BIG DATA MINING: ANÁLISE DE DESEMPENHO DE CLASSIFICADORES

BIG DATA MINING: CLASSIFIER PERFORMANCE ANALYSIS

Ede Miranda Junior
https://orcid.org/0000-0001-5151-6742
CPF: 453.419.928-73
Centro Paula Souza – Fatec Indaiatuba/SP
ede.miranda@fatec.sp.gov.br

Orientador: Profa Dra. Maria das Graças J. M. Tomazela
https://orcid.org/0000-0002-5471-2658
CPF: 085.107.058-28
Centro Paula Souza – Fatec Indaiatuba/SP
gtomazela@fatecindaiatuba.edu.br

RESUMO: A Mineração da dados é um campo vasto que combina diferentes áreas como inteligência artificial, gerenciamento de banco de dados, visualização de dados, aprendizado de máquina, algoritmos matemáticos e técnicas estatísticas. Big data diz respeito a grandes quantidades de dados provenientes de origens variadas, como redes sociais, sensores, dispositivos, terceiros, aplicativos da web e mídia social e em uma variedade de formatos, como texto, vídeo, áudio, diagramas, imagens e combinações de dois ou mais formatos. A execução de tarefas de mineração de dados em conjuntos de dados de grande escala, Big Data, é chamada de Big Data Mining, e é considerada uma atividade complexa. Para realizar tarefas de mineração de big data, os dados são dimensionados muito além da capacidade de um único computador. A plataforma de mineração de Big Data precisa contar com computadores em cluster em plataforma de computação de alto desempenho. Esse estudo teve como objetivo realizar uma avaliação de desempenho de classificadores complexos em um ambiente local e em um ambiente na nuvem, utilizando-se de ferramentas de mineração de dados e de processamento distribuído. A Metodologia usada para essa pesquisa foi a experimental. O conjunto de dados utilizado no experimento foram referentes a alunos da Fatec dos campi de Sorocaba e Indaiatuba, a plataforma de computação em nuvem utilizada foi a Azure da Microsoft. Foram criadas 5 máquinas virtuais para a construção do cluster. Por meio desse estudo foi possível observar que houve uma queda média de 66% no tempo de processamento dos classificadores na nuvem em relação a máquina local, também foi possível estipular uma configuração de máquina virtual ideal para o conjunto de dados analisados, evitando assim desperdício de recursos.

ABSTRACT: Data mining is a vast field that combines different areas like artificial intelligence, database management, data visualization, machine learning, mathematical algorithms, and statistical techniques. Big data is about large amounts of data coming from a variety of sources such as social networks, sensors, devices, third parties, web applications and social media and in a variety of formats such as text, video, audio, diagrams, images, and combinations of two or more formats. Performing data mining tasks on large-scale data sets, Big Data, is called Big Data Mining, and is considered a complex activity. To perform big data mining tasks, data scales far beyond the capacity of a single computer. The big data mining platform needs to rely on clustered computers on high-

performance computing platform. This study aimed to perform a performance evaluation of complex classifiers in a local environment and in a cloud environment, using data mining and distributed processing tools. The methodology used for this research was experimental. The data set used in the experiment were related to Fatec students from the Sorocaba and Indaiatuba campuses, the cloud computing platform used was Microsoft's Azure. 5 virtual machines were created to build the cluster. Through this study it was possible to observe that there was an average drop of 66% in the processing time of the classifiers in the cloud in relation to the local machine, it was also possible to stipulate an ideal virtual machine configuration for the analyzed dataset, thus avoiding wasted resources.

PALAVRAS-CHAVE: Big data. Computação em nuvem. Mineração de dados. Análise de desempenho. Classificadores.

KEYWORD: Big data. Cloud computing. Data mining. Performance analysis. Classifiers.

1 INTRODUÇÃO

A mineração de dados (*Data Mining*) é um campo interdisciplinar que combina inteligência artificial, gerenciamento de banco de dados, visualização de dados, aprendizagem de máquina, algoritmos matemáticos e técnicas estatísticas (HAN e KAMBER e PEI, 2011; UNIVASO, ALE e GURLEKIAN, 2015), faz parte de um processo maior denominado descoberta de conhecimento em base de dados ou KDD, do inglês *Knowledge Discovery in Database.*

Big Data é um campo emergente de pesquisa que usa análise de dados para apoiar as decisões. Trata-se de grandes volumes de dados provenientes de várias fontes, tais como redes sociais, sensores, dispositivos, terceiros, aplicativos da *Web* e mídias sociais, e em uma variedade de formatos, como texto, vídeo, áudio, diagramas, imagens e combinações de dois ou mais formatos (SIN e MUTHU, 2015).

