Aprendizado por Agrupamento

Francisco Romes da Silva Filho[1], **Marcos Vinicius de Lima Venancio**[1]
Igor Alan Albuquerque de Sousa[1]

[1]Universidade Federal do Ceará (UFC)
Campus Quixadá

{romesfilho_cc, marcos.vincius, igorigor}@alu.ufc.br

Abstract. *The Ensemble Learning has been used in several applications in our daily lives and presents a promising future in Artificial Intelligence. Today, it is already possible to reap its rewards with predictions, from lithium batteries to the prediction of business failure. Such elements affect our daily lives. Therefore, the present work presents the theme and its applications.*

Resumo. *O Aprendizado por Agrupamento vem sendo usado em várias aplicações do nosso cotidiano e apresenta um futuro promissor na Inteligência Artificial. Atualmente, já é possível colher seus frutos com predições, desde a baterias de lítio à previsão de falência de empresas. Tais elementos afetam o nosso dia a dia. Portanto, o presente trabalho apresenta o tema e suas aplicações.*

1. Introdução

Aprendizado por Agrupamento (*Ensemble Learning*) foi uma ideia inicialmente apresentada em [Nilsson 1965] e com aplicação inicial em Aprendizagem Supervisionada, tarefa de aprendizado de máquina que consiste em aprender uma função que mapeia uma entrada para uma saída com base em pares de entrada e saída de exemplo, como descrito em [Russell and Norvig 2009].

O conceito é apresentado em [Sammut and Webb 2011] como sendo procedimentos empregados para treinar várias máquinas de aprendizagem e combinar seus resultados, tratando-os como um "grupo" de tomadores de decisão. O princípio é que a decisão do grupo, com previsões individuais combinadas de forma adequada, deve ter melhor acurácia geral, em média, do que qualquer membro individual do grupo. Numerosos estudos empíricos e teóricos têm demonstrado que os modelos de *Ensemble* muitas vezes atingem maior acurácia do que os modelos individuais [Lv et al. 2019].

Portanto, o presente trabalho se propõe a apresentar o Aprendizado por Agrupamento e suas aplicações. Este artigo está organizado de forma a apresentar: A seção 2 com a fundamentação teórica do tema; A seção 3 com as aplicações do tema, e, por fim, a Seção 4 com as conclusões.

2. Fundamentação Teórica

A construção de bons modelos preditivos é condicionada por vários fatores, especialmente a relação com o Viés (*Bias*) e a Variância [Geman et al. 1992]. O *Bias* é a diferença entre o valor esperado da predição e o valor real pretendido, e a Variância captura as variações

aleatórias do algoritmo, de uma amostra a outra. Os modelos têm que balancear esses valores e evitar a alta deles.

Um *Ensemble* consiste em um conjunto de modelos e um método que os combinam, para a produção de um modelo mais forte. Os métodos mais populares são o *Bagging* e o *Boosting*. Ambos são introduzidos nos próximos parágrafos, além disso, é apresentado também o *Random Forest*.

2.1. *Bagging*

Na técnica *Bagging* – acrônimo derivado de *Bootstrap AGGregatING* –, apresentada em [Breiman 1996], cada membro do grupo de aprendizagem é construído a partir de um conjunto de dados (*dataset*) diferente de treino. Cada *dataset* é uma amostra de um total de *N* exemplos do *dataset* original, escolhendo *N* itens de forma aleatória. Os modelos são combinados pelo valor que mais aparece – votação –, em caso de classificação, ou por uma média uniforme.

Bagging funciona melhor com *Unstable learners*, ou seja, aqueles que produzem padrões de generalização diferentes com pequenas alterações nos dados de treinamento. Portanto, *Bagging* tende a não funcionar bem em modelos lineares.

2.2. *Boosting*

A ideia do *Boosting* surgiu de questionamentos e hipóteses apresentadas por Kearns (1988) e Kearns e Valiant (1989), onde é levantada a questão: "Pode um conjunto de aprendizes fracos criarem um aprendiz forte?". Em [Schapire 1990], é defendido que sim e como consequência é apresentado o *Boosting*. O *Boosting* é uma família de métodos de *Ensemble Learning* e da mesma forma como *Bagging*, os métodos de *Boosting* se baseiam em diversos modelos mais simples com o objetivo de produzir um modelo final robusto, o aprendiz forte.

Os modelos são treinados de forma sequencial, a partir de uma análise dos modelos treinados anteriormente [Schapire 1999]. Para otimizar o desempenho do modelo final, o *Boosting* treina de forma iterativa novos modelos com ênfase nas dificuldades dos modelos anteriores, desta forma, fazendo a predição mais resistente a *bias*. Em seguida, atualizamos o modelo para priorizar as predições com maior acurácia nas observações do *dataset* de teste.

2.3. *Random Forest*

Abordada em [Breiman 2001], *Random Forest*, ou Floresta Aleatória, é uma técnica de aprendizagem por agrupamento. É um híbrido do algoritmo *Bagging* e do método *Random Subspace*, apresentado em [Ho 1998], que tenta reduzir a correlação entre estimadores em um conjunto treinando-os em amostras aleatórias de recursos, e usa árvores de decisão, que é uma representação de uma tabela de decisão sob a forma de árvore, como classificador base. Cria muitas árvores de decisão aleatoriamente, formando o que podemos enxergar como uma floresta, em que cada árvore será utilizada na escolha do resultado final [Sammut and Webb 2011].

