ANÁLISE DO DESEMPENHO DO ALGORITMO JRIP UTILIZANDO A BASE DE DADOS "MONK'S PROBLEMS"

Alan Ehrich de Moura [1]

INTRODUÇÃO

A base de dados "MONK's Problems" ("Problemas dos Monges", tradução do autor) foi desenvolvida para a primeira comparação internacional de algoritmos de aprendizagem. Consiste em três problemas envolvendo robôs, sendo a tarefa de aprendizagem uma tarefa de classificação binária. Cada problema é fornecido por uma descrição lógica de uma classe. Os robôs pertencem ou não a esta classe, mas em vez de fornecer uma descrição completa da classe para o problema de aprendizagem, apenas um subconjunto de todos os 432 robôs possíveis com sua classificação é fornecido. A tarefa de aprendizagem é, então, generalizar sobre esses exemplos e, se a técnica de aprendizagem específica disponível permitir, derivar uma descrição de classe simples (THRUN et al., 1991).

Os "MONK's Problems" têm sido extensivamente utilizados para testar a qualidade dos métodos de aprendizagem (PARVIN et al., 2013; BUENO-CRESPO et al., 2017). Esses métodos são implementados por meio de algoritmos computacionais que se originam a partir da confluência de múltiplas disciplinas, são elas: inteligência artificial, aprendizado de máquina, estatística, reconhecimento de padrões e sistema de banco de dados (LAUSCH; SCHMIDT; TISCHENDORF, 2015). A partir dos padrões descobertos com a aplicação desses métodos, pode-se gerar conhecimento útil para dar suporte ao processo de tomada de decisão.

Analisar corretamente as informações fornecidas por bases de dados está entre os requisitos fundamentais para uma boa tomada de decisão. Nesse cenário, os algoritmos computacionais são capazes de captar um conjunto de dados ocorridos no mundo real e produzir um padrão de comportamento, o qual pode ser manifesto, por exemplo, como uma regra de associação, uma função de mapeamento ou a modelagem de um perfil (SILVA; PERES; BOSCARIOLI, 2016). Dessa forma, os algoritmos fornecem bons índices de precisão e padronização na análise dos dados, permitindo ao tomador de decisão ter clareza e eficácia naquilo que for decidir.

[1] Mestrando do Programa de Pós-graduação em Modelos de Decisão e Saúde da Universidade Federal da Paraíba – UFPB, alanehrich@gmail.com.

Existem diversos algoritmos para diferentes finalidades, entre as técnicas mais utilizadas estão as árvores de decisão, redes neurais e regras de classificação. Essas últimas preveem um valor que um determinado atributo do conjunto assumirá dado um conjunto de valores dos demais atributos (SOCZEK; ORLOVSKI, 2014). Um dos algoritmos de aprendizagem de regras de classificação mais poderoso é o *Repeated Incremental Pruning to Produce Error Reduction* (RIPPER) que se trata de uma versão otimizada do Incremental *Reduced Error Pruning* (IREP), pois obtém menos erros de generalização e consegue processar com eficiência conjuntos extensos de dados ruidosos (COHEN, 1995).

O RIPPER (também conhecido como JRip) baseia-se em regras de associação com podas para redução de erros, técnica muito comum e eficaz utilizada em árvores de decisão. Esse algoritmo implementa a estratégia "dividir-para-conquistar", de modo que elenca linearmente o número de exemplos para treinamento em sua base de regras, até que as chances de erro sejam as menores possíveis de serem detectadas pelo sistema (ASADI; SHAHRABI, 2016). A regra produzida com menor incidência de erro é eleita para a classificação, ou seja, a classe que se sobressai é escolhida como padrão e auxilia na determinação da classe minoritária (FRANK; WITTEN, 1998).

