IFES - INSTITUTO FEDERAL DO ESPÍRITO SANTO
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

LUAN GRILLO SILVA

ARQUITETURA PARA APIS DE MACHINE LEARNING

Cachoeiro de Itapemirim - ES
2022

LUAN GRILLO SILVA

ARQUITETURA PARA APIS DE MACHINE LEARNING

Trabalho de Conclusão de Curso apresentado à Coordenadoria do Curso de Sistemas de Informação do Instituto Federal do Espírito Santo, Campus Cachoeiro de Itapemirim, como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: PhD Rafael Silva Guimarães

Cachoeiro de Itapemirim - ES

2022

RESUMO

Esse artigo aborda o desenvolvimento de arquiteturas distribuídas para uso em APIs de machine learning. Visando alta escalabilidade, baseia se principalmente na problemática de projetos que demandam alta quantidade de modelos computacionais de machine learning, buscando uma melhor distribuição compartilhada de todo o projeto.

Palavras-chave: Sistemas distribuídos, Machine Learning, API, Computação distribuída

ABSTRACT

This article covers the development of distributed architectures for use in machine learning APIs. Aiming at high scalability, it is mainly based on the problem of projects that demand a high amount of computer models for machine learning, given a better shared distribution of the entire project.

Keywords:

Distributed Systems, Machine Learning, API, Distributed Computing

SUMÁRIO

1 INTRODUÇÃO 5
1.1 Motivação 5
1.2 A problematica 6
2 SISTEMAS DISTRIBUIDOS 7
3 A APRENDIZAGEM DE MÁQUINA 8
3.1 O classificador 9
4 OBJEÇÕES 11
5 CONCLUSÃO 13
REFERÊNCIAS 14

1 INTRODUÇÃO

A construção de um projeto de aprendizado de máquina envolve alta complexidade, cada projeto é único, incluindo seus requisitos. A busca por metodologias que melhorem as lacunas de uma arquitetura é essencial para grandes projetos, após a fase de prototipagem, a arquitetura geral do projeto deve ser pensada com cuidado, pois resultará diretamente no custo final de implantação. Para isso, uma pesquisa com uma forma distribuída e compartilhada de recursos para o projeto pode se tornar essencial para a execução de projetos.

1.1 MOTIVAÇÃO

A motivação de pesquisa para esse tema surgiu durante a atuação em um projeto que utiliza um algoritmo de machine learning para fornecer uma predição, aplicação disponibilizada a partir de uma API RESTful. O resultado da predição deve ser entregue com baixo tempo de resposta. A criação de uma arquitetura estável e escalável torna-se indispensável para o crescimento de grande parte dos projetos.

Na fase de expansão do projeto foi constatado um aumento do custo computacional em relação à sua expansão a novos clientes. O custo computacional tornou exponencial a quantidade de clientes atendidos, gerando grandes problemas, desde a dificuldade de escala do projeto, problemas de custo e baixa eficiência.

Durante a análise foi levantado que um dos principais problemas de escala desta aplicação é durante o armazenamento de estruturas de dados na memória. Os chamados modelos, são gerados a partir de uma etapa denominada treinamento, no qual uma grande quantidade de dados são processados por algoritmos de machine learning e, o retorno desse processamento é utilizado em para determinar a predição de certos tipos de informações.

1.2 A PROBLEMATICA

O intuito é fornecer ao final deste artigo um levantamento de referencial bibliográfico para construção de uma arquitetura distribuída de que contemple os seguintes requisitos, a escalabilidade, que define um projeto com capacidade de crescimento baseado na demanda e carga de trabalho, redundância, capacidade de ser flexível e tolerante a problemas, e uma boa relação custo-benefício.

Os desafios de construção da arquitetura estão diretamente relacionados à proposta do projeto, em alguns casos, os recursos variam como processamento e memória, em alguns casos específicos, existe a necessidade de processadores gráficos para o processamento de alguns algoritmos de machine learning.

Em projetos que são compostos de inúmeros modelos a escalabilidade é afetada diretamente, pois exigem requerimentos mais altos para a execução do projeto. Visando a solucionar esse problema, a criação de uma arquitetura que distribua esses modelos de forma descentralizada e compartilhada será a solução em estudo, o que poderá aumentar a eficiência desse ecossistema.

A distribuição de recursos computacionais em aplicações que demandam baixos tempos de resposta é diretamente atrelado ao algoritmo utilizado e suas complexidades. Suas excentricidades são desde o custo computacional, uso de memória primária para execução, uso da memória primária para armazenamento dos modelos e tempo de execução.

