BIG DATA, MINERAÇÃO DE DADOS E APRENDIZAGEM DE MÁQUINA: FORMAS DE EXTRAIR INFORMAÇÃO EM GRANDES VOLUMES DE DADOS

BIG DATA, DATA MINING AND MACHINE LEARNING: WAYS TO EXTRACT INFORMATION IN LARGE VOLUMES OF DATA

André Bessa da Silva[1]

RESUMO

Devido aos avanços tecnológicos das últimas duas décadas, principalmente no que se refere à grande produção de dados (Big Data), fez-se necessário a aplicação de métodos que pudessem extrair informações desse volume de dados. Pensando nisso, este artigo apresenta por meio de uma revisão de literatura, os conceitos centrais da Aprendizagem de Máquina, Mineração de Dados e Big Data. A Metodologia utilizada na pesquisa foi à revisão teórica, dos principais autores (MONARD, 2003; SOCZEK; ORLOVSKI, 2014) e produções dos últimos 25 anos na área da Aprendizagem de Máquina. As vantagens esperadas na aplicação do Aprendizado de Máquina e da Mineração de Dados na obtenção de informações em bases de dados de grande volume, será realizada de maneira automática.

Palavras-chave: Aprendizado de Máquina. Mineração de dados. Inteligência Artificial. Big Data.

ABSTRACT

Due to the technological advances of the last two decades, especially with regard to the large data production (Big Data), it was necessary to apply methods that could extract information from this volume of data. Thinking about this, this article presents through a literature review, the central concepts of Machine Learning, Data Mining and Big Data. The methodology used in there search was the theoretical revision of the main authors (MONARD, 2003; SOCZEK; ORLOVSKI, 2014) and productions of the last 25 years in the area of Machine Learning. The expected advantages in the

[1] Mestrando em Pesquisa Operacional e Inteligência Computacional pela Universidade Candido Mendes e Professor do Curso do Bacharelado em Sistemas de Informações da Multivix – Cachoeiro de Itapemirim/ES. E-mail: andrebessax@gmail.com.

application of Machine Learning and Data Mining in obtaining information in large databases will be performed automatically.
Keywords: Machine Learning. Data Mining. Artificial Intelligence. Big Data.

1 INTRODUÇÃO

Devido ao intenso avanço tecnológico das últimas duas décadas, no que tange ao grande volume de dados que se é produzido e as diversas maneiras criadas para gerarmos e extrairmos informação, o desenvolvimento de algoritmos, os métodos e as aplicações conseguem tratar esse avanço de maneira eficaz e se faz cada vez mais necessário. E, é neste cenário que, um dos ramos da Inteligência Artificial, a Aprendizagem de Máquina, tem crescido e se desenvolvido em conjunto com a mineração de dados, possibilitando a criação de soluções tecnológicas inovadoras.

Neste panorama, a aplicação de Aprendizagem de Máquina (Machine Learning) sobre este grande volume de dados (Big Data) pode oferecer soluções e propor ferramentas metodológicas a serem aplicados nos dados a fim de que possam ser gerados informações, que poderão ser transmitidas e aprendidas a sistemas que implementem algum nível relevante de inteligência artificial.

Desde a década de 2000 produzimos muitos dados, o que faz existir uma gama de maneiras de processá-los, tabulá-los e inferir conhecimento sobre esses resultados e, por conseguinte, no que se refere à inovação, a criação de máquinas cada vez mais inteligentes seria uma das maneiras produtivas para que se possa aprender com essas informações.

O objetivo do presente estudo é apresentar os conceitos centrais relacionados à Aprendizagem de Máquina e a Mineração de Dados, bem como sobre aqueles que são de suma importância para obtenção de informação no Big Data, tendo como metodologia de pesquisa a revisão de literatura pertinente aos temas abordados.

2 BIG DATA

Conceitualmente, definimos Big Data como sendo o termo mais comum para a grande massa de dados, estruturados ou não, que a sociedade como um todo vem produzindo. Esses são oriundos das mais diversas fontes, desde as mais tradicionais como livros, revistas, periódicos, sensores até dados de sensoriamento, geolocalização e mídias digitais, sobretudo, a internet e as redes sociais.

De acordo com Chen et al. (2012)o termo Big data é um conceito um tanto abstrato que nos permite definir a grande massa de dados produzidas de fontes e formatos diverso.

