UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA

Wallace Soares dos Santos Júnior

Proposta de Trabalho de Graduação

Adição de um seletor de atributos ao Ranking-based instance selection

Recife

2021

1 INTRODUÇÃO

Com o aumento exponencial da quantidade de dados produzidos pela humanidade aumenta também a necessidade de possuir ferramentas de processamento cada vez mais robustas para processar estes dados. Entretanto, a cada dia que passa fica evidente que a indústria de *hardware* tem ficado para trás e não consegue em tempo hábil produzir dispositivos capazes de processar estes dados. Se faz necessário a criação de mecanismos para evitar esta dependência de *hardware*. Tais mecanismos já foram propostos por diversos pesquisadores da área: redução de dimensionalidade, redução de instâncias e redução de atributos são alguns deles. Estes mecanismos evitariam, no geral, a necessidade contínua de mais recursos computacionais para gerar um resultado. Por exemplo, como dito por Miao e Niu (2016), a seleção de atributos geralmente pode levar a um melhor desempenho na aprendizagem, isto é, maiores acurácias, menores custos computacionais e uma melhor interpretação de modelos. O mesmo raciocínio se aplica à seleção de instâncias. Como demonstrado por Cavalcanti e Soares (2020), é possível, sem fase de treinamento, selecionar um sub-conjunto de instâncias a partir de conjunto de treinamento para serem usadas durante o processo de classificação. Este procedimento inclusive obteve melhores resultados em alguns casos na acurácia da classificação. Entretanto a combinação destas técnicas de seleção não é bem explorada visto que o resultado é frequentemente afetado(TSAI et al., 2021).

Neste trabalho exploraremos os resultados da combinação entre o algoritmo de seleção de atributos, RelieF (KONONENKO, 1994), e o algoritmo de seleção de instâncias, Ranking-based Instance Selection (RIS) (CAVALCANTI; SOARES, 2020). Em particular faremos uma análise comparativa com os resultados encontrados por Cavalcanti e Soares (2020). Será observado que para alguns casos abordados por Cavalcanti e Soares (2020), a combinação do RelieF ao RIS se mostrou positiva, melhorando os resultados de classificação mesmo utilizando uma menor quantidade de dados ao selecionar as instâncias.

2 OBJETIVO

O objetivo geral deste trabalho é, a partir dos estudos realizados por Cavalcanti e Soares (2020), entender a influência do uso do RelieF combinado ao RIS. Entender como os resultados na acurácia e na redução do número de instâncias é afetado. Para os objetivos específicos o foco será entender qual seria a ordem dos processos que traria maiores benefícios ao resultado. Este estudo será realizado permutando cada algoritmo em dois experimentos diferentes. No primeiro será realizado o RIS para depois executar o RelieF. E um segundo experimento realizando o inverso. Estes dois experimentos terão como base de comparação a execução do RIS isoladamente e utilizarão o próprio classificador do RIS para gerar as métricas de acurácia.

3 METODOLOGIA

Para avaliar como o RIS se comportaria ao receber um conjunto de dados reduzidos em atributos foram realizadas duas abordagens.

- Primeiramente colocamos o RIS para executar a seleção de instâncias. Após a seleção de instâncias a seleção de atributos é feita colocando um limite na quantidade de atributos selecionados. Este limite foi, para todas as bases testadas, de 50% dos atributos.
- A segunda abordagem é realizar o mesmo processo de forma inversa. Isto é, primeiramente selecionamos os atributos, colocando um limite para selecionar aproximadamente 50% dos atributos. E após a seleção de atributos realizar a seleção de instâncias.

Apesar do RelieF possuir um parâmetro *t - threshold* que limitaria quantos atributos seriam selecionados, Urbanowicz et al. (2018) identifica que limitar por um número de atributos já escolhido *a priori* com base no poder computacional é mais efetivo.