3. Aplicações

[Santos 2020] análisa técnicas de ciência de dados para entender como estudos observacionais podem contribuir para a área das políticas públicas de saúde. Em seus resultados,

apresenta três artigos como resultados. Em [Santos et al. 2020], ainda não publicado, realiza uma análise de inteligência artificial para ausência laboral por motivos de doença com uma amostra populacional de Inquérito Nacional de Saúde, aplicando o algoritmo *Random Forest*.

[Xiao 2019] ratifica a importância e uso de aprendizado de máquina na detecção de incidentes de tráfego. Porém, reafirma que muitos métodos utilizados nem sempre tem bom desempenho. Então, é proposto o uso de aprendizagem por agrupamento para melhorar a robustez na detecção de incidentes de tráfego. O método treina individualmente modelos e depois estrategicamente os combinam para melhorar o resultado final. Como resultados, demonstraram que há superioridade do método proposto aos demais.

A previsão de falência é um tema relevante para pesquisar e de grande impacto socioeconômico. Em [Chen et al. 2020], a solução proposta é otimizada com uso de estratégias de aprendizado por agrupamento. Como contribuição, este trabalho apresentou dois novos métodos de predição que utilizam as estratégias de *Ensemble*, *Bagging* e *Boosting*. Os experimentos demonstraram eficiência e superioridade na solução do problema de previsão de falência.

[Shen et al. 2020] realizam um estudo sobre a estimativa de capacidade de baterias recarregáveis de íon de lítio (*Li-Ion*). Para lidar com o custo e demora de tratar uma grande base de dados, é aplicado uma solução em aprendizagem profunda que incorpora os conceitos de aprendizagem por transferência e aprendizagem por agrupamento. Na aplicação do *Ensemble*, o *Bias* e a Variância são considerados para a melhora do modelo. Os resultados de verificação e comparação demonstram que o método – com aprendizagem por conjunto e aprendizagem de transferência – proposto pode produzir uma maior precisão e robustez do que esses outros métodos baseados em dados na estimativa das capacidades das células de íons de lítio na tarefa alvo.

4. Conclusões

Aprendizado por Agrupamento é um conjunto de procedimentos para otimizar a construção de bons modelos preditivos de Aprendizado de Máquina. Trabalhando com um "grupo" de modelos mais fracos, para a obtenção de um modelo mais robusto e eficiente. O presente trabalho procurou abordar os principais tópicos referentes a temática, como: relação entre *Bias* e Variância com a eficiência dos modelos e a necessidade de correção desses aspectos; os principais métodos para implementação de aprendizagem por agrupamento, o *Bagging* e o *Boosting*, além de introduzir a descrição do *Random Forest*; e, por fim, fornecer exemplos de aplicações da temática.

Ensemble Learning foi concebido inicialmente para trabalhos com Aprendizagem Supervisionada. Inclusive, os principais exemplos citados nesse trabalho, em especial na seção de Aplicações, tratam exclusivamente desse tipo de tarefa, como na aplicação de energia, no caso, as baterias de lítio; na parte socioeconômica com a falência de empresas, assim como também na área da saúde; e também em incidentes de tráfego.

Recentemente, existem aplicações em tarefas de Aprendizagem Não-Supervisionada. Portanto, como trabalhos futuros, pretende-se pesquisar acerca exclusivamente da aplicação de *Ensemble* em tarefas não-supervisionadas, entender os benefícios e os eventuais desafios encontrados.

Referências

Breiman, L. (1996). Bagging predictors. *Mach. Learn.*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chen, Z., Chen, W., and Shi, Y. (2020). Ensemble learning with label proportions for bankruptcy prediction. *Expert Systems with Applications*, 146:113155.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4(1):1–58.

Ho, T. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20:832–844.

Kearns, M. (1988). Thoughts on hypothesis boosting. Unpublished.

Kearns, M. and Valiant, L. G. (1989). Crytographic limitations on learning boolean formulae and finite automata. In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, STOC '89, page 433–444, New York, NY, USA. Association for Computing Machinery.

Lv, Y., Peng, S., Yuan, Y., Wang, C., Yin, P., Liu, J., and Wang, C. (2019). A classifier using online bagging ensemble method for big data stream learning. *Tsinghua Science and Technology*, 24(4):379–388.

Nilsson, N. J. (1965). *Learning Machines*. McGraw-Hill, New York.

Russell, S. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, USA, 3rd edition.

Sammut, C. and Webb, G. I., editors (2011). *Encyclopedia of Machine Learning*. Springer Reference. Springer, New York.

Santos, J. R. R. (2020). *Ciência de Dados e políticas publicas de saúde: exemplos práticos*. PhD thesis, Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo.

Santos, J. R. R., Dias, C. M., and Filho, A. D. P. C. (2020). A machine learning approach for classifying work absenteeism.

Schapire, R. E. (1990). The strength of weak learnability. *Mach. Learn.*, 5(2):197–227.

Schapire, R. E. (1999). A brief introduction to boosting. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, page 1401–1406, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Shen, S., Sadoughi, M., Li, M., Wang, Z., and Hu, C. (2020). Deep convolutional neural networks with ensemble learning and transfer learning for capacity estimation of lithium-ion batteries. *Applied Energy*, 260:114296.

Xiao, J. (2019). Svm and knn ensemble learning for traffic incident detection. *Physica A: Statistical Mechanics and its Applications*, 517:29–35.