Tópicos Especiais em Reconhecimento de Padrões [2COP329]

Mestrado em Ciência da Computação

Sylvio Barbon Jr
barbon@uel.br

DC
Departamento de Computação

Tema

Aula 5
Dados Desbalanceados

Introdução

- Lidar com distribuição desbalanceada (imbalanced) nas amostras é um problema para o reconhecimento de um determinado padrão ou classe;
- Isso ocorre quando o número de exemplos que representam uma classe é muito menor que outra;
- Tal problema está presente em problemas do mundo real: casos raros ou complexos, mas necessários para descrição completa de um problema;
- Esta característica ocorre em problemas de uma ou mais classes.
- A maioria dos sistemas de aprendizado de máquina não estão preparados para tratar classes desbalanceadas;

Introdução

- Os algoritmos de classificação tem o seu poder de generalização prejudicado quando as bases estão desbalanceadas.
- As soluções para o problema de **class imbalance** são categorizadas em quatro grupos:
  - Re-amostragem (resampling) para balanceamento de dataset;
  - Modificação ou ajuste nos algoritmos de aprendizado;
  - Métricas de avaliação em domínios desbalanceados;
  - Relação entre classes desbalanceadas e características complexas.

Introdução

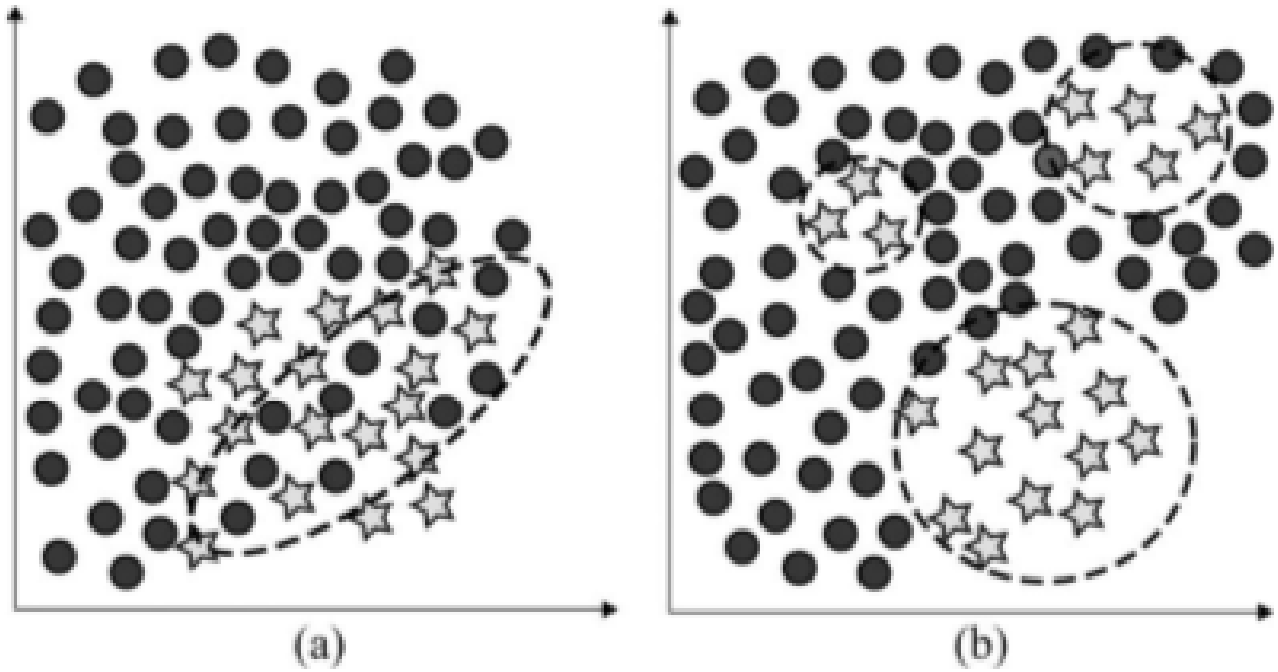


Fig. 1. Example of difficulties in imbalanced data-sets. (a) Class overlapping. (b) Small disjuncts.