2 SISTEMAS DISTRIBUIDOS

Com a exponencial de crescimento do poder de processamento dos computadores juntamente com a baixa do custo desses equipamentos forneceu um cenário ideal para a construção de redes extremamente complexas e descentralizadas. Segundo (STEEN; HOMBURG; TANENBAUM, 1999), o resultado da tecnologia atual é que agora se torna ainda mais possível uma arquitetura composta por inúmeros computadores em rede, de proporções inimagináveis.

Esse avanço possibilitou a conexão de todo o globo, encurtando distâncias, as redes de computadores distribuídas se tornaram um pilar essencial para a possibilidade de distribuição de uma complexa gama de serviços, hoje, fenômeno irreversível.

Portanto, os sistemas distribuídos são conjunto de computadores independentes que se apresentam ao usuário como um sistema único e coerente (STEEN; TANENBAUM, 2016).

A utilização de sistemas distribuídos nesta arquitetura será o ponto chave como objeto de estudo. Possibilitando a utilização de clusters para processamento escalável da aplicação. Será analisado o impacto no uso de servidores cache para a distribuição dos modelos de machine learning de forma descentralizada aos clusters, o que possibilitará uma maior eficiência e custo-benefício final.

3 A APRENDIZAGEM DE MÁQUINA

Desde os primórdios, a humanidade busca maneiras de aperfeiçoar suas técnicas e métodos para melhorar seus processos, seja com a descoberta do fogo para a preparação dos alimentos, a ferramentas para auxiliar o trabalho e qualidade de vida humana. A tecnologia é presente desde os primeiros resquícios de vida humana, sua notável evolução e a que proporcionou o conjunto tão completo de conhecimento e ciência, que hoje, proporciona uma melhoria essencial na qualidade de vida.

Em resumo, a informática e uso de diversas técnicas que proporcionam a automação da informação, desde uma calculadora manual, qual seus complexos sistemas de engrenagem proporcionam a utilização de funções matemáticas de forma simplificadas, a invenção do transistor, que proporcionou uma exponencial na evolução tecnológica humana.

Arbitrariamente, os computadores podem ser vistos como um livro aberto para a criatividade humana, são ınfimos as utilidades e aplicações, que dependem de grande parte da imaginação humana para a solução de problemas reais.

Como a maioria das criações, a aprendizagem de máquina nasce a partir da busca de uma solução para um problema real, mais especificamente, um simples jogo de damas. (SAMUEL, 1959), cientista da computação, pioneiro em aprendizado de máquina, construiu um jogo de damas na qual seu adversário seria o computador, entretanto, o computador não conseguiu ganhar nenhuma das partidas, Arthur decide escrever um algoritmo no qual o computador analisa as partidas anteriores e aprendia as melhores estratégias dos jogos históricos, foram feitas diversas rodadas, e a partir de um certo momento, a máquina ganhava todas as rodadas. Com isso podemos iniciar o surgimento da aprendizagem de máquina, um problema aparentemente simples, que ao final se torna um novo ramo da computação.

Segundo a definição clássica de (SAMUEL, 1959), O aprendizado de máquina é um campo de estudo no qual computadores têm a habilidade de aprender sem ter sido

explicitamente programado para tal.

Diante da imensidão de aplicações, a aprendizagem de máquina torna-se uma aplicação de alta complexidade, incluindo a arquitetura distribuída, os principais problemas envolvem a dificuldade de escala dos projetos, gerando um alto custo.

3.1 O CLASSIFICADOR

Neste tópico iremos abordar o classificador utilizado para a confecção da aplicação de machine learning utilizada como base de estudo, que no caso um dos classificadores clássicos, a floresta randômica. O algoritmo preditivo baseia-se na combinação de diversas árvores de decisão para chegar em um resultado único (BREIMAN, 2001). Uma simples analogia é de uma pessoa realizando a comparação de determinado computador, no caso a pessoa irá dispor a analisar o produto com seus concorrentes, comparando o preço, funcionalidades, qualidade, desempenho, com base nessas informações históricas, determina qual será a melhor escolha a ser feita baseado nesses dados.

