Universidade Federal de Pernambuco

Graduação em Ciência da Computação

Centro de Informática

2015.2

Análise comparativa de técnicas de amostragem aplicada ao problema de dados desbalanceados

Proposta de Trabalho de Graduação

Aluno: Romero Fernando Almeida Barata de Morais (rfabm@cin.ufpe.br)

Orientador: Germano Crispim Vasconcelos (gcv@cin.ufpe.br)

Contexto

Nos últimos 10 anos, foi visível o crescente interesse da comunidade acadêmica em torno da questão de como aprender a partir de dados desbalanceados [1]. O desbalanceamento dos dados pode ser de natureza tanto intrínseca como extrínseca. No primeiro caso, o desbalanceamento provém da própria natureza dos dados, um exemplo comum são dados de exames de câncer, é de se esperar que uma minoria dos pacientes consultados de fato sejam diagnosticados com a doença. No caso do desbalanceamento extrínseco, embora os dados venham de uma fonte balanceada, por algum motivo a amostragem pode ter gerado um desbalanceamento.

As principais técnicas para lidar com tal problema incluem (mas não são limitadas a) algoritmos de amostragem e algoritmos sensíveis ao custo [2], [3]. Algoritmos de amostragem atuam diretamente nos dados, e em geral, super amostram a classe minoritária ou sub amostram a classe majoritária. Algoritmos sensíveis ao custo, por sua vez, penalizam os classificadores de maneira mais severa quando um exemplo da classe minoritária é classificado erroneamente.

Além do mais, métricas para avaliar e comparar o desempenho de diferentes classificadores que aprendem a partir de dados desbalanceados ainda não estão bem estabelecidas [1]. De acordo com [1], curvas ROC, curvas PR e curvas de custo devem ser empregadas simultaneamente para uma melhor comparação entre diferentes técnicas e classificadores.

Objetivo

O presente trabalho tem por objetivo comparar e avaliar as mais relevantes técnicas de amostragem de dados. Dentre elas podemos citar: random oversampling, random undersampling, SMOTE [2], Borderline-SMOTE [5] e o ADASYN [6].

Inicialmente, as técnicas serão aplicadas a dados sintéticos bidimensionais para que o comportamento de cada técnica possa ser visualizado e entendido. Em seguida, uma ou mais aplicações reais serão submetidas a todas as técnicas para identificarmos o quão relevante o balanceamento dos dados é para o processo de aprendizagem. Finalmente, todos os resultados serão analisados e interpretados seguindo as recomendações de [1].

Um estudo similar já foi proposto por [4] mas as técnicas comparadas não incluem as mais relevantes atualmente, e além disso a visualização e interpretação das técnicas sob um cenário bidimensional artificial não é feito. É válido também ressaltar que a métrica principal de comparação foi a AUC, não incluindo técnicas como curvas PR e curvas de custo.

Cronograma

Atividades	Agosto				Setembro				Outubro				Novembro				Dezembro				Janeiro			
Levantamento bibliográfico																								
Geração dos dados artificiais																								
Implementação das técnicas																								
Aplicação das técnicas aos dados artificiais																								
Aplicação das técnicas aos dados reais																								
Análise dos resultados																								
Escrita do relatório																								
Elaboração da apresentação																								

Possíveis Avaliadores

1º Tsang Ing Ren (tir@cin.ufpe.br)

2º Patricia Cabral de Azevedo Restelli Tedesco (pcart@cin.ufpe.br)

Referências

[1] H. He and E. A. Garcia, “Learning from Imbalanced Data,” *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009. doi: 10.1109/TKDE.2008.239

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artificial Intell. Research*, vol. 16, no. 1, pp. 321–357, Jan. 2002.

[3] Z. H. Zhou, and X. Y. Liu, “Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem,” *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 63–77, Jan. 2006. doi: 10.1109/TKDE.2006.17

[4] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data,” *SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, June 2004. doi: 10.1145/1007730.1007735

[5] H. Han, W. Y. Wang, and B. H. Mao, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning,” in *Proc. 2005 Int. Conf. Advances Intelligent Computing (ICIC’05)*, pp. 878–887. doi: 10.1007/11538059_91

[6] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,” in *IEEE Int. Joint Conf. Neural Networks (IJCCN 2008)*, June, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969

Assinaturas

Recife, Pernambuco. Brasil

30 de Setembro de 2015

_____

Germano Crispim Vasconcelos

(orientador)

_____

Romero Fernando Almeida Barata de Morais

(proponente)