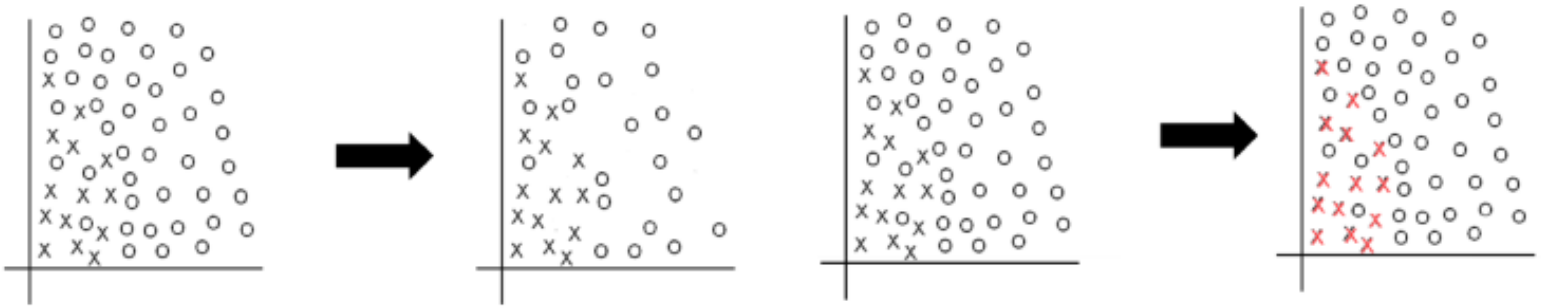
Problemas

- **Small sample size**: Um número baixo de amostras em uma classe minoritária diminui a sua representatividade, generalização e avaliação do desempenho do classificador;
- **Overlapping e class separability**: Quando isso ocorre regras discriminativas são de difícil indução. Regras mais gerais acarretam erros na classificação;
- **Small disjuncts**: Ocorre quando a classe minoritária é composta por diversos **subconcepts**. Problemas do mundo real tem esta características, o problema está na formação dos subconcepts, que geralmente são formados por poucas amostras.

Abordagens para solução

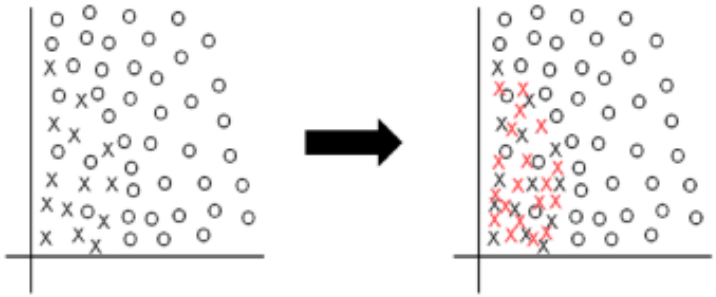
- **Sampling** (métodos de amostragem): Balanceamento alcançado por meio de re-amostragens e subamostragens dos datasets.
- **Ensemble** (conjunto de métodos): Explora de maneira supervisionada a classe majoritária, buscando equalizar a quantidade de amostras.
- **Cost based** (métodos de custo): Considera os custos dos erros de classificação na distribuição do treinamento.
- **Outros métodos**: buscam ajustar o problema de desbalanceamento considerando amostras irrelevantes para o treinamento.

Sampling (métodos de amostragem)



