A Preliminary Study on the Applied Machine Learning for Detection of the Predominant Factor of Big Five Personality Test

Caio Christian da Rocha
caiochristian28@gmail.com
CEFET/RJ–*Campus* Petrópolis

Diogo Fagundes Pereira
diogofagundes.psi@gmail.com
FASE

Andre Felipe de Almeida Monteiro
andre.monteiro@cefet-rj.br
CEFET/RJ–*Campus* Petrópolis

Felipe da Rocha Henriques
felipe.henriques@cefet-rj.br
CEFET/RJ–*Campus* Petrópolis

ABSTRACT

Nowadays, the importance of mental health has become an increasingly relevant theme. Psychological assessments are part of the daily life of clinical psychologists in order to identify possible issues to be explored. Therefore, this work presents a preliminary study which aims to evaluate the accuracy of machine learning algorithms for the detection of the predominant factor of big five personality test. Real answers from a dataset were considered in the computational experiments, and two machine learning algorithms were evaluated: the K-Nearest Neighbors (KNN) and the K-means. Results show that both algorithms could accurately detect the pedominant factor of the big five test, and KNN obtained better results than the other algorithm.

KEYWORDS

Machine learning, Big five, KNN, K–means

1 INTRODUÇÃO

Atualmente, técnicas de aprendizado de máquina estão por toda parte. Elas são usadas em pesquisas do Google para definir padrões de usuários através da mineração de dados, são usadas em classificação de imagens para a detecção de objetos e pessoas, na predição de doenças como o câncer, na classificação de textos, e até em carros autônomos de modo que aprendam rotas e desafios do tráfego urbano.

No ano de 2018 foi noticiado o desenvolvimento de um algoritmo capaz de detectar a possibilidade de suicídios a partir de textos escritos. Este estudo utilizou textos da escritora britânica Virginia Woolf, antes dela cometer suicídio. Assim como no referido exemplo, é de interesse neste trabalho estudar de que maneira o aprendizado de máquina pode auxiliar na detecção de possíveis problemas psicológicos em indivíduos e na realização de um apoio ao diagnóstico.

1.1 Trabalhos Relacionados

Em [1], os autores desenvolveram uma ferramenta para auxílio na avaliação psicológica de crianças. Mais especificamente, os autores consideram a avaliação do desenho da figura humana, e usam aprendizado profundo (*deep learning*) para identificar possíveis problemas de saúde mental, em testes de triagem infantil.

Um esquema de aprendizado baseado em imagens do cérebro utilizado em [2] tenta caracterizar persistência e remissão de TDAH em jovens que outrora, quando crianças, tenham sido diagnosticados com tal transtorno. Ainda sobre TDAH, os autores de [3] aplicam aprendizado de máquina em uma plataforma de aprendizado virtual para inferir sobre indicadores de TDAH.

1.2 Objetivos e Contribuições

Este trabalho traz alguns resultados preliminares acerca do uso de dois algoritmos de aprendizado de máquina na detecção do fator preponderante do teste de personalidade *Big Five* [4]. Segundo [5], índices altos de alguns dos fatores, como o neuroticismo, por exemplo, podem sugerir propensão a sofrimentos psicológicos. Para os experimentos computacionais, os algoritmos KNN (*K–Nearest Neighbors*) [6] e o K–means [7] foram avaliados, considerando uma base de dados reais disponível na Internet. Os resultados iniciais apresentaram uma boa acurácia de detecção do fator preponderante, tendo o KNN obtido melhores resultados.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 O Teste *Big Five*

O teste de personalidade *Big Five* [4] apresenta cinco marcadores de personalidade presentes em indivíduos. Através deste teste, que se dá por um conjunto de perguntas, conseguimos identificar traços de personalidade que se apresentam através de percentis. Cada percentil quantifica um determinado traço (ou marcador) de personalidade. Os marcadores indicados para os traços de personalidades, de acordo com [4], são os seguintes:

(1) *Surgency*: Extroversão;
(2) *Agreeableness*: Socialização;
(3) *Conscientiousness*: Realização;
(4) *Emotional Stability*: Neuroticismo;
(5) *Openness*: Abertura.

2.2 Algoritmos de Aprendizado de Máquina

Nesta seção, apresentaremos brevemente os dois algoritmos de aprendizado de máquina considerados para os experimentos computacionais realizados neste estudo preliminar.

Um deles é chamado de algoritmo supervisionado, em que ao inserirmos dados de entrada e, após o processamento desses dados e dos resultados serem apresentados, é necessária a comparação entre a resposta obtida e a resposta esperada referente aos dados de entrada. Tendo o resultado dessa comparação, o algoritmo toma a decisão do que se deve fazer para que a acurácia do algoritmo aumente.

Como algoritmo supervisionado, consideramos o **KNN**. O algoritmo assume que dados similares tendem a ficar c oncentrados na mesma região no espaço de dispersão dos dados, como objetos no espaço. Quando um novo dado é inserido no espaço, a distância euclidiana desse objeto em relação a todos os outros precisa ser calculada previamente, considerando um número K de vizinhos mais próximos. Desse modo, os dados vão sendo agrupados a cada iteração até que se convirja em resultados satisfatórios.

No caso do algoritmo não–supervisionado, o aprendizado não depende de comparações entre resultados obtidos e esperados. Inicialmente, seus dados de entrada dispõem-se de maneira completamente aleatória, em *clusters*. A cada iteração, os dados são reagrupados de acordo com o que o algoritmo aprende dos dados, baseado em similaridade.