Várias áreas utilizam o método RIPPER para tomadas de decisões, entre elas: saúde, genética, computação, economia, psicologia e educação. Destacam-se os estudos sobre diagnóstico e definição do tratamento para diabetes (EYASU; JIMMA; TADESSE, 2020), identificação dos fatores que afetam a visita de cuidados pós-natal (SAHLE, 2016), diagnóstico de doença de Alzheimer a partir dos escores em testes neuropsicológicos (SHREE; SHESHADRI; MURALIKRISHNA, 2016), previsão dos genes relacionados ao câncer colorretal (QING et al., 2015), detecção de comportamento de malware (KAUR; SINGH, 2017; FAIZAL; YASSIN; NUR HIDAYAH, 2019), definição de concessão de crédito bancário para pessoas jurídicas (STEINER et al., 2007), previsão de preços de imóveis (PARK; BAE, 2015), diagnóstico de transtornos psicóticos envolvendo fatores de vulnerabilidade social (CARVALHO et al., 2018) e previsão de desempenho dos alunos do curso de ciências sociais (HUSSAIN et al., 2018).

Diante do exposto, este estudo teve como objetivo geral verificar a eficácia do algoritmo JRip com a base de dados "MONK's problems". Mais especificamente, pretendeu-se estimar os melhores parâmetros para o modelo proposto, em termos de: coeficiente kappa, matriz de confusão e precisão de acertos e erros.

METODOLOGIA

Trata-se de uma pesquisa do tipo experimental exploratória, transversal de abordagem quantitativa, realizada a partir de experimentos e simulações de parametrização do algoritmo JRip. Este estudo utilizou o banco de dados "MONK's Problems" disponível em repositório internacional (http://archive.ics.uci.edu/ml/datasets/MONK%27s+Problems) em formato arff. O banco possui três problemas diferentes que descrevem robôs usando seis atributos distintos. Cada problema fornece um conjunto de exemplos de treino e teste que diferem no tipo de conceito de destino a ser aprendido e na quantidade de ruído nos dados possíveis (216 positivos e 216 negativos). Os conceitos 'verdadeiros' subjacentes a cada problema são dados por: MONK-1: ($attribute_1$ = $attribute_2$) ou ($attribute_5$ = 1); MONK-2: ($attribute_i$ = 1) para EXATAMENTE DOIS $i \in \{1, 2, 3, 4, 5, 6\}$; e MONK-3: ($attribute_5$ = 3 e $attribute_4$ = 1) ou ($attribute_5 \neq 4$ e $attribute_2 \neq 3$).

Existem 124, 169 e 122 amostras nos conjuntos de treinamento dos MONK-1, MONK-2 e MONK-3, respectivamente. O conjunto de teste tem 432 instâncias. Ademais, no conjunto de treinamento do MONK-3 há 5% de ruído adicional (classificações erradas). Para o presente estudo foi utilizado apenas o MONK-1 pois é facilmente aprendido por todos os algoritmos de aprendizagem simbólica (THRUN et al., 1991).

Os dados foram analisados através do software *Waikato Environment for Knowledge Analysis* (WEKA), versão 3.8.4. No WEKA estão incluídos quatro modos de teste para avaliar o resultado da aplicação do classificador, são eles: Use *Training Set*, *Supplied Test Set*, *Cross Validation* e *Percentage Split* (HALL et al., 2009). Para este estudo decidiu-se utilizar o *Supplied Test Set*, uma vez que esta opção permite que o usuário determine os conjuntos de treinamento e teste. Essa decisão deveu-se às características oriundas da base de dados supracitada.

Utilizou-se o JRip como classificador baseado em regras SE - ENTÃO (IF - THEN). Uma regra SE - ENTÃO é uma expressão da forma: *SE condição ENTÃO conclusão*. Este algoritmo divide-se em duas fases: 1) gera uma lista de regras para a comparação; e 2) otimiza o conjunto de regras iniciais para diminuir erros e tornar o processo mais seletivo (ASADI; SHAHRABI, 2016).

Para a realização das análises, foram padronizados alguns parâmetros do método JRip como: pesos mínimos das instâncias dentro de uma divisão (*minNo* = 2.0), número de execuções otimizadas (*optimizations* = 2) e semente de randomização (*seed* =1). Contudo, foram feitos agrupamentos de 2 a 4 *folds*. Ademais, foram reportadas as porcentagens de acertos e erros, o

coeficiente Kappa e a matriz de confusão, com intenção de obter os melhores valores na quantidade de regras geradas.

