Fichamento artigo “An experimental methodology to evaluate machine learning methods for fault diagnosis based on vibration signals”

https://authors.elsevier.com/c/1cZKp3PiGTFJtU

Dentro do contexto de diagnóstico de falhas com base em sinais de vibração, o artigo propõe uma metodologia experimental de avaliação de abordagens de *machine learning*, apresentando um procedimento sistemático para comparar de maneira imparcial os *scores* de desempenho experimental para os métodos utilizados. O artigo aborda também outras questões que visam isolar completamente o conjunto de testes, evitando desta forma o viés de similaridade, verificando diferenças estatisticamente significativas e se preocupa também em permitir a reprodutibilidade. A preocupação em evitar o viés de similaridade é um diferencial, pois até então essa questão não foi aplicada no contexto de diagnóstico de falhas com base em sinais de vibração. Os outros pontos já foram trabalhados em outros artigos, porém a preocupação com todos esses pontos em um mesmo trabalho também é um diferencial.

Uma das principais hipóteses deste trabalho é mostrar que uma validação cruzada aninhada, embora não obtenha *scores* de desempenho superotimistas, é sempre melhor do que a validação cruzada convencional. Para demonstrar essa hipótese foram selecionados 4 classificadores (KNN, SVM, RF e MLP, esta nos experimentos foi substituída por uma Rede Neural Convolucional) e de 2 a 4 hiperparâmetros de cada classificador para comparação da validação cruzada simples (chamada no artigo de enviesada) com a validação cruzada aninhada. Também foram definidos 4 *datasets* com limite de precisão de Bayes conhecido. Após realizar diversas comparações ao utilizar a validação cruzada aninhada e a enviesada, observou-se que para um pequeno número de amostras em uma tarefa de aprendizagem supervisionada, a estimativa de desempenho é quase sempre superotimista, independente do método de validação cruzada; que para um grande número de amostras, as técnicas mais simples e as técnicas mais sofisticadas de validação cruzada apresentam resultados semelhantes; e que para um pequeno número de amostras a validação cruzada aninhada fornece geralmente scores mais conservadores. Foram consideradas as métricas de precisão e *F1 score*. Os testes realizados foram uma aplicação prática da validação cruzada aninhada proposta. Também foi criado um cenário mais realista para o diagnóstico de falhas através da fusão de várias condições da máquina em

uma única classe. Adicionalmente foi avaliada a capacidade de generalização do classificador através da submissão de dados que nunca foram vistos durante o treinamento. A partir daí mostrou-se qualitativamente que as pontuações de alto desempenho que são apresentadas na maioria das publicações são irrealistas para um ambiente de diagnóstico de falhas no mundo real.

A metodologia apresentada no artigo não foi utilizada completamente em nenhum outro artigo publicado (uso de validação aninhada, reprodutibilidade, evitar o viés de similaridade e aplicar testes estatísticos para mostrar diferenças significativas entre métodos propostos e outros métodos), tornando-se um artigo de peso em relação ao conteúdo/metodologia apresentado. Apresenta também que muitos trabalhos já realizados apresentam resultados questionáveis, uma vez que empregam uma divisão simplista de treino/teste único, reutilizam os mesmos dados em vários estágios do aprendizado do classificador ou apresentam métodos que não podem ser verificados. Outro ponto forte é a inclusão dos endereços para acesso código fonte produzido de forma a facilitar as comparações experimentais de métodos envolvendo os dados CWRU. Outro ponto forte é o apoio dado a trabalhos futuros apresentando técnicas que ajudam a chegada de resultados mais próximos da realidade.

No artigo é citado: “Um artigo de boa qualidade deve fornecer uma descrição meticulosa do projeto experimental para permitir que a comunidade científica verifique a veracidade dos critérios de desempenho.” Acredito que este seja um ponto fraco (porém, de maneira geral, para a comunidade é um ponto forte), pois torna o artigo mais longo sendo necessário ser mais criterioso em sua leitura, porém isso é negativo apenas para quem não está interessado na reprodutibilidade e outras características existentes neste artigo. Também vejo como um ponto fraco outra questão citada que refere-se ao não acompanhamento de um manual dos dados utilizados, sobre como realizar experimentos com eles, deixando um grande espaço para a interpretação de quem estiver trabalhando com eles, não existindo assim um direcionamento mínimo por quem produziu o *dataset* de como se trabalhar com ele. Isso permitiria a existência de interpretações incorretas que poderiam invalidar algum trabalho realizado.

Meu projeto de dissertação tem como título “Uso de heurísticas para o ajuste dos hiperparâmeros do *XGBoost* aplicado à recomendação de propostas de recomendação de

refinanciamento”. Como citado no artigo, o ajuste dos hiperparâmetros raramente é feito, com poucos trabalhos utilizando um procedimento automático. Acredito que esta não seja uma situação comum apenas no caso do diagnóstico de falhas baseado nos sinais de vibração mas também em vários outros cenários. Meu projeto tem como um dos objetivos a implementação de heurísticas para a calibração de hiperparâmetros selecionados do classificador XGBoost, baseando-se nas seguintes meta-heurísticas: Grasp, bary search, simulated annealing, genético e random search. Entendo que esse trabalho pode ajudar a aprimorar a forma como a questão do ajuste dos hiperparâmetros é realizado, buscando conseguir melhores resultados na classificação. Após ser feito o trabalho de calibração utilizando as técnicas citadas, certamente irei utilizar técnicas utilizadas neste artigo para aplicação na classificação que será feita sobre os dados, como por exemplo a validação cruzada aninhada.