Figura 1 - Big Data “Harvest Value from Petabytes”

Fonte: http://www.dtiers.com/big-data-2

Mesmo em rápida evolução o Big data foi sendo caracterizado em cinco valores fundamentais (5vs) (MAÇADA et al., 2015):

- Volume - Grande quantidade informação a ser processada;
- Variedade – Os diferentes tipos de dados analisados;
- Velocidade – No que se refere ao tempo hábil para recuperar e processar a informação;
- Valor – o grau de importância deste dado para compor uma informação; e
- Veracidade – O quão confiável é o dado.

O Big Data é um conceito usado atualmente em larga escala por todo tipo de organização, tanto na esfera privada quanto na esfera governamental, pois a análise correta sobre esses dados permite que os gestores e administradores possam tomar decisões mais acertadas e (ou) possam corrigir decisões tomadas anteriormente de forma equivocada.

Essa massa volumosa de dados possui tamanha complexidade que é praticamente impossível realizar operações sobre elas como, por exemplo, a ordenação, a sumarização e as consultas. Entre outras formas eficientes, utilizar dos Gerenciadores de Banco de dados (SGBD) que possuímos atualmente, com a quantidade de informação que é gerada na ordem de terabytes e (ou) pentabytes, vem a promover novos desafios no que diz respeito às maneiras de armazenarmos e processarmos todo esse volume de dados para extração de informações relevantes para geração de conhecimento (VIEIRA et al., 2012).

Mas, como coletar informação neste volume absurdo de dados? Haja vista que dados não faltam, desde as bases de dados tradicionais (transacionais) a informações de que estamos consumindo ou interesses que estamos compartilhando, onde estamos no mapa, quem estamos conhecendo, o que estamos assistindo ou lendo (IAMARINO, 2014), ou seja, nunca se produziu tanta informação e de maneiras tão diversa.

Como solução para a extração de informações dessas fontes de dados tão diversa e rica existe uma série de técnicas, ferramentas e estratégias. Contudo, neste estudo o foco estará voltado para a:

- Mineração de dados: trata-se de um conjunto de processos e técnicas computacionais que procuram por padrões em um conjunto de dados. Este processo de extração de informação utiliza técnicas de inteligência artificial, bancos de dados relacionais, aprendizado de máquina e estatística e
- Aprendizado de máquina: faz menção a parte da inteligência artificial que objetiva treinar o computador a reconhecer padrões de informações e, com base nisso, “inferir” conhecimento, ou seja, aprender.

3 MINERAÇÃO DE DADOS (DATA MINING)

Como a busca por informações é de suma importância em qualquer que seja a área de estudo e conhecimento, torna-se importante ter ferramentas e mecanismos que propicie o auxílio à tomada de decisão, bem como a busca de informações seguras e confiáveis.

Graças aos avanços tecnológicos que viemos experimentando desde o início da década de 2000, no que tange as Implementações de a Base de Dados, Inteligência Computacional e Redes de Comunicação, nunca se fez tão necessário a aplicação de técnicas como a Mineração de Dados como uma ferramenta de busca de informações úteis que possibilite a tomada de decisões em condições de certeza limitada (SOCZEK; ORLOVSKI, 2014).

A Mineração de Dados é o ramo da disciplina Banco de Dados, que utiliza técnicas e algoritmos para extrair informações relevantes de uma base de dados densamente povoadas. Portanto, nada mais é que uma das técnicas para obtermos conhecimento em base de dados, permitindo que possamos descobrir conhecimento que esteja implícito no agrupamento de dados (CARDOSO; MACHADO, 2008).

A mineração de informações em bases de dados é uma área de estudo relativamente nova, surgidas na década de 1980, objetivando a obtenção de informações relevantes aos negócios, utilizando computadores e que coletem informações embasadas nos dados contidos nas gigantescas bases de dados que já haviam na época.