A realização de tarefas de mineração de dados em conjuntos de dados de larga escala, *Big Data*, tem sido denominada de *Big Data Mining* (mineração de *Big Data*). Amma (2016) apresenta a seguinte definição para o termo: o processo de extração de informações úteis de grandes conjuntos de dados ou fluxos de dados, devido ao seu volume, velocidade, variedade, validade, veracidade, valor e visibilidade é denominado como *Big Data Mining*.

Amma (2016) e Wu et al. (2013) destacam ainda que, para lidar com tarefas de mineração de dados em pequena escala um único computador desktop é suficiente. Mas para a realização de tarefas de mineração de *Big Data*, os dados são dimensionados muito além da capacidade de um único computador. Assim, a plataforma de mineração de *Big Data* depende de computadores em *cluster* com plataforma de computação de alto desempenho e ferramentas de programação distribuída como *Map Reduce.*

A partir deste contexto, o objetivo desta pesquisa foi realizar uma avaliação de desempenho de classificadores complexos em um ambiente o*n-premise* (local) e um ambiente IaaS (Infraestrutura como um Serviço), utilizando-se de ferramentas de mineração de dados e de processamento distribuído.

2 METODOLOGIA

Neste trabalho, utilizou-se a pesquisa experimental, que implica em determinar o objeto de estudo, selecionar as variáveis que podem afetá-lo, definir as formas de controle e de observação dos efeitos que a variável produz no objeto (GIL, 2007).

O conjunto de dados utilizado no experimento foram referentes a alunos da Fatec dos campus de Sorocaba e Indaiatuba, dos cursos de Análise e Desenvolvimento de Sistemas e Gestão Empresarial, esse conjunto de dados contém 144.468 instâncias, contendo os atributos: Curso; Turno; Status do Aluno (em curso, concluído etc.); Escola Pública (sim ou não); Raça; Nota do Vestibular; Idade; Ano; Semestre; Ano de Início do Curso; Semestres Cursados; Disciplina; Nota; Frequência; Campus e Conceito (Aprovado, Reprovado etc.)

Esses dados foram pré-processados e *clusterizados* localmente utilizando a ferramenta Weka, gerando 5 *clusters*. Em seguida foi criada uma tabela contendo o conjunto de dados anterior com seus respectivos *clusters* gerados. Assim o último atributo adicionado foi o atributo Cluster, totalizando 17 atributos no conjunto de dados final. Os classificadores têm como objetivo prever em qual *cluster* um determinado aluno será alocado.

A plataforma de computação em nuvem utilizada foi a Azure da Microsoft. Foram criadas 5 máquinas virtuais para a construção do *cluster*. Sendo 4 *workers*, e 1 *master*. O quadro 1 apresenta as configurações das máquinas virtuais e de uma máquina *On-Premise*.

Quadro 1: Especificações das máquinas

Máquinas	CPUs Core	RAM	CPU	SO
Master	4	16	Intel Xeon E5-2673 v3	Linux
Worker 1	4	16	Intel Xeon E5-2673 v4	Linux
Worker 2	4	16	Intel Xeon E5-2673 v4	Linux
Worker 3	4	16	Intel Xeon E5-2673 v4	Linux
Worker 4	4	16	Intel Xeon E5-2673 v4	Linux
Local	4	12	AMD Ryzen 5 2500U	Windows

Fonte: Autor

As CPUs Intel Xeon E5-2673 possuem um *Clockspeed* base entre 2,3 Ghz e 2,4 Ghz, com elevadas cargas de trabalho esse valor pode chegar a 4,0 Ghz. Enquanto as CPUs AMD Ryzen 5 2500U possuem um valor base de 2,0 Ghz, podendo chegar até 3,6 Ghz. Um valor elevado do *Clock* aponta um maior número de ciclos executados por segundo pela CPU, em GigaHertz. Em todas as 5 máquinas foram realizadas as configurações do Hadoop 3.2 e do Spark 3.1.