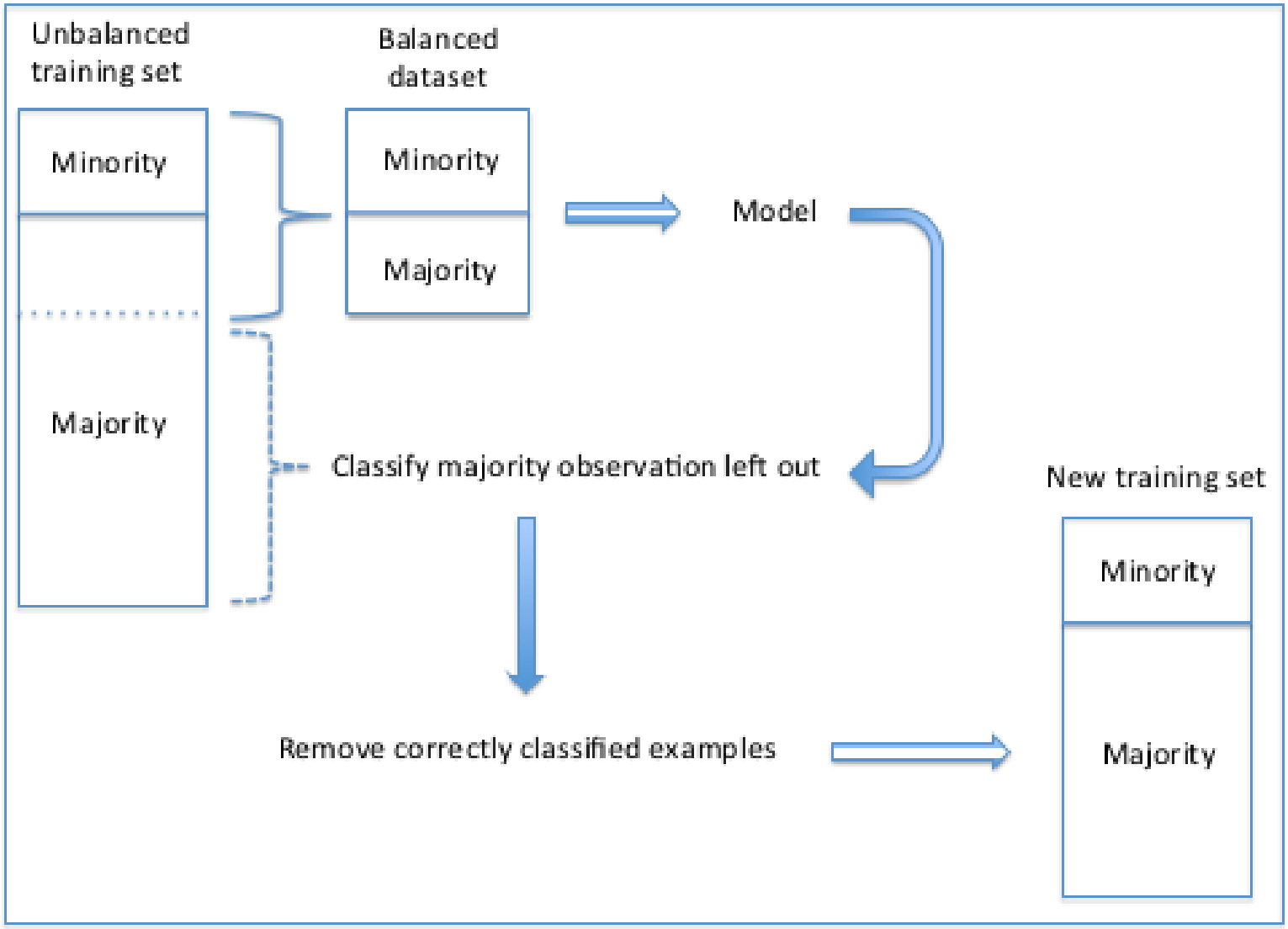
Undersampling

Oversampling

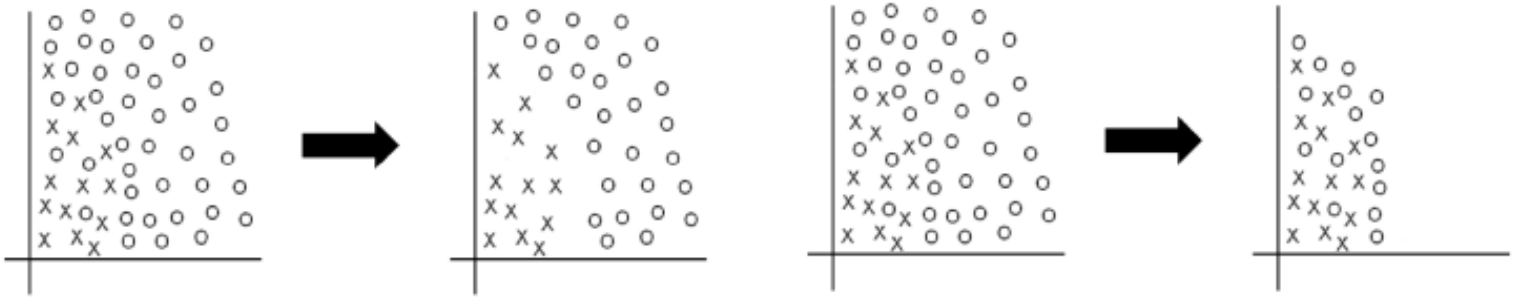


Smote[1]

Ensemble (conjunto de métodos)