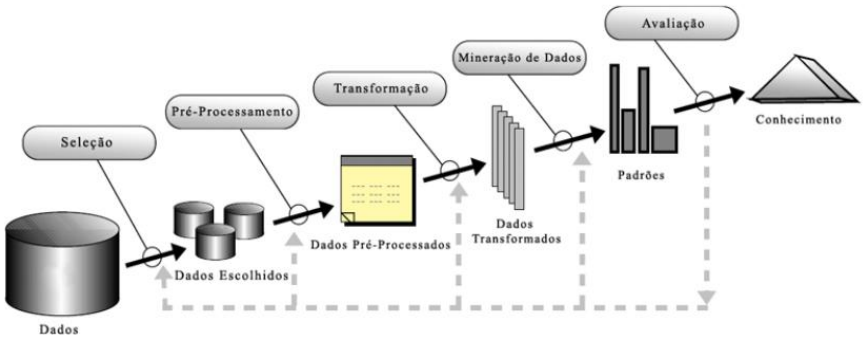
O processo mais tradicional de coleta de dados consiste basicamente no processamento de informação usando técnicas manuais de processamento de informação por especialistas que geram uma série de relatórios que deverão ser analisados e interpretados pelos tomadores de decisão. Em muitos casos, este processo se torna impraticável devido ao volume de dados e, é nesse momento que a mineração de dados torna-se uma alternativa de solução a esse problema de sobrecarga de dados. (CAMILO; CARLOS, 2009). Todo este processo é denominado *Knowledge Discovery in Database* (KDD).

É importante salientar que, a mineração de dados é parte de um processo maior de descoberta de conhecimento em bases de dados ou KDD, onde há na literatura quem considere Data Mining e KDD sinônimos (FAYYAD et. al., 1996).

O Data Mining é capaz de maneira automatizada encontrar informações escondidas em grandes massas de dados, possibilitando uma agilidade maior na tomada de decisão baseadas nas informações extraídas (CARDOSO; MACHADO, 2008).

Como explicitado na Figura 2, no processo de Mineração de dados, ocorre à seleção de um conjunto de dados provenientes de um banco de dados demasiadamente grande, como um banco da Receita Federal com milhões de dados de contribuintes. Sobre esse conjunto de dados é aplicado um pré-processamento que objetiva tratar os dados antes de se aplicar os algoritmos que irão minerar as informações. Deste processo serão definidos padrões e características a serem analisados que irão se tornar informação relevante para os tomadores de decisão.

Figura 2 - Processo de Mineração de Dados com base no KDD



Fonte: Camilo e Carlos (s.d.).

O principal objetivo da Mineração de Dados é a transformação dos dados armazenados em conhecimento, devendo ser apresentado de maneira formal, com regras preestabelecidas, para que haja um relacionamento entre os dados, haja vista que é possível sim, extrair informações relevantes de em bases de dados sem nenhum tipo de ferramenta ou técnica de tratativa. Conquanto, há também uma gama de informações que não ficam explícitas no volume de dados em tratamento, devido às interrelações dos dados, se fazendo necessária a aplicação de algumas técnicas especiais.

A mineração de dados é muito utilizada por empresas que possuem um grande volume de dados. Aplicar a mineração pode ser muito útil para que a mesmas possam conhecer e planejar melhor suas operações e conhecer melhor seu público, compreendendo, portanto, suas preferências.

A título de exemplo, pode-se citar uma rede de supermercados que querem saber a quantidade de produtos “X” e produtos “Y” que são vendidos em conjunto para melhorar o posicionamento destes nas prateleiras, ou ainda empresas como Google, que utilizam de mineração de dados em bases não convencionais que armazenam as preferências de buscas de seus usuários, para que assim possam sugerir termos relacionados às buscas ou a exibição de anúncios direcionados a um público especifico com base nos seus históricos de pesquisa.

Importante ressaltar que, a mineração de dados no Big Data tem como finalidade ajudar as pessoas e organizações a extraírem informações relevantes para uma tomada de decisão que será útil ao negócio. Por isso que um dos maiores benefícios da mineração de dados é a criação de uma inteligência de negócio sobre determinada área, permitindo maior competitividade, propiciando o desenvolvimento de melhores produtos e serviços as empresa e (ou) organização (SANTANA, 2017).

4 APRENDIZADO DE MÁQUINA

A Aprendizagem de Máquina é uma das áreas da Inteligência Artificial que de maneira considerável vem ganhando espaço no mercado, graças aos avanços dos estudos e tecnologias de Internet das Coisas (IoT) e do Big Data. Com este aprendizado, os computadores podem identificar padrões entre os dados analisados e, por meio da aplicação de algoritmos especiais, serem treinados a aprender sozinhos, a fim de executar uma tarefa (SAP, 2017).

Softwares capazes de aprender com a experiência e informações inerentes a um grande volume de dados nos ajudam a definir o aprendizado de máquina (MITCHELL, 1997). O objetivo é a criação de técnicas computacionais que visam o aprendizado e a construção de sistemas inteligentes com a capacidade de adquirir conhecimento de

forma automática. Um sistema aprendiz é aquele que consegue tomar decisões com base em soluções bem-sucedidas aplicadas a problemas anteriores (MONARD; BARANAUSKAS, 2003).