Utilizou-se da máquina alocada para *Master* para a instalação do Weka 3.8.5. O módulo utilizado da ferramenta foi o *KnowledgeFlow*, porque esse módulo apresenta a possibilidade de uso do Weka em uma estrutura de processamento distribuído, como Spark ou Hadoop MapReduce, conforme apresentado em Hall (2015).

2 DESENVOLVIMENTO

Para Koliopoulos et al. (2015) uma dificuldade no desenvolvimento de ferramentas de mineração de dados em grande escala é como expressar os algoritmos de forma a torná-los tão fáceis de usar quanto as ferramentas sequenciais existentes. A maioria das tentativas expõe um conjunto restrito de primitivas de baixo nível, mas geralmente tendem a ser proibitivas devido à sua natureza complexa e à incapacidade de acomodar os padrões de algoritmos de mineração de dados. Implementações mais recentes tentam fornecer interfaces de alto nível para mineração de dados e algoritmos associados que são compilados para primitivas de baixo nível. Tais desenvolvimentos tendem a exigir conhecimento do sistema distribuído subjacente, mudando efetivamente o foco da mineração de dados para a implementação de algoritmo individual.

Amma (2016) lista os principais desafios impostos ao se desenvolver um projeto de *Big Data Mining*: análise de arquitetura; avaliação de desempenho; mineração distribuída; dados de evolução do tempo; compressão dos dados; Visualização; dados que se perdem (*Big Data* Oculto).

Wu et al. (2014) apresentam os desafios da realização de mineração de *Big Data* sob uma visão conceitual da estrutura de processamento de *Big Data*, que inclui três camadas, com considerações sobre acesso e computação de dados (Camada I), privacidade de dados e conhecimento de domínio (Camada II) e algoritmos de mineração de *Big Data* (Camada III).

Visando a analisar as vantagens da utilização de ambientes de computação na nuvem para problemas que demandam grande carga de processamento, e assim, contribuir para a criação de uma Arquitetura de *Big Data* para monitoramento de desempenho de alunos

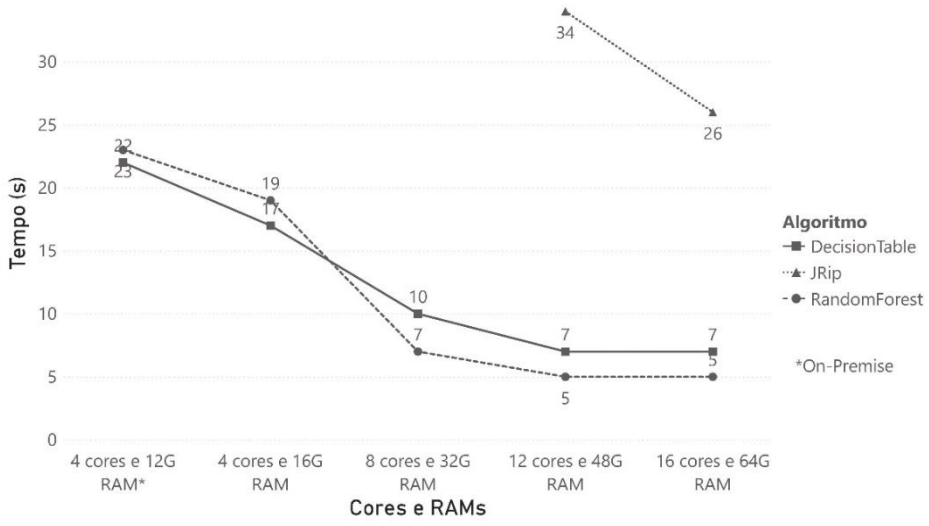
de ensino superior público foi realizado um experimento com classificadores disponíveis na ferramenta Weka. Os classificadores selecionados para o experimento foram: DecisionTable, JRip, RandomForest, SMO e MultilayerPerceptron. Foram realizadas um total de 5 rodadas de classificação para cada classificador em uma configuração de máquina diferente.

4 RESULTADOS OBTIDOS

Os resultados foram divididos em 2 gráficos para facilitar a visualização, pois os classificadores MultilayerPerceptron e SMO possuem escalas maiores do que o restante dos classificadores. O classificador JRip não foi capaz de executar a classificação nas três primeiras máquinas, devido a limitação de memória RAM, sendo assim ele foi executado somente nas duas últimas máquinas.