Para critérios de comparação foi realizado também o experimento do RIS isoladamente acoplado ao processo descrito no primeiro item acima. Ou seja, as mesmas instâncias selecionadas, quando executamos o RIS como primeira parte do processo, foram utilizadas tanto para recolhimento de métricas utilizando somente o RIS, como também para RIS + RelieF. Desta forma obteríamos uma comparação justa no impacto da seleção dos atributos já que o RIS é não-determinístico.

Para cada um dos experimentos foram recolhidas métricas para fins de comparação. As métricas são: acurácia de classificação, redução no número de instâncias, redução no número de atributos e tamanho da matriz de dados resultante. Para fins de clareza, o tamanho da matriz de dados merece uma explicação mais detalhada. Esta métrica está relacionada diretamente a quantidade de informação que foi reduzida. Este dado se resume a equação abaixo:

$$tamanhoMatriz = QtdInstancias * QtdAtributos \tag{3.1}$$

Desta forma, se uma base de dados de 100 instâncias e 100 atributos for reduzida em 40 instâncias e 60 atributos, o tamanho resultante desta matriz de dados será de 24% da original.

Com essas métricas será possível traçar o impacto causa pela seleção de atributos combinada ao RIS.

4 CRONOGRAMA

O cronograma das atividades está descrito na tabela 1 abaixo. Vale destacar que as fases de implementação e estudo do algoritmo vem sendo realizadas desde abril do ano de 2020, por isso não estão contempladas neste cronograma.

Tabela 1 – Cronograma de Atividades

Atividade	**Junho**	**Julho**	**Agosto**
Experimentos	X | X | |		
Avaliação de resultados	| | X | X		
Revisão bibliográfica		X | X | |	
Escrita do TG		| | X | X	
Preparação da apresentação			X | X | |
Implementação	NA	NA	NA
Estudo do algoritmo	NA	NA	NA

REFERÊNCIAS

CAVALCANTI, G. D.; SOARES, R. J. Ranking-based instance selection for pattern classification. *Expert Systems with Applications*, v. 150, p. 113269, 2020. ISSN 0957-4174.

KONONENKO, I. Estimating attributes: Analysis and extensions of relief. In: BERGADANO, F.; RAEDT, L. D. (Ed.). *Machine Learning: ECML-94*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994. p. 171–182. ISBN 978-3-540-48365-6.

MIAO, J.; NIU, L. A survey on feature selection. *Procedia Computer Science*, v. 91, p. 919–926, 2016. ISSN 1877-0509. Promoting Business Analytics and Quantitative Management of Technology: 4th International Conference on Information Technology and Quantitative Management (ITQM 2016). Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050916313047>.

TSAI, C.-F.; SUE, K.-L.; HU, Y.-H.; CHIU, A. Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction. *Journal of Business Research*, v. 130, p. 200–209, 2021. ISSN 0148-2963. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0148296321001776>.

URBANOWICZ, R. J.; MEEKER, M.; La Cava, W.; OLSON, R. S.; MOORE, J. H. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, v. 85, p. 189–203, 2018. ISSN 1532-0464. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1532046418301400>.

5 POSSÍVEIS AVALIADORES

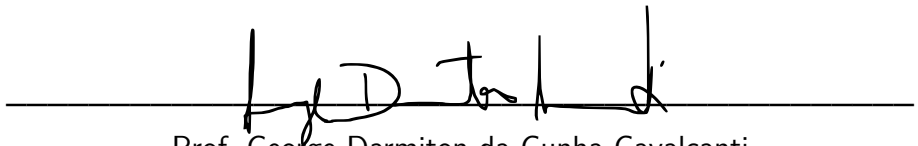
Os possíveis avaliadores escolhidos para este trabalho de graduação são:

- Tsang Ing Ren
- Fernando Maciano

6 ASSINATURAS

Wallace Soares dos Santos Júnior

Aluno



Prof. George Darmiton da Cunha Cavalcanti

Orientador

Recife, 2 de Junho de 2021