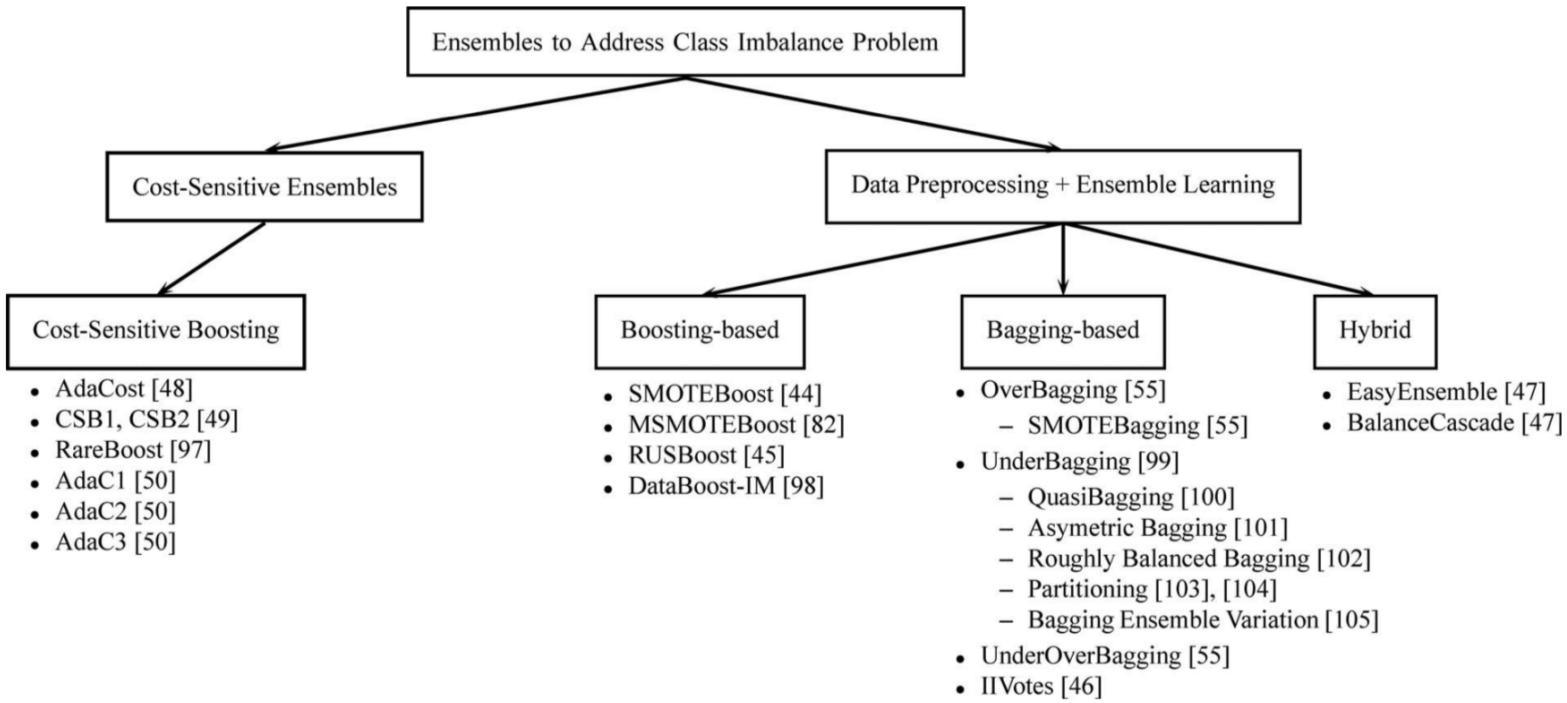
Outros métodos



Tomek link [7]

Condensed Nearest Neighbor [3]

