2018

Study and comparison of automatic learning tools: TensorFlow with Google Cloud and Azure Machine Learning

Diogo Manuel Ribeiro Pinto
Instituto Politécnico do Cávado e do Ave, diogoo_mrp@hotmail.com

Nuno Lopes
Instituto Politécnico do Cávado e do Ave, nlopes@ipca.pt

Joaquim P. Silva
Instituto Politécnico do Cávado e do Ave, jpsilva@ipca.pt

Estudo e comparação de ferramentas de aprendizagem automática: TensorFlow da Google Cloud e Azure Machine Learning

Study and comparison of automatic learning tools: TensorFlow with Google Cloud and Azure Machine Learning

Diogo Manuel Ribeiro Pinto, Instituto Politécnico do Cávado e do Ave, Portugal, diogoo_mrp@hotmail.com

Nuno Lopes, Instituto Politécnico do Cávado e do Ave, Portugal, nlopes@ipca.pt

Joaquim P. Silva, Instituto Politécnico do Cávado e do Ave, Portugal, jpsilva@ipca.pt

Resumo

Com o evoluir do mundo tecnológico, devido ao grande volume e diversidade de dados gerados por diversos tipos de sistema de informação, as ferramentas de análise tradicional têm sido insuficientes para compreender o valor que cada um dos dados apresenta. Por isso, a aprendizagem automática é perfeita para explorar as oportunidades dissimuladas, ou seja, é capaz de descobrir e exibir padrões e identificar relações entre dados. Esses dados apresentam um valor importante para diversos tipos de empresas. A necessidade de estar na vanguarda da tecnologia é um passo extremamente importante, pois cada empresa pode se ajustar à mudança do mercado, evitando que os seus clientes mudem para empresas concorrentes.

O objetivo deste trabalho é estabelecer um conjunto de critérios para comparação de um conjunto de ferramentas de aprendizagem automática para o tratamento de grandes de diversas quantidades de informação em plataformas de Big Data. Com base nesses critérios, o utilizador poderá selecionar a melhor ferramenta que se enquadra com o seu objetivo.

Palavras-chave: Big Data; Aprendizagem automática;

Abstract

With the evolution of the technological world, due to the large volume and diversity of data generated by various types of information systems, the traditional analysis tools have been insufficient to understand the value that each data presents. For this reason, automatic learning is perfect for exploiting hidden opportunities, that is, it is able to discover and display patterns and identify relationships between data. These data present an important value for different types of companies. The need to be at the forefront of technology is an extremely important step, as each company can adjust to changing the market by preventing its customers from switching to competing companies.

The objective of this work is to establish a set of criteria for comparing a set of automatic learning tools for the treatment of large quantities of information on Big Data platforms. Based on these criteria, the user can select the best tool that fits his purpose.

Keywords: Big Data; Machine Learning;

1. INTRODUÇÃO

Atualmente, a abundância de dados e a velocidade com que são gerados têm provocado mudanças de planeamento e operação em diversas instâncias organizacionais. Segundo Gantz & Reinsel (2011), *Big Data* é um corte horizontal do universo digital e pode incluir dados transacionais, dados armazenados, metadados e outros dados que residem em grandes arquivos. Estas explosões de dados digitais podem trazer enormes oportunidades e potencial transformador para vários setores, como empresas de manufatura do setor de saúde e serviços educacionais.

Segundo Oussous (2017), anteriormente à revolução do *Big Data*, as empresas não podiam armazenar informação por longos períodos, nem gerir eficientemente enormes conjuntos de dados. As tecnologias tradicionais, além de terem uma capacidade de armazenamento limitada, apresentam uma gestão rígida e de elevado custo. Ao contrário dos dados tradicionais, o termo *Big Data* refere-se a grandes conjuntos de dados que incluem formatos heterogéneos: estruturados, não estruturados e semi-estruturados.

Quando se fala de *Big Data*, são identificadas várias caraterísticas, iniciadas pela letra V que constituem a sua definição, sendo três as mais consensuais. O primeiro V refere-se ao volume dos dados, o segundo à variedade dos diversos tipos de dados e informações, e o terceiro à velocidade com que os dados são gerados. Toda esta informação deriva de três tipos de dados: *Social Data*, *Enterprise Data* e *Personal Data*.

Para Gandomi (2015), as definições de *Big Data* são relativas e variam de acordo com fatores, como o tempo e o tipo de dados. Pode ser considerado que os maiores dados de hoje podem não representar o limite no futuro, porque as capacidades de armazenamento aumentam, permitindo que conjuntos de dados ainda maiores sejam capturados.

Para Keim (2013), como muitas novas tecnologias de informação, *Big Data* pode trazer reduções drásticas de custo e melhorias substanciais no tempo necessário para executar uma tarefa de computação, ou novas ofertas de produtos e serviços. Como a análise tradicional, também pode suportar decisões internas de negócio.