RESULTADOS E DISCUSSÃO

As simulações evidenciaram resultados fortemente satisfatórios a partir do *Supplied Test Set*. Observou-se que o modelo com poda incremental e três *folds* apresentou dados adequados com um coeficiente Kappa de 100 % e 0% de erros, evidenciado na matriz de confusão. Identificou-se, ainda, uma precisão de 100% de acertos na sua diagonal principal, apresentando o melhor resultado de classificação.

O conjunto de treinamento para o problema representou um tipo de classificação para o robô. O conjunto de teste foi então usado para determinar quão bem o algoritmo aprendeu a reconhecer aquela classe. Nesse sentido, o algoritmo JRip gerou como resultado um conjunto de seis regras: SE ($attribute_5$ >= 2) AND ($attribute_1$ <= 2) AND ($attribute_2$ >= 3) ENTÃO *class = positive*; SE ($attribute_2$ <= 2) AND ($attribute_5$ >= 2) AND ($attribute_1$ >= 3) ENTÃO *class = positive*; SE ($attribute_1$ <= 1) AND ($attribute_5$ >= 3) AND ($attribute_2$ >= 2) ENTÃO *class = positive*; SE ($attribute_2$ <= 1) AND ($attribute_1$ >= 2) AND ($attribute_5$ >= 2) ENTÃO *class = positive*; SE ($attribute_1$ <= 1) AND ($attribute_2$ >= 2) AND ($attribute_5$ >= 2) ENTÃO *class = positive*; CASO CONTRÁRIO ENTÃO *class = negative*. Percebe-se que foram cinco regras para a classe 0 (que representa exemplos positivos do conceito) e uma regra para a classe 1 (que representa os exemplos negativos).

O JRip classificou corretamente todas as 432 instâncias do conjunto de teste. Considera-se que estes resultados estão em consonância com a literatura atual que salientam o poder do JRip em classificar corretamente sem produzir maiores erros. De acordo com Steiner et al. (2007), o JRip apresentou melhor desempenho em praticamente todas as vezes em que o mesmo foi comparado com outros algoritmos de aprendizagem. Além disso, dentre os dez métodos para obtenção das regras de classificação contidos no software WEKA, o método JRip apresentou a maior taxa de precisão.

Ademais, devido ao MONK-1 ser considerado fácil de aprender por todos os algoritmos de aprendizagem simbólicos, outras pesquisas também evidenciaram modelos de domínio perfeito (acurácia de 100%) para outros métodos de classificação, entre eles: AQ17-DCI, AQ17-HCI, *Assistant Professional*, mFOIL, CN2, *Backpropagation* e *Cascade Correlation* (THRUN et al., 1991).

CONSIDERAÇÕES FINAIS

Conclui-se que o algoritmo JRip mostrou-se um poderoso algoritmo de aprendizagem de regras de classificação. Ademais, a base de dados "MONK's problems" foi criada especialmente para testar a qualidade dos métodos de aprendizagem. Portanto, as conclusões sobre ela são verossímeis e valiosas. Sugere-se que em estudos futuros os problemas MONK-2 e MONK-3 também sejam utilizados para conceder maior eficiência do algoritmo JRip.

Assim, confirmou-se a utilidade e eficiência do JRip, podendo ser utilizado por diversas áreas para produção de redução de erros em determinados estudos através de regras e poda incremental repetida. O JRip aprendeu um pequeno conjunto de regras com um nível de precisão muito bom. Tal combinação (ou seja, conjunto de regras gerenciáveis e alta precisão entre as abordagens simbólicas) torna a escolha de JRip perfeitamente adequada para sua utilização em pesquisas.

Palavras-chave: Machine learning, JRip, Algoritmo, WEKA, Data mining.

REFERÊNCIAS

ASADI, S.; SHAHRABI, J. RipMC: RIPPER for Multiclass Classification. **Neurocomputing**, v. 191, p. 19-33, mai. 2016.