De acordo com (LUIZ e et. al., 2003) dentre os algoritmos de Aprendizado de Máquina alguns se inspiram em sistemas biológicos como: Algoritmos Genéticos; Redes Neurais; Raciocínio Baseado em Casos; Árvores de Decisão e as Teorias Estatísticas. Uma das maneiras de se aprender é observando as propriedades das coisas e eventos com base nos valores destas propriedades e experiências para deduzir informações e denominamos a capacidade de deduzir como inferência. Na aprendizagem de máquina também é utilizado inferências sobre valores de dados, a fim de obtermos conclusões genéricas sobre um conjunto particular de dados (MONARD; BARANAUSKAS, 2003).

O aprendizado indutivo se dá através da aplicação de raciocínio lógico nos dados fornecidos por um processo externo ao processo de aprendizado. A tecnologia de Aprendizado de Máquina, que é indutiva, categoriza o aprendizado em: Supervisionada e Não-supervisionada.

Na categoria supervisionada existe o fator humano para a entrada e a saída de dados para os algoritmos de aprendizagem. As informações obtidas durante o processo de aprendizado serão aplicadas a um novo conjunto de dados, ou seja, a máquina é treinada a reconhecer os resultados satisfatórios. Na não-supervisionada, os algoritmos não recebem nenhuma entrada de dados previamente, é utilizado neste caso uma denominada aprendizagem profunda (SAP, 2017).

No Aprendizado não-supervisionado é realizada uma análise nos exemplos de dados fornecidos, a fim de se determinar o agrupamento entre eles. Com este agrupamento estabelecido, é feita uma análise posterior para identificar o que cada agrupamento significa no contexto da solução do problema (MONARD; BARANAUSKAS, 2003).

Ainda no processo de aprendizado de máquina temos o Aprendizado por recompensa, no qual, o sistema é compensado de acordo com desempenho apresentado. (LUIZ et al., 2003).

4.1. Aplicabilidade do Aprendizado de Máquina

Atualmente é muito utilizado o aprendizado de máquina para avaliação de preferências de buscas online dos usuários. Por exemplo, após a realização de uma pesquisa web sobre algum tema de interesse, nos dias seguintes os perfis online serão bombardeados com anúncios relacionados às pesquisas realizadas.

Mas, além disso, a aprendizagem de máquina pode ser utilizada para ensinar os sistemas a detectar fraudes em sistemas bancários, buscar falhas em sistemas de redes e realizar avaliações de rotinas de manutenção, bem como a aplicabilidade do aprendizado de máquina, que é muito grande.

Gigantes da tecnologia a Google e Facebook investem massivamente no desenvolvimento de tecnologias de Aprendizado de máquina e vem aplicando com sucesso não só aos seus motores de busca, mas também em pesquisas em diversas áreas. Recentemente, a Google pela sua divisão de Inteligência Artificial Google Brain, estuda a possibilidade de aliar aprendizagem de máquina e realidade aumentada para a detecção de células cancerígenas (SATURNO, 2018).

4.2. Mineração de Dados e Aprendizado de Máquina

Segundo Carvalho e Dallagasa (2014), o processo KDD compreende uma série de disciplinas como: estatística, banco de dados, inteligência artificial e aprendizado de máquina, sendo o KDD criado a partir dos conceitos destas áreas. E a aprendizagem de máquina seria a automação do processo de aprendizagem.

Sendo assim, usando algoritmos de aprendizado de máquina, são criados padrões de generalizações com base nos dados minerados para analise, possibilitando que os mesmos sejam agrupados e, por fim, seja criadas regras de associações sobre eles, a fim de inferir conhecimento.

Todavia, cabe salientar que, a Aprendizado de máquina e Mineração de dados não são a mesma coisa, pois cada uma possui suas particularidades e objetivos. Porém,

são disciplinas complementares, quando se parte das primícias que a intenção de ambos é extrair conhecimento de maneira automatizada e otimizada.

No Quadro 1 são exemplificadas as principais diferenças entre os dois temas apresentados neste artigo.