Os pilares do *Big Data* estão na sua definição, os V, mas toda a inteligência está na análise de dados. Sem uma análise correta e criteriosa, é impossível gerar entendimentos/perspetivas e direcionar o caminho mais acertado. O processo da análise passa por inspecionar os dados e criar hipóteses para realizar testes com o objetivo de melhorar ou entender um determinado cenário e os seus padrões.

Para Xue-Wen Chen (2014), embora o *Big Data* ofereça um enorme potencial de revolucionar todos os aspetos da nossa sociedade, a coleta de conhecimento do *Big Data* não é uma tarefa comum. O crescendo de informação oculta nos volumes de dados não tradicionais requer o desenvolvimento de tecnologias avançadas. A aprendizagem automática, juntamente com os avanços no poder

computacional, passou a desempenhar um papel vital na análise de *Big Data* e na descoberta de conhecimento. Ambos são aplicados amplamente para alcançar o poder preditivo do *Big Data* em campos como mecanismos de pesquisa, medicina e astronomia.

Os sistemas de aprendizagem automática aprendem muitos modelos complexos, com bilhões de parâmetros, onde prometem capacidade adequada para digerir conjuntos de dados massivos e oferecem uma análise preditiva poderosa, com recursos latentes de alta dimensionalidade, representações intermediarias e funções de decisão, segundo Xing et al. (2016).

A camada que vai ser explorada será a camada de processamento pois é onde ocorre uma análise real, ou seja, permite processar estruturas e inclui ferramentas para movimentação e interação de dados.

2. TRABALHOS RELACIONADOS

2.1. Aprendizagem automática

Segundo Xing et al. (2016), a ascensão do *Big Data* levou à implementação de sistemas de aprendizagem automática para aprender modelos mais complexos, com milhões de parâmetros, que prometem uma capacidade adequada para digerir conjuntos de dados massivos e oferecer análises preditivas, como recursos latentes de alta dimensionalidade, representações intermediárias e funções de decisão.

A aprendizagem automática tornou-se um mecanismo primário para destilar informação estruturada e conhecimento a partir de dados brutos, permitindo previsões automáticas e hipóteses acionáveis para diversas aplicações, tais como: análise de redes sociais, raciocínio sobre o comportamento do cliente, interpretações de texto, identificação de doenças e conduzir carros sem condutor.

Para Wang (2016), a aprendizagem automática é uma das áreas mais importantes da inteligência artificial. O objetivo é descobrir conhecimento e tomar decisões inteligentes. Os algoritmos da aprendizagem automática podem ser categorizados em supervisionados, não supervisionados e semi-supervisionados.

Segundo Nilsson (2010), a aprendizagem automática é um método de análise de dados que automatiza o desenvolvimento de modelos analíticos. Para isso, utiliza algoritmos que aprendem interactivamente a partir de dados, e por sua vez, permite que os computadores encontrem intuições ocultas sem serem explicitamente programados para procurar algo específico.

Big Data, segundo Chunhe et al. (2016), tem uma baixa densidade de valor, portanto, a amostra completa é frequentemente adotada aquando da análise de dados, o que significa que o volume de dados em larga escala traz desafios sem precedentes para *a* aprendizagem automática.

2.2. Análise de estudos semelhantes

Nos últimos anos tem sido frequente observar que a temática de *Big Data* e Aprendizagem automática tem vindo a ser bastante abordada por diversos académicos e profissionais.

Para Gandomi & Haider (2015) a realização do seu estudo sobre conceitos, métodos e análises em *Big Data*, possibilitou uma consolidação do tema até então fragmentado sobre o que constitui o *Big Data*, quais as métricas e outras características e também quais as ferramentas e tecnologias que existem para aproveitar o potencial do *Big Data*.

Outro estudo realizado por Landset et al. (2015), veio fornecer uma revisão abrangente do estado de arte de ferramentas escaláveis para ser aplicadas na aprendizagem automática. São oferecidas recomendações para critérios das ferramentas, comparações de mecanismos e também bibliotecas e estruturas de aprendizagem automática.

3. MÉTODO COMPARATIVO

3.1. Descrição e caracterização

Segundo Landset et al. (2015) , existe uma variedade de ferramentas de aprendizagem automática criadas para facilitar o processo de aprendizagem, mas muitos pesquisadores e profissionais rejeitam por varias razões, na maioria das vezes porque carecem de recursos necessários ou são difíceis de integrar num ambiente existente. Uma questão é que a aprendizagem automática é um amplo campo de estudo e muitas das ferramentas disponíveis carecem de funcionalidades importantes.