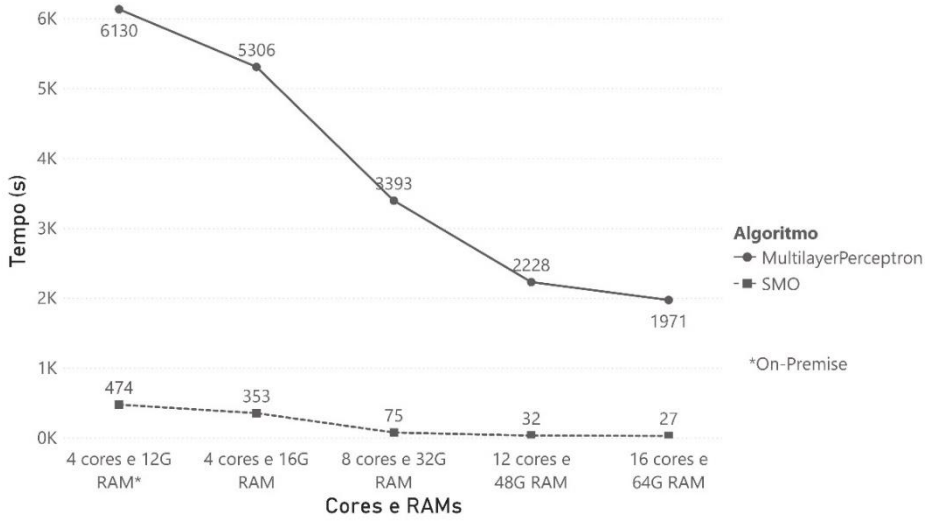
Os resultados são exibidos nas Figura 1 e Figura 2 com o primeiro gráfico apresentando o desempenho dos classificadores DecisionTable, JRip e RandomForest, enquanto o segundo apresenta o desempenho dos classificadores MultilayerPerceptron e SMO.

Figura 1: Desempenho dos Classificadores



Fonte: Autor

Figura 2: Desempenho dos Classificadores 2



Fonte: Autor

A primeira rodada de classificação foi realizada na máquina local (*On-Premise*). O tempo total de execução da primeira rodada foi de 6.649 segundos (1h51min). Em seguida foi realizado a segunda rodada de classificação, utilizando a primeira máquina *worker* na nuvem. O tempo total de execução da segunda rodada foi de 5.695 segundos (1h35min). Notou-se que houve uma pequena queda no tempo de execução de todos os classificadores em comparação com a máquina local. O classificador com maior queda foi o SMO com 25% de queda em relação a primeira rodada, e o classificador com menor queda foi o RandomForest com 17%. A média de queda de todos os classificadores foi de 19%. Mesmo possuindo a mesma quantidade de *Cores*, essa pequena queda é explicada devido a maior quantidade de memória RAM disponível e pela CPU com maior poder de processamento na máquina *worker* da nuvem.

Na terceira rodada de classificação foi utilizada a primeira e a segunda máquina *worker* na nuvem. As duas máquinas puderam executar as tarefas em 3.485 segundos (58min). Nessa rodada houve uma expressiva queda no tempo de execução da classificação. O classificador que sofreu a maior queda foi o SMO novamente, com 79% de queda, enquanto a menor queda foi do MultilayerPerceptron, com 36%. A média de queda de todos os classificadores foi de 55%. Essa foi a maior queda registrada no experimento.

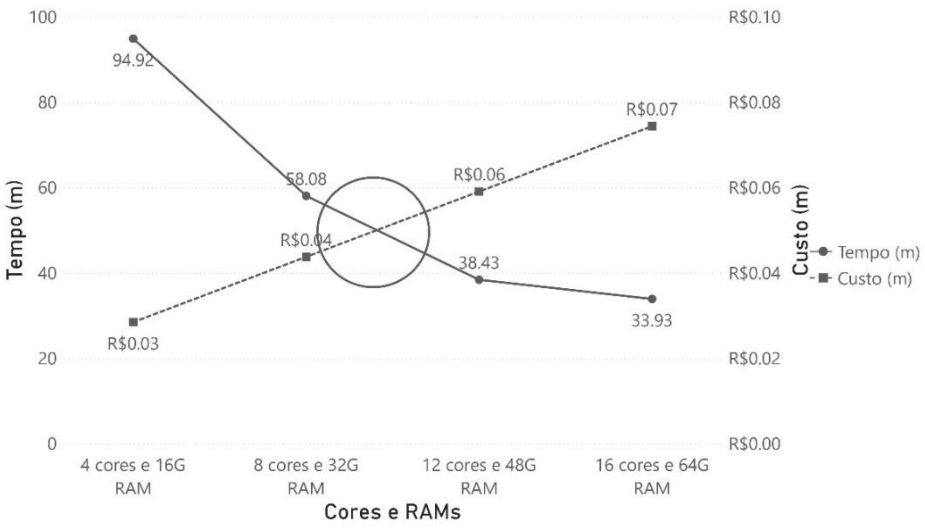
A quarta rodada de classificação foi realizada utilizando três máquinas *workers* na nuvem. As três máquinas puderam realizar todas as tarefas em 2.272 segundos (38min), dessa vez o classificador JRip também foi executado, devido a quantidade maior de RAM

