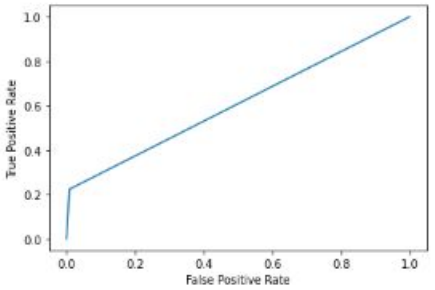


Figura 2. Curva ROC SMOTEBAGGING.

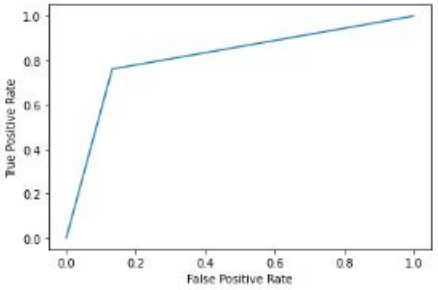


Figura 3. Curva ROC UNDERBAGGING.

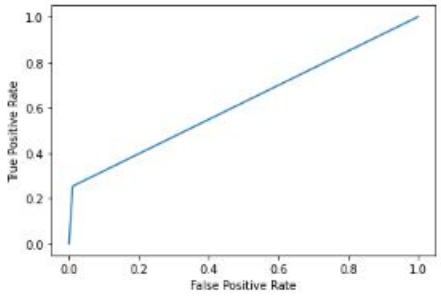


Figura 4. Curva ROC SMOTEBOOST.

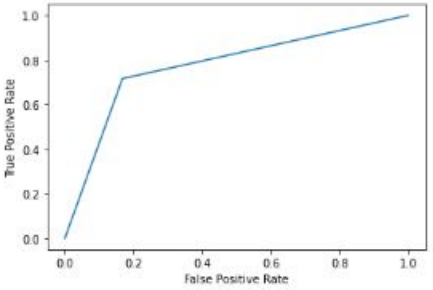


Figura 5. Curva ROC RUSBOOST.

- Novamente, a diferença é maior comparando técnicas de amostragem do que os ensembles
- Com RUS, mais instâncias positivas foram classificadas corretamente, porém mais instâncias negativas foram classificadas como positivas

Centro de Informática UFPE

Conclusão

- Dados desbalanceados são uma tarefa desafiadora
- Contudo conseguimos obter os resultados demonstrados no artigo
- Comparando SMOTEBAGGING, UNDERBAGGING, SMOTEBOOST e RUSBOOST, evidenciamos que SMOTEBAGGING e SMOTEBOOST possuíram os melhores resultados
- Logo, SMOTE superou RUS

Referências

[1] K. UlagaPriya and S. Pushpa, "A Comprehensive Study on Ensemble-Based Imbalanced Data Classification Methods for Bankruptcy Data," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 800-804, doi: 10.1109/ICICT50816.2021.9358744.

[2] UCI Machine Learning Repository: Polish companies bankruptcy data Data Set. Disponível em https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data. Acesso em: 27 de abril, 2021.

[3] Python-based implementations of algorithms for learning on imbalanced data. Disponível em https://github.com/dialnd/imbalanced-algorithms. Acesso em: 27 de abril, 2021.

[4] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse and A. Napolitano, "RUSBoost: Improving classification performance when training data is skewed," 2008 19th International Conference on Pattern Recognition, 2008, pp. 1-4, doi: 10.1109/ICPR.2008.4761297.

[5] A. Amin et al., "Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study," in IEEE Access, vol. 4, pp. 7940-7957, 2016, doi: 10.1109/ACCESS.2016.2619719.

[6] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 4, pp. 463-484, July 2012, doi: 10.1109/TSMCC.2011.2161285.

[7] S. Ahmed, A. Mahbub, F. Rayhan, R. Jani, S. Shatabda and D. M. Farid, "Hybrid Methods for Class Imbalance Learning Employing Bagging with Sampling Techniques," 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), 2017, pp. 1-5, doi: 10.1109/CSITSS.2017.8447799.

[8] Chawla N.V., Lazarevic A., Hall L.O., Bowyer K.W, “SMOTEBoost: Improving Prediction of the Minority Class in Boosting”, n: Lavrač N., Gamberger D., Todorovski L., Blockeel H. (eds) Knowledge Discovery in Databases: PKDD 2003. PKDD 2003. Lecture Notes in Computer Science, vol 2838. Springer, Berlin, Heidelberg, 2003, doi: 10.1007/978-3-540-39804-2_12.

[9] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," 2009 IEEE Symposium on Computational Intelligence and Data Mining, 2009, pp. 324-331, doi: 10.1109/CIDM.2009.4938667.

fim