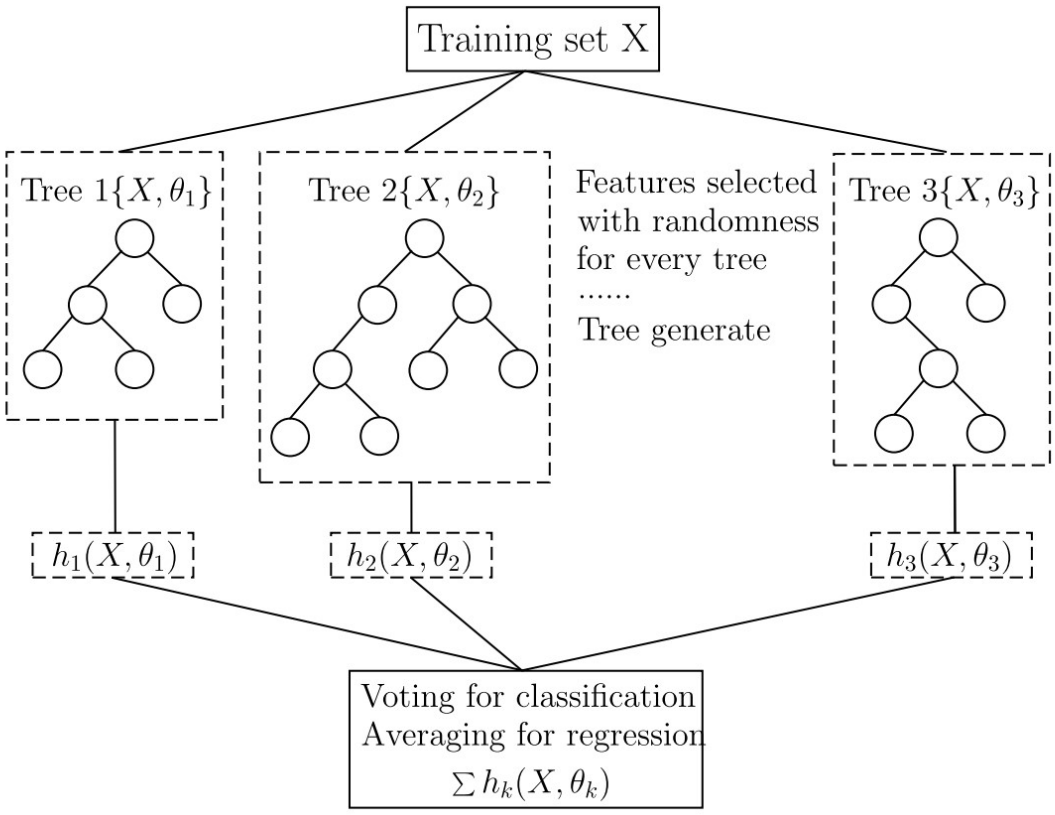

Figura 1 – Fluxograma classificatório da florestas randômica, por (ZHANG et al., 2018).

No caso do algoritmo, será feito uma análise dos dados, histórico para realização do treinamento do modelo, este modelo é composto de padrões matemáticos, ou seja, métricas gerais do comportamento histórico das informações. Com esses modelos gerados a partir de uma montanha de dados, será capaz o algoritmo realizar a predição de um novo dado, possibilitando a avaliação, por exemplo, de uma escolha boa ou ruim de determinado computador.

4 OBJEÇÕES

Durante a elaboração deste trabalho, foi constatado uma objeção, o tamanho dos modelos em relação a quantidade de dados utilizados na etapa de treinamento.

Os modelos de machine learning geralmente possuem crescimento logarítmico em relação a entrada de dados, ou seja, a partir de certa quantidade de dados processados, o espaço gasto para alocar os modelos tendem a se estabilizar em uma curva logarítmica. O mesmo fenômeno também pode ser observado durante a comparação de performance do modelo versus a quantidade de dados utilizados no treinamento.

