TRATAMENTO DE DADOS DESBALANCEADOS

Leandro Almeida

Ricardo Prudêncio

Introdução

- É comum nos depararmos com problemas de classificação com dados desbalanceados
  - i.e., Presença de classes majoritárias com frequência muito maior que as outras classes minoritárias
- Desbalanceamento de classes pode ser prejudicial dependendo do problema e algoritmo

Desbalanceamento de Dados

- Consequência:
  - Maior tendência para a responder bem para as classes majoritárias em detrimento das minoritárias
  - Entretanto, em muitos casos, o que importa é ter um bom desempenho para as classes minoritárias!!!
    - Ver exemplos no próximo slide

Desbalanceamento de Dados

- Exemplo: Detecção de Fraude
  - Menos de 1% das transações de cartão de crédito são fraudes
  - Em um conjunto de exemplos relacionados a transações, teremos:
    - 99% dos exemplos para a classe negativa (não-fraude)
    - 1% dos exemplos para a classe positiva (fraude)

Desbalanceamento de Dados

- Exemplo: Detecção de Fraude
  - Classificadores terão uma tendência a dar respostas negativas para transações com fraude
    - i.e. Alto número de falsos negativos
  - Problema: o custo de um falso negativo é muito maior que o custo de um falso positivo
    - Falso negativo: fraude que não foi detectada em tempo
      - i.e., prejuízo a operadora de cartão
    - Falso positivo: transação normal bloqueada
      - i.e., aborrecimento para o usuário do cartão

Desbalanceamento de Dados

- Exemplo: Diagnóstico Médico
  - Pacientes doentes são em geral menos comuns que pacientes saudáveis
  - No diagnóstico médico, novamente a classe positiva (dos pacientes doentes) tem uma frequência muito menor que a classe negativa (pacientes saudáveis)