Com a realização deste estudo pretende sobretudo identificar, caracterizar e fornecer uma visão detalhada dos pontos fortes e fracos que cada uma das ferramentas de aprendizagem automática apresentada, baseando em datasets derivados das redes sociais, sobretudo do Twitter e Facebook.

Para Gandomi (2015), a analise das redes sociais refere-se à análise de dados estruturados e não estruturados. A principal característica da análise moderna das *mídia social* é a sua centralização em dados. A pesquisa sobre análise de redes socais abrange várias disciplinas, incluindo, psicologia, sociologia, antropologia, ciência da computação, matemática, física e economia.

A revisão de outros estudos realizados sobre a temática será um ponto de partida para delinear quais os critérios de avaliação e classificação de cada uma das ferramentas.

3.2. Ferramentas de aprendizagem automática

Para a realização deste trabalho as ferramentas escolhidas recaíram sobre duas das maiores empresas do mundo tecnológico, Google e Microsoft. Quanto ao Google Cloud será utilizado o TensorFlow e na Microsoft Azure será utilizada a ferramenta Machine Learning.

Quanto ao Google Cloud, segundo Vishnu et al. (2017), a Google lançou o TensorFlow em novembro de 2015 como uma plataforma para construir e desenvolver implementações de *Deep Learning*. É capaz de usar várias *threads*, de modo a que sistemas *multi-core* possam ser efetivamente utilizados. Além disso, foi projetado para ser executado em vários computadores para distribuir as cargas de trabalho.

A plataforma Google Cloud, em termos de aprendizagem automática, apresenta algumas características adequadas para aprendizagem, como por exemplo, possui baixo custo de inscrição, alto desempenho, grande escalabilidade, custos compartilhados e apenas é pago o que é utilizado. O TensorFlow será executado em várias instâncias de uma máquina virtual, onde o uso do Cloud ML Engine com o Cloud Datalab será usado para realizar previsões. Além disso, a sua integração com o Google Cloud Dataflow ajuda no pré-processamento, quanto que o Google Cloud Storage permite aceder a dados facilmente. Os principais benefícios da sua utilização podem ser bastantes interessantes, por exemplo, apresenta uma integração forte com outros serviços do Google, como o Cloud DataFlow, o Cloud Storage e o Google Datalab; utilização do HyperTune para deteção automática os modelos viáveis; a plataforma permite que a pessoa gaste mais tempo no desenvolvimento e menos no provisionamento e na monitorização dos recursos, por fim, os modelos portáteis permitem modelar os modelos ML localmente no TensorFlow e fazer download para execução local. Quanto à utilização do TensorFlow, o *google command line* permite controlar os procedimentos usando o gcloud ml-engine. Para uma implementação mais rápida dos modelos mais simples, o Google oferece uma API de previsão por meio da interface da API REST.

Quanto ao Azure Machine Learning, Chappel (2015) afirma que o processo de aprendizagem automática não é especialmente simples. Para facilitar a sua utilização pelos programadores, o Azure ML fornece diferentes componentes. O Azure ML é uma solução integrada de análise de dados, que permite que os cientistas de dados preparem dados e desenvolvam experiências escaláveis na nuvem.

A plataforma Azure Machine Learning é produzida para cientistas de dados onde oferece espaço para executar tarefas como exploração de dados, escolha de métodos além de validação de pré-processamento. Em termos gerais, a plataforma Azure Machine Learning apresenta uma maior abrangência de componentes em relação ao TensorFlow, mas para este trabalho será focado principalmente na componente de aprendizagem automática. Os principais benefícios da sua utilização podem ser vários, como por exemplo, apresenta uma excelente capacidade de realizar operações manuais para aprender os fundamentos do ML, suporta uma enorme variedade de métodos, suporta múltiplas classificações como binários, deteção, regressão e análise de texto. Umas das maiores vantagens é a enorme lista de algoritmos predefinidos que os utilizadores podem aplicar nos seus próprios dados, utilizando as linguagens de programação R e Python. Além disso, permite uma facilidade de implementação, preços acessíveis consoante a sua necessidade, alto desempenho e escalabilidade, soluções compartilhadas e apresenta uma plataforma interativa.

3.3. Critérios de avaliação

Diversas ferramentas de aprendizagem automática foram desenvolvidas no intuito de tornar a aplicação da aprendizagem automática uma tarefa menos técnica.

As ferramentas de aprendizagem automática têm características únicas, onde os seus pontos fortes apresentam vantagens em relação a outras ferramentas.

Para a avaliação de cada uma das ferramentas de aprendizagem automática é necessário delinear um conjunto de critérios para que possa ser concluída qual a mais vantajosa num determinado conjunto de dados.