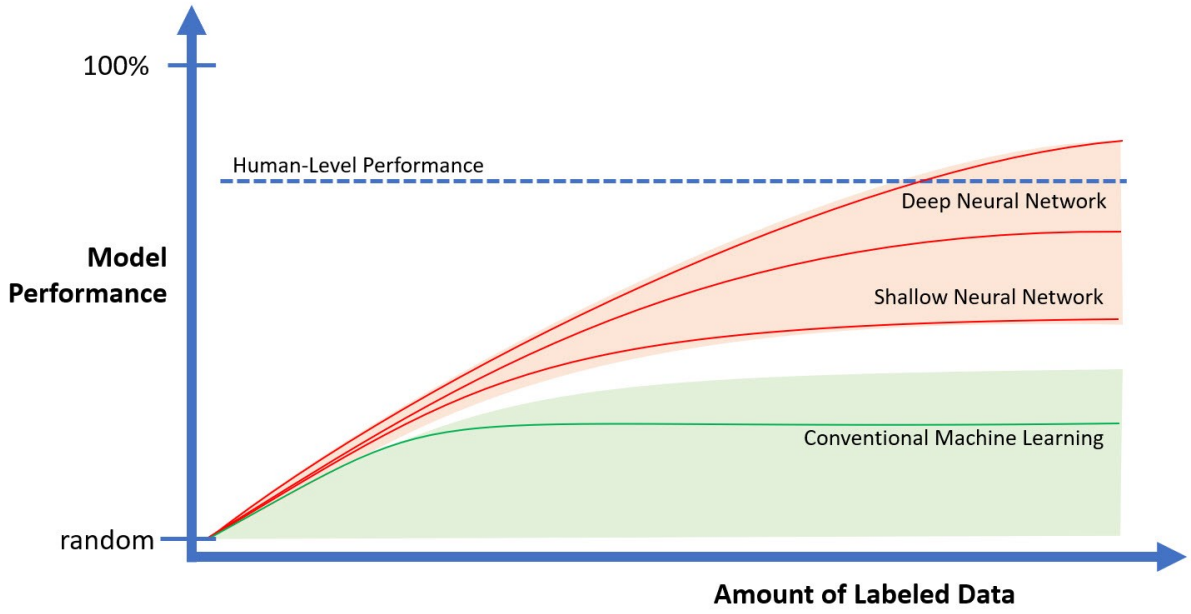

Figura 2 – Comparativo entre quantidade de dados versus performance generalista do modelo, por (HACKATHORN, 2018).

Com os estudos realizados durante a produção deste trabalho, constatamos que a ideia de resolver o problema de distribuição dos modelos de machine learning não parece ser a melhor forma de solucionar esse problema específico. Para isso, existem outros tipos de otimizações a serem feitas que reduzem o tamanho de funções desse modelo, podendo ser reduzido em espaço e contendo performance compatível ou superior a modelos mais otimizados.

A definição dos modelos como uma curva logarítmica foi crucial, pois com essa definição foi possível constatar que no início dos testes e treinamento o modelo geralmente vai se comportar de forma crescente, mas tente a se estabilizar em decorrer do aumento dos dados, por isso, torna a questão como solucionada e não tendo a necessidade desse tipo de projeto para a possível solução a curto prazo de uma modelagem com falta de otimização.

5 CONCLUSÃO

A arquitetura de uma aplicação de machine learning é composta por inúmeros desafios. Seu arranjo preserva os principais pilares dos sistemas distribuídos, como a disponibilidade, escalabilidade, confiabilidade e tolerância à falha, como descrito por van Steen, M.

Existirá grandes desafios durante a elaboração desta arquitetura, para isso, será fundamental o estudo e pesquisa de novas tecnologias, a comparação entre os principais tipos de soluções nas quais irá auxiliar para a execução desta arquitetura.

Essa revisão proporcionou uma nova perspectiva de pensamento sobre a problemática principal do trabalho, uma solução para uma optimização de modelos de machine learning em produção, entretanto, conclui se que não existe uma real preocupação como problema a longo prazo, pois o tamanho dos modelos se estabilizam em relação a quantidade de dados na etapa de treinamento.

Ainda existem temas a serem abordados dentro do tema principal do trabalho, arquitetura para API's de machine learning, o tema ficará em estudo de uma problemática principal mais condizente com a realidade e que possua boa aplicabilidade na realidade.

REFERÊNCIAS

BREIMAN, L. *Machine Learning*, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, 2001. Disponível em: <https://doi.org/10.1023/a:1010933404324>.

HACKATHORN, R. How managers should prepare for deep learning: New values. *Published in Towards Data Science*, p. 1, 08 2018.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, IBM, v. 3, n. 3, p. 210–229, jul. 1959. Disponível em: <https://doi.org/10.1147/rd.33.0210>.

STEEN, M. van; HOMBURG, P.; TANENBAUM, A. Globe: a wide area distributed system. *IEEE Concurrency*, Institute of Electrical and Electronics Engineers (IEEE), v. 7, n. 1, p. 70–78, jan. 1999. Disponível em: <https://doi.org/10.1109/4434.749137>.

STEEN, M. van; TANENBAUM, A. S. A brief introduction to distributed systems. *Computing*, Springer Science and Business Media LLC, v. 98, n. 10, p. 967–1009, ago. 2016. Disponível em: <https://doi.org/10.1007/s00607-016-0508-7>.

ZHANG, D. et al. A data-driven design for fault detection of wind turbines using random forests and XGboost. *IEEE Access*, Institute of Electrical and Electronics Engineers (IEEE), v. 6, p. 21020–21031, 2018. Disponível em: <https://doi.org/10.1109/access.2018.2818678>.