O **K-means** foi usado como algoritmo não–supervisionado. Primeiramente deve ser informado um número K, relativo a quantidade de classes a ser considerada para a divisão dos objetos (dados) no espaço. Após isso, o algoritmo atribui uma posição aleatória para um objeto, de sorte que ele esteja na distância média entre todos os pontos de uma mesma classe. A partir daí, o algoritmo agrupa os pontos mais próximos formando *clusters*, e vai refinando a posição dos objetos no espaço. A vantagem deste tipo de estratégia é que não se faz necessário conhecer *a priori* os resultados dos dados.

3 SOLUÇÃO PROPOSTA

A Figura 1 apresenta a ideia básica do sistema proposto neste trabalho. Os dados de entrada dizem respeito aos resultados do questionário, que são divididos em dois grupos: um grupo é usado para treinar o algoritmo de aprendizado de máquina, e o outro grupo é usado para testar a acurácia do algoritmo para a detecção do fator preponderante do teste *big five*. O s r esultados t razem o percentual de detecção.

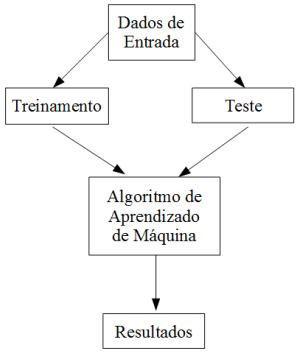

Figura 1: Diagrama da solução proposta.

4 SIMULAÇÕES E RESULTADOS

Para realizar as simulações, consideramos os dados de respostas disponíveis em http://openpsychometrics.org/_rawdata/, dos quais utilizamos dois conjuntos de um total de 50 resultados. Dessa amostra, 30 resultados foram usados no conjunto de treinamento e os outros 20 foram usados no conjunto de testes. Além disso, consideramos $K = 3$. Cada resposta do questionário traz como resultados os percentis dos cinco fatores para aquele indivíduo. Neste estudo inicial, consideramos apenas o resultado do fator preponderante; ou seja, o que apresentou o maior percentil, como dado de entrada para os conjuntos de treinamento e teste. Estes dados de entrada não foram previamente categorizados, por exemplo, em idade, sexo ou nacionalidade, de modo a avaliarmos, neste primeiro momento, a capacidade de generalização dos algoritmos.

Como resultados, obtivemos os seguintes percentuais de acurácia de detecção do fator preponderante:

- 60% de acurácia para o K–means;
- 70% de acurácia para o KNN.

Como este teste se deu com uma amostra não muito grande (30 dados de treinamento e 20 de teste), acreditamos que isso pode ter favorecido o algoritmo supervisionado.

5 CONSIDERAÇÕES FINAIS

Este trabalho realizou um estudo preliminar acerca do uso de dois algoritmos de aprendizado de máquina (um supervisionado e um não–supervisionado) na detecção do fator preponderante do teste de avaliação de personalidade *big five*. Utilizamos resultados reais do teste para os conjuntos de treinamento e teste utilizados nos experimentos computacionais [1].

Acreditamos que a amostra ainda pequena utilizada neste trabalho levou ao resultado de 60% de acurácia, no caso do K–means, além do fato dos dados de entrada não estarem agrupados por idade ou sexo. Contudo, pôde-se observar a aplicação de técnicas de aprendizado de máquina em testes psicológicos, o que pode ser muito útil no auxílio aos profissionais da área de saúde mental.

Como trabalhos futuros, pretendemos aumentar os dados de ambos os conjuntos, além de investigarmos outros algoritmos, como o *Random Forest* e o SVM (*Support Vector Machine*). Ademais, caso haja a identificação de altos índices de um determinado fator, pretendemos incluir mais um estágio ao sistema, de modo que se possa detectar problemas específicos, a depender do fator preponderante associado.

6 AGRADECIMENTOS

Os autores agradecem ao CEFET/RJ (DIPPG) pela bolsa PIBIC disponibilizada para este projeto.

REFERÊNCIAS

[1] Wesley da Silva and Mateus Raeder. Melampus: um modelo deep learning para triagem psicológica infantil. *Revista Brasileira de Computação Aplicada*, 10(3): 21–33, nov. 2018. doi: 10.5335/rbca.v10i3.8471. URL http://seer.upf.br/index.php/rbca/article/view/8471.

[2] Hazel McCarthy, Jessica Stanley, Richard Piech, Norbert Skokauskas, Aisling Mulligan, Gary Donohoe, Diane Mullins, John Kelly, Katherine Johnson, Andrew Fagan, Michael Gill, James Meaney, and Thomas Frodl. Childhood-diagnosed adhd, symptom progression, and reversal learning in adulthood. *Journal of Attention Disorders*, 22(6):561–570, 2018.

[3] L. P. M Valetts, S. M. B. Navarro, and V. B Chicué. Indicadores das sintomas tdah na aprendizagem virtual contexto usando técnicas de aprendizagem automática. *Rev. esc.adm.neg*, (79), 2015.

[4] L. R. Goldberg. The development of markers for the big-five factor structure. *Psychological Assessment*, (1), 1992.

[5] Carlos Henrique Sancineto S. Nunes. A construÇÃo de um instrumento de medida para o fator neuroticismo / estabilidade emocional dentro do modelo de personalidade dos cinco grandes fatores. *Dissertação de Mestrado–UFRGS*, 2000.

[1] Disponível em https://openpsychometrics.org/tests/IPIP-BFFM/.

[6] P. Soucy and G. W. Mineau. A simple knn algorithm for text categorization. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 647–648, Nov 2001. doi: 10.1109/ICDM.2001.989592.

[7] S. Na, L. Xumin, and G. Yong. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, pages 63–67, April 2010. doi: 10.1109/IITSI.2010.74.