Segundo Landset et al. (2015), a seleção da escolha das ferramentas baseia-se nas seguintes características: a *escalabilidade* deve ser considerada com a relação ao tamanho e à complexidade dos dados; a *velocidade* refere a rapidez com que a plataforma de processamento na qual o algoritmo está sendo executado; a *cobertura* refere-se à gama de opções contidas nas ferramentas nas diferentes classes de aprendizagem automática, bem como a variedade de implementação em cada classe; a *usabilidade* pode ser considerada em termos de configuração inicial, manutenção contínua, linguagens de programação disponíveis, interface de usuário disponível e quantidade de documentação.

Quanto aos critérios definidos para a caracterização física de cada uma das ferramentas, são definidos os seguintes critérios: a *linguagem de programação* que cada umas delas suporta, os *sistemas operativos* onde podem ser executados; a *documentação* disponibilizada, *base de dados* compatíveis, quantidade e qualidade de *algoritmos* que apresentam para uma melhor previsão e representação gráfica para concluir os resultados finais.

Para Gomes (2014), as ferramentas de aprendizagem automática também podem ser caracterizadas pelos seguintes critérios: última versão estável que apresenta menos erros, ano da *última versão* de lançamento e também termos de uso, ou seja, se é Open Source ou apresenta alguns *custos* de uso.

4. DISCUSSÃO

Com o evoluir do mundo tecnológico, a procura de ferramentas de aprendizagem automática tem crescido exponencialmente, na medida que a escolha da melhor ferramenta é um dos fatores importantes.

Atualmente, existem vários tipos de ferramentas consoante a escolha do utilizador, podendo ser *Open Source* ou com custos. Para este trabalho, a escolha recaiu sobre plataformas de *Big Data* das duas das maiores empresas tecnológicas. Como resultado, é esperado que seja criado um guia de apoio para o utilizador, para auxílio na escolha de uma ferramenta. Este trabalho pretende identificar um conjunto de características que permitirá tomar uma decisão sobre a escolha das ferramentas

consideradas. Outro tópico a desenvolver no decorrer deste trabalho será um aperfeiçoamento dos critérios de escolha das ferramentas de aprendizagem automática, pois com a utilização frequente e uma maior pesquisa, os critérios definidos inicialmente poderão ser revistos e melhorados. Além disso, não está fechada a utilização de mais ferramentas de aprendizagem automática, pois quanto maior a utilização de diferentes plataformas maior será a recolha de informação.

REFERÊNCIAS

Chappel, D. and A. (2015). Introducing Azure Machine Learning. *Microsoft Azure*, 17. Retrieved from http://download.microsoft.com/download/3/b/9/3b9fba69-8aad-4707-830f-6c70a545c389/introducing_azure_machine_learning.pdf

Chunhe, S., Chengdong, W., Xiaowei, H., Yinghong, X., & Zhen, L. (2016). Machine Learning under Big Data. *Conference on Electronic, Mechanical, Information and Management*, (Emim), 301–305.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. http://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gantz, J., & Reinsel, D. (2011). Extracting Value from Chaos State of the Universe: An Executive Summary. *IDC iView*, (June), 1–12. http://doi.org/10.1007/s10916-016-0565-7

Gomes, T. (2014). Ferramentas Open Source de Data Mining, 163. Retrieved from http://comum.rcaap.pt/handle/10400.26/14084

Keim, D., Qu, H., & Ma, K.-L. (2013). Big-Data Visualization. In IEEE Computer Graphics and Applications, 33.

Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, *2*(1), 1–36. http://doi.org/10.1186/s40537-015-0032-1

Nilsson, N. J. (2010). *Introduction to Machine Learning Second Edition*. *Machine Learning* (Vol. 56). http://doi.org/10.1016/j.neuroimage.2010.11.004

Oussous, A., Benjelloun, F. Z., Ait Lahcen, A., & Belfkih, S. (2017). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*. http://doi.org/10.1016/j.jksuci.2017.06.001

Vishnu, A., Manzano, J., Siegel, C., & Daily, J. (2017). User-transparent Distributed TensorFlow, 1–9. Retrieved from http://arxiv.org/abs/1704.04560

Wang, L. (2016). Machine Learning in Big Data. *International Journal of Advances in Applied Sciences*, *4*(4), 117–123. Retrieved from http://www.iaescore.com/journals/index.php/IJAAS/article/view/868/5800

Xing, E. P., Ho, Q., Xie, P., & Wei, D. (2016). Strategies and Principles of Distributed Machine Learning on Big Data. *Engineering*, *2*(2), 179–195. http://doi.org/10.1016/J.ENG.2016.02.008

Xue-Wen Chen, & Xiaotong Lin. (2014). Big Data Deep Learning: Challenges and Perspectives. *IEEE Access*, *2*, 514–525. http://doi.org/10.1109/ACCESS.2014.2325029