Outros métodos



Técnicas de Pré-processamento

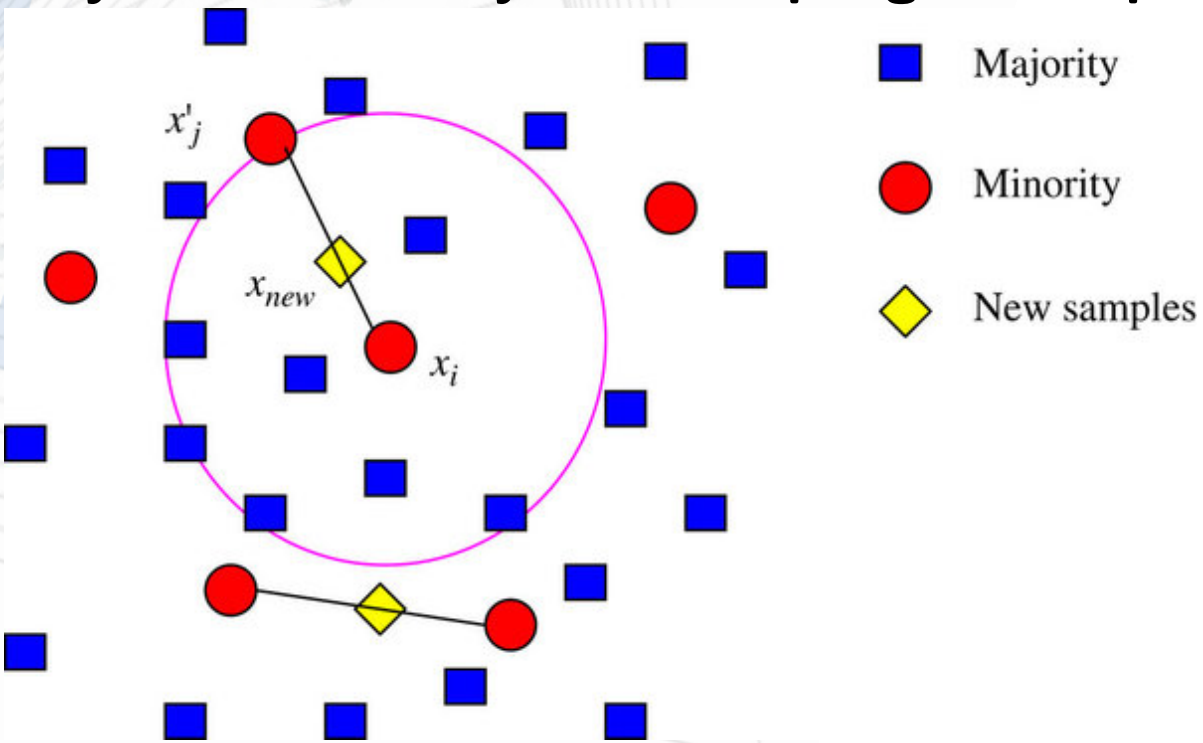
- **Random Undersampling**: É um método não heurístico que tem como objetivo a distribuição balanceada de classes eliminando aleatoriamente amostras da classe majoritária. O principal problema é o descarte de dados potencialmente úteis para o processo de indução.
- **Random Oversampling**: É um método semelhante ao Random Undersampling, porém são replicadas instâncias das classes minoritárias. No entanto este método aumenta a "vizinhança" podendo ocorrer o "**overfitting**" devido às cópias idênticas.

Técnicas de Pré-processamento

- **SMOTE - Sythetic Minority Oversampling technique**: É uma técnica de **oversampling**, criando novas amostras baseadas na interpolação de instâncias da classes minoritárias. Baseado em kNN (k nearest neighbors), aleatóriamente selecionam amostras das classes minoritárias e geram as novas.
  O problema de overfitting é evitado, no entanto os limites da classe minoritária pode ser sinteticamente modificada.

Técnicas de Pré-processamento

SMOTE - Sythetic Minority Oversampling technique



Técnicas de Pré-processamento

- **MSMOTE - Modified Sythetic Minority Oversampling technique**: É uma versão modificada do SMOTE, cuja instâncias das classes minoritárias são dividas entre *safe*, *border* e *latent noise*.
  Quando MSMOTE gera as novas instâncias considerando o kNN entre as classes previamente rotuladas, evitando a criação de possíveis instâncias que causem ruído.

Técnicas de Pré-processamento

- **SPIDER - Selective preprocessing of imbalanced data**: Esta técnica combina o oversampling local da classe minoritária com filtragem baseada na classe majoritária. Esta técnica é baseada em duas fases:
  - Identificação;
  - Pré-processamento;
- A identificação verifica instâncias que são potenciais ruídos;
- O pré-processamento depende de parâmetros (weak, relabel ou strong) e uma estrutura de decisão modificação a criação das amostras sintéticas.

Técnicas de Pré-processamento

- **SMOTEBoost**:
- Combina o algoritmo SMOTE com procedimento padrão de Boosting.
- SMOTE atuando sobre a classe minoritária;
- Boosting atribui pesos iguais para todos os exemplos classificados incorretamente. Avaliando FP e FN, o objetivo é reduzir o viés aumentando os pesos das classes minoritárias melhorando o poder de decisão sobre a classe minoritária.

Lista de Referências

1. N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: synthetic minority over-sampling technique. Arxiv preprint arXiv:1106.1813, 2011.
2. C. Elkan. The foundations of cost-sensitive learning. In International Joint Conference on Artificial Intelligence, volume 17, pages 973–978. Citeseer, 2001.
3. P. E. Hart. The condensed nearest neighbor rule. IEEE Transactions on Information Theory, 1968 . Japkowicz and S. Stephen. The class imbalance problem: A systematic study. Intelligent data analysis, 6(5):429–449, 2002.
4. B. D. Ripley, "Pattern Recognition and Neural Networks", Cambridge University Press, 1996.
5. C.X. Ling and V.S. Sheng. Cost-sensitive learning and the class imbalance problem. Encyclopedia of Machine Learning, 2008.