BUENO-CRESPO et al. Bioinspired Architecture Selection for Multitask Learning. **Frontiers in Neuroinformatics**, v. 11, n. 39, p. 1-9, jun. 2017.

CARVALHO et al. Rule induction algorithms for classification of psychotic disorders involving social vulnerability features. **Procedia Computer Science**, v. 138, p. 49-55, 2018.

COHEN, W. W. Fast Effective Rule Induction. In: KAUFMANN, M. (Org.). **Twelfth International Conference on Machine Learning**. 12ª ed. Burlington, MA: Morgan Kaufmann Publishers, 1995, p. 115-123.

EYASU, K.; JIMMA, W.; TADESSE, T. Developing a Prototype Knowledge-Based System for Diagnosis and Treatment of Diabetes Using Data Mining Techniques. **Ethiopian Journal Of Health Sciences**, v. 30, n. 1, p. 115-124, jan. 2020.

FAIZAL, M. A.; YASSIN, W.; NUR HIDAYAH, M. S. Scrutinized System Calls Information Using J48 And Jrip For Malware Behaviour Detection. **Journal Of Engineering Science And Technology**, v. 14, n. 1, p. 291-304, 2019.

FRANK, E.; WITTEN, I. H. Generating accurate rule sets without global optimization. In: SHAVLIK, J. **Proceedings of the Fifteenth International Conference on Machine Learning**. 15ª ed. Burlington, MA: Morgan Kaufmann Publishers, 1998, p. 144-151.

HALL et al. The WEKA Data Mining Software: An Update. **ACM SIGKDD Explorations Newsletter**, v. 11, n. 1, p. 10-18, nov. 2009.

HUSSAIN et al. Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. **Computational Intelligence and Neuroscience**, v. 2018, p. 1-21, out. 2018.

KAUR, G.; SINGH, J. Classification of Malicious URLs for Web Using Ripper Algorithm. **International Journal of Advanced Research in Computer Science**, v. 8, n. 7, p. 137-139, ago. 2017.

LAUSCH, A.; SCHMIDT, A.; TISCHENDORF, L. Data mining and linked open data – New perspectives for data analysis in environmental research. **Ecological Modelling**, v. 295, p. 5-17, jan. 2015.

PARK, B.; BAE, J. K. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. **Expert Systems with Applications**, v. 42, n. 6, p. 2928-2934, abr. 2015.

PARVIN et al. A new classifier ensemble methodology based on subspace learning. **Journal of Experimental & Theoretical Artificial Intelligence**, v. 25, n. 2, p. 227-250, jun. 2013.

QING et al. Prediction of Colorectal Cancer Related Genes Based on Gene Ontology. **Current Bioinformatics**, v. 10, n. 1, p.22-30, 2015.

SAHLE, G. Ethiopic maternal care data mining: discovering the factors that affect postnatal care visit in Ethiopia. **Health Information Science and Systems**, v. 4, n. 4, p. 1-8, may. 2016.

SHREE, S. R. B.; SHESHADRI, H. S.; MURALIKRISHNA. Diagnosis of Alzheimer's Disease using Rule based Approach. **Indian Journal of Science and Technology**, v. 9, n. 12, p. 1-6, abr. 2016.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados: com aplicações em R**. Rio de Janeiro: Elsevier, 2016.

SOCZEK, F. C.; ORLOVSKI, R. Mineração de Dados: Conceitos e aplicação de algoritmos em uma Base de Dados na área da saúde. **Revista Científica Semana Acadêmica**, v. 01, p. 1-25, 2014.

STEINER et al . Extração de regras de classificação a partir de redes neurais para auxílio à tomada de decisão na concessão de crédito bancário. **Pesquisa Operacional**, v. 27, n. 3, p. 407-426, set. 2007.

THRUN et al. The Monk's Problems-A Performance Comparison of Different Learning Algorithms. **CMU-CS-91-197**, Pittsburgh, PA, 1991, p. 2-80.