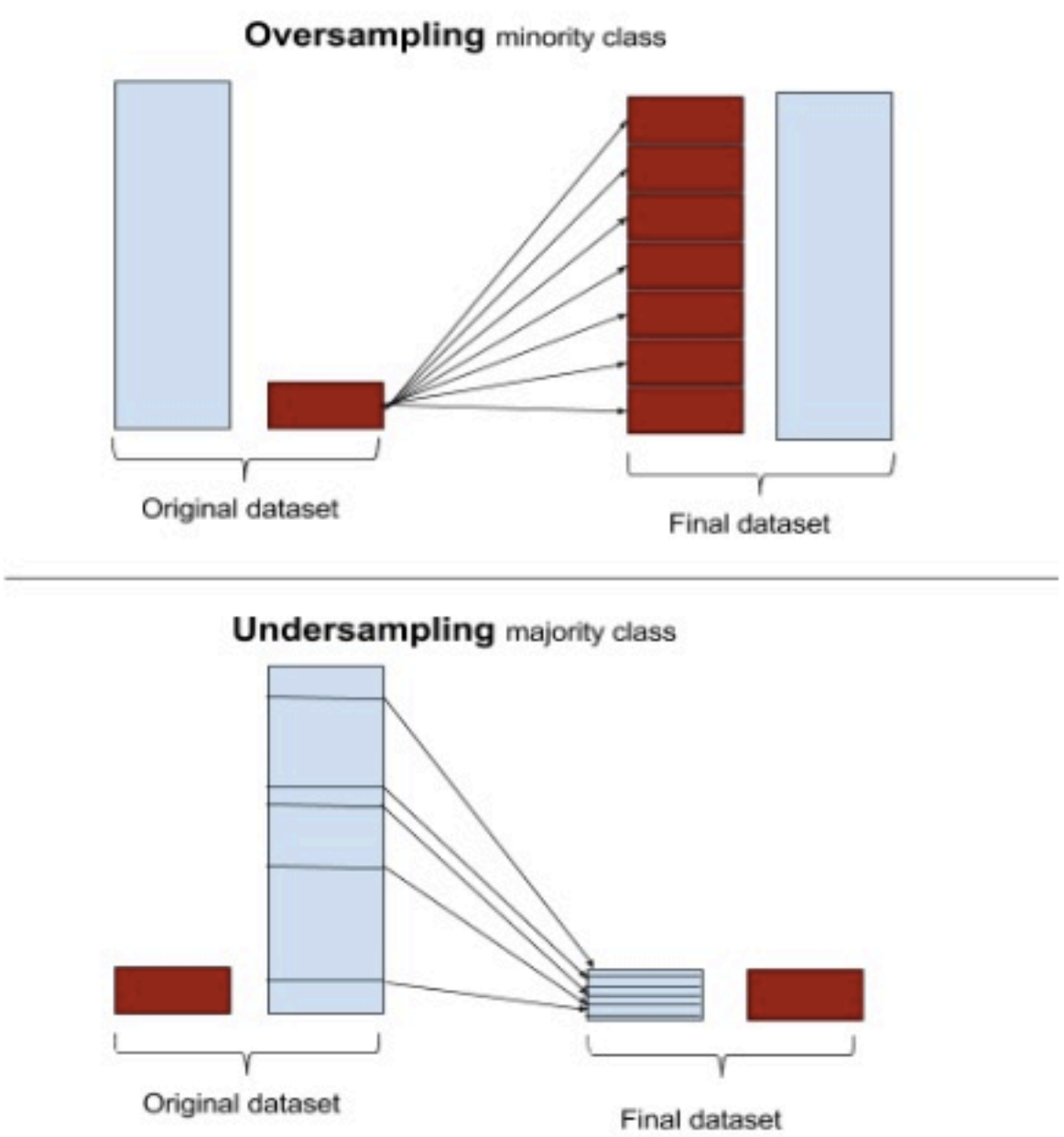
Desbalanceamento de Dados

- Exemplo: Diagnóstico Médico
  - Classificadores terão uma tendência a classificar doentes reais como supostamente saudáveis
    - Novamente, alto número de falsos positivos
  - Consequência :
    - Diagnóstico tardio e dano para o paciente

Abor
de D

- Resar
  - Rear
    conj
  - Und
    - Re
    - Pc
  - Over
    - Re
    - Se

erar



Abordagens para Tratamento de Dados Desbalanceados

- SMOTE (Chawla et al., 2002)
  - Oversampling usando exemplos sintéticos da classe minoritária
  - Exemplos sintéticos extraídos ao longo dos segmentos que unem vizinhos mais próximos da classe minoritária
- One-side-selection (Kubat & Matwin, 1997)
  - Undersampling de exemplos redundantes da classe majoritária

Abordagens para Tratamento de Dados Desbalanceados

- Wilson's editing (Barandela et al., 2004)
  - Usa kNN para classificar instâncias da classe majoritária e exclui as classificadas erroneamente
- Cluster-based oversampling (Jo & Japkowicz, 2004)
  - Realizam cluster dos exemplos e replicam exemplos dos clusters mais desbalanceados

Abordagens para Tratamento de Dados Desbalanceados

- Comparação entre métodos
  - Undersampling aleatório tem se mostrado superior a oversampling (Chawla et al. 2002) (Hulse et al. 2007)
  - Entretanto o melhor método de sampling depende do algoritmo sendo utilizado e da métrica de avaliação (Hulse et al. 2007)

Considerações finais

- Em geral, conjuntos de dados tem algum grau de desbalanceamento das classes
- Desbalanceamento pode ser um problema difícil dependendo da complexidade das classes
- Conjuntos desbalanceados podem ser lidados com técnicas de amostragem ou introduzindo custos diretamente
  - Não há “o melhor” método

Referências

- Barandela, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for learning in class imbalance problems, Pattern Recognition 36(3) (2003) 849-851.
- N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In Proceedings of the Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases, pages 107-119, Dubrovnik, Croatia, 2003.
- M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: One sided selection. In Proceedings of the Fourteenth International Conference on Machine Learning, pages 179-186, Nashville, Tennesse, 1997. Morgan Kaufmann
- Taeho Jo and N. Japkowicz (2004), Class Imbalances versus Small Disjuncts, Sigkdd Explorations. Volume 6, Issue 1 - Page 40-49
- Gary M. Weiss, Kate McCarthy, and Bibi Zabar (2007), Cost-Sensitive Learning vs. Sampling: Which is Best for. Handling Unbalanced Classes with Unequal Error Costs?. Proceedings of the 2007 International Conference on Data Mining, DMIN.
- Jason Van Hulse, Taghi M. Khoshgoftaar, Amri Napolitano: Experimental perspectives on learning from imbalanced data. ICML 2007: 935-942
- Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas (2006), Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, Vol.30, 2006

Trabalhando em python

- from collections import Counter
- from sklearn.datasets import make_classification
- from imblearn.under_sampling import RandomUnderSampler
- from imblearn.over_sampling import RandomOverSampler

- X, y = make_classification(n_classes=2, class_sep=2,weights=[0.1, 0.9], n_informative=3, n_redundant=1, flip_y=0,n_features=20, n_clusters_per_class=1, n_samples=1000, random_state=10)
- print('Original dataset shape {}'.format(Counter(y)))

- rus = RandomUnderSampler(random_state=42)
- X_res, y_res = rus.fit_sample(X, y)
- print('Resampled dataset shape {}'.format(Counter(y_res)))

- ros = RandomOverSampler(random_state=42)
- X_res, y_res = ros.fit_sample(X, y)
- print('Resampled dataset shape {}'.format(Counter(y_res)))