UNIVERSIDADE FEDERAL DO CEARÁ

PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CKP8277 - APRENDIZAGEM AUTOMÁTICA - 2018.2

PROPOSTA DE PROJETO FINAL

Armando Soares, Erick Barros, Fabiano Gadelha, Lucas Mapurunga

Descrição do Problema

O Kickstarter[1] é uma plataforma que permite a criação de projetos onde um indivíduo ou organização pode solicitar apoio financeiro de uma comunidade de apoiadores. A plataforma surgiu com o intuito de viabilizar campanhas para projeto inovadores e criativos. Outra característica importante do serviço é a política do "Tudo ou Nada", ou seja, uma vez que o projeto é iniciado ele só será realmente financiado caso todo o valor necessário seja arrecadado até o fim da campanha. A plataforma defende esse tipo de política como uma garantia para os responsáveis e para os apoiadores. Porém, é provável que muitos projetos tenham falhado por conta dessa política estrita como também por conta dos riscos associados a projetos inovadores.

Nesse sentido, o atual trabalho tem como objetivo auxiliar os criadores de projetos, da plataforma Kickstarter, na escolha das características de suas propostas através de modelos de aprendizagem automática para então tentar aumentar sua chance de sucesso. Este trabalho terá um objetivo duplo, sendo o primeiro deles tratar da classificação de sucesso/insucesso de projetos baseados em suas características. O segundo objetivo diz respeito a clusterização de projetos em um determinado intervalo de tempo e, para com isso, tentar identificar padrões nos dados obtidos.

Descrição dos dados

O conjunto de dados selecionado foi o "Kickstarer Projects"[2], disponível na plataforma Kaggle. Além disso, também vamos usar o serviço Webrobots[3] que tem o registro de todos projetos do Kickstarter desde 2015 contendo mais de 38 características. Especificamente serão utilizados os dados coletados de projetos finalizados em 2018 em que os atributos encontram-se listados na Tabela 1.

[1] https://www.kickstarter.com/

[2] https://www.kaggle.com/kemical/kickstarter-projects

[3] https://webrobots.io/kickstarter-datasets

Tabela 1 - Atributos do Conjunto de Dados

Nome do Atributo	Descrição
ID	Identificador do projeto
name	Nome do projeto
category	Sub-categoria da campanha
main_category	Categoria da campanha
currency	A moeda utilizada (ex: USD)
deadline	Prazo final para o *crowdfunding* do projeto
goal	Montante de dinheiro necessário para o projeto
launched	Dia de lançamento da campanha
pledged	Montante de dinheiro que os apoiadores colaboraram para a campanha
backers	Quantidade de apoiadores do projeto
country	País de origem
usd_pledged	Montante de dinheiro que os apoiadores colaboraram para a campanha em USD
state	Estado final do projeto

Para as saídas dos modelos estão previstos dois tipos de resultados: para o modelo de classificação, será fornecida uma resposta de sucesso/insucesso do projeto comparando esse resultado com a coluna "state" da Tabela 1; e para o modelo de clusterização será realizada uma análise dos agrupamentos obtidos para identificar a possível existência de padrões nos dados analisados.

Métodos utilizados

O método selecionado para a classificação dos projetos será o algoritmo de redes neurais Multilayer Perceptron (MLP), pois o mesmo apresenta as características necessárias para lidar com a modelagem de problemas complexos de classificação. [1]

Para a clusterização será utilizado o algoritmo K-médias [2], pois após uma análise prévia dos dados foi possível identificar a existência de alguns agrupamentos relacionados às categorias dos projetos.

Métricas utilizadas

A utilização do conjunto de dados para o treinamento do modelo MLP seguirá o método da divisão do *dataset* em 3 segmentos com as seguintes proporções: 50% para treino, 25% para validação, 25% para teste. [2]

Neste trabalho serão usadas as métricas para avaliação de qualidade dos modelos baseado na estimativa de taxa de erros dos dados de validação em relação a taxa de acerto dos testes [2].

Além disso, será calculada a acurácia da qualidade do resultado, que consiste em calcular a relação entre o somatório da quantidade de verdadeiros positivos e verdadeiros negativos sobre

o total de casos; a precisão, que quantifica a capacidade de predição do modelo; *recall*, que quantifica a capacidade de realizar o correto reconhecimento; e F1 *Score*, que representa a média harmônica entre a taxa de precisão e *recall*, resultando em uma taxa geral de qualidade do modelo. [3]

Referências

1. Haykin, Simon. Redes Neurais: princípios e prática / Simon Haykin; Tradução Paulo Martins Engel. 2a.ed. Porto Alegre: Bookman. 2001.

2. FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Robert. The elements of statistical learning. New York, NY, USA:: Springer series in statistics, 2001.

3. MARQUES, Visctor. Avaliação do desempenho das redes neurais convolucionais na detecção de ovos de esquistossomose. 2017. 38f. Trabalho de Conclusão de Curso - Universidade Federal de Pernambuco, Recife, 2017.