disponível. Nessa quarta rodada a queda média foi de 37%, uma queda menor que a rodada anterior. Novamente o classificador com maior queda foi o SMO com 57%, e o com menor queda foi o RandomForest, com 28%.

A quinta e última rodada foi realizada utilizando todas as quatro máquinas *workers* na nuvem. As quatro máquinas puderam realizar todas as tarefas em 2.037 segundos (34min). Nessa última rodada a queda média foi de apenas 12%, a menor queda registrada no experimento. Não houve alteração no tempo de execução para os classificadores DecisionTable e RandonForest, expondo que esse é o limite para esses classificadores para o tamanho desse conjunto de dados. O classificador que registrou a maior queda foi o JRip, com 23%.

A Figura 3 exibe o gráfico contendo a relação entre custo e tempo de execução por quantidade de cores e RAM na nuvem, ambos em minutos.

Figura 3: Relação Custo e Tempo em minutos



Fonte: Autor

Essa relação revela que a quantidade de *cores* e RAM ideal para esse experimento e para esse conjunto de dados específico, seria de aproximadamente 10 *cores* e 40G RAM. É importante analisar esse tipo de relação para que não haja gastos desnecessários com recursos que não farão grande diferença no tempo de execução das tarefas dos classificadores.

5 CONSIDERAÇÕES FINAIS

Por meio das análises de desempenho realizadas sobre os classificadores foi possível concluir que os algoritmos sofreram uma queda total média de 66% no tempo de processamento das tarefas de classificação, e que, o algoritmo mais sensível na adição de novas máquinas *workers* foi o SMO, com uma queda total de 94%, enquanto os algoritmos menos sensíveis foram o JRip e o MultilayerPerceptron, com uma queda de 23% e 68% respectivamente. Isso indica que seria necessário a adição de mais máquinas *workers* para se visualizar uma queda maior no tempo de processamento desses classificadores menos sensíveis. Também foi possível concluir que nem sempre uma grande quantidade de recursos de hardware disponível será vantajosa para o processamento das tarefas dos classificadores, pois poderá haver desperdício financeiro nos casos em que não houver grandes quedas no tempo de processamento. Sendo assim análises como essas são indispensáveis para a avaliação da necessidade de implementação de uma arquitetura de Big Data eficiente.

REFERÊNCIAS

AMMA, N. Big Data Mining. **Effective Big Data Management and Opportunities for Implementation**. 2016.

GIL, A. C. **Métodos e técnicas de pesquisa social**. São Paulo: Atlas, 2007.

HALL, Mark. Weka and Spark. **Mark Hall on Data Mining & Weka**, 2015. Disponível em: <http://markahall.blogspot.com/2015/03/weka-and-spark.html>. Acesso em: 15/08/2021.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. São Francisco, EUA: Morgan Kaufmann Publishers, 2011.

KOLIOPOULOS, A.; YIAPANIS, P.; TEKINER, F.; NENADIC, G.; KEANE, J. A Parallel Distributed Weka Framework for Big Data Mining Using Spark. **IEEE International Congress on Big Data**. 2015.

SIN, K.; MUTHU, L. Application of Big Data in Education Data Mining and Learning Analytics-A Literature Review. **SOCO.** 2015.

UNIVASO, P.; ALE, J. M.; GURLEKIAN, J. A. Data mining applied to forensic speaker identification. **IEEE**, v. 13, n. 4, p. 1098–1111, 2015.

WU, X.; ZHU, X.; WU, G.; DING, W. Data mining with big data. **IEEE Transactions on Knowledge and Data Engineering**. vol. 26, no. 1, pp. 97-107. 2014.