Quadro 1 - Diferença entre Data mining e Machine learning

	Mineração de Dados	**Aprendizado de Máquina**
Definição	Processo de extração de informação de um conjunto de dados e transformação de uma estrutura entendível para posterior uso.	Tem como objetivo a construção e estudo de sistemas que podem aprender com os dados.
Foco	Tem o foco na descoberta de propriedades desconhecidas dos nos dados.	Tem foco na predição, baseado em características conhecidas e aprendidas pelos dados de treinamento.
Tamanho da Base de Dados	É um processo automático ou semi-automático para performar em bases com grandes quantidades de dados.	É geralmente performada em bases de dados pequenas para o aumento da acurácia.
Tipos	Regras de Associação, Classificação, Clustering (Agrupamento), Padrões Sequenciais, Sequência de Similaridade.	Supervisionado, Não-Supervisionado, Reforço.
Relacionamento	A Mineração de Dados usa diversas técnicas provenientes de Aprendizado de Máquina, mas com objetivos distintos.	O Aprendizado de Máquina também usa técnicas de mineração de dados como “Aprendizado Não-Supervisionado” ou como “Passo de Pré-Processamento” para melhoria do modelo de aprendizado.
Aplicações	Previsão, Classificação, Associação, Clustering (Agrupamento), Geração de Sequências.	Automação de Controle de Acesso de Funcionários, Proteção da Fauna, Predição de tempo de espera em salas de emergência, Identificação de falha cardíaca.

Fonte: Clésio (2015).

Conforme apresentado no Quadro 1, as características de cada um dos temas ficam evidenciadas e é possível visualizar que em alguns pontos estes temas se complementam, como no relacionamento onde o Aprendizado de máquina é trabalhado com princípios de mineração de dados.

Portanto, o objetivo central é que, conhecendo cada uma das técnicas apresentadas e suas particularidades, poderá utilizá-las em conjunto da melhor forma para o processo de extração de conhecimento de base de dados heterogênea, como o Big Data, de maneira eficaz e automatizada.

5 CONSIDERAÇÕES FINAIS

O aludido tema vem assumindo cada vez maior a importância no que tange ao volume de dados que produzimos que são oriundos de diversas fontes o qual pessoas e organizações podem extrair e gerar conhecimento sobre o negócio que será útil a sua operação.

Sendo assim, utilizando as técnicas e algoritmos de mineração de dados, torna-se possível atuar sobre o Big Data para tratar esse conjunto de dados, pois as maneiras tradicionais de armazenagem e classificação não são suficientes para trabalhar sobre esse volume de formatos tão variados.

Neste processo de mineração, para se poder encontrar um padrão nos dados de maneira automática e usando ferramentas de disciplinas diversas como inteligência artificial, estatística entre outras, a melhor opção é usar as técnicas de aprendizado de máquina que, quando combinado com os algoritmos de mineração podem permitir que se gere informação e conhecimento relevantes para os tomadores de decisão.

6 REFERÊNCIAS

CAMILO, Cássio Oliveira; CARLOS, João. **Mineração de Dados:** Conceitos, Tarefas, Métodos e Ferramentas. RT-INF_001-09. p. 29, 2009. Instituto de Informática. Universidade Federal de Goiás. Goiás-GO.

CARDOSO, Olinda Nogueira Paes; MACHADO, Rosa Teresa Moreira. Knowledge management using data mining: a case studyofthe Federal Universityof Lavras. **Revista de Administração Pública,** v. 42, n. 3, p. 495–528, Jun 2008.

CARVALHO, Deborah Ribeiro; DALLAGASSA, Marcelo Rosano. Mineração de dados: aplicações, ferramentas, tipos de aprendizado e outros subtemas. **AtoZ:** novas práticas em informação e conhecimento, v. 3, n. 2, p. 82–86, 31 Dez 2014.

CHEN, Hsinchun; CHIANG, Roger H L; STOREY, Veda C. Business intelligence and analytics: from big data to big impact. **MIS Quarterly,** v. 36, n. 4, p. 24, 2012.

CLÉSIO, Flávio. Diferença entre Data Mining (Mineração de Dados) e Machine Learning (Aprendizado de Máquina). Data Mining / Machine Learning / Data Analysis. [S.l: s.n.]. Disponível em: <https://mineracaodedados.wordpress.com/2015/01/06/diferenca-entre-data-mining-mineracao-de-dados-e-machine-learning-aprendizado-de-maquina/>. Acesso em: 2 mai 2018. 6 Jan 2015

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining toKnowledge Discovery in Databases. p. 18, 1996. AI Magazine.

IAMARINO, Atíla. BIG DATA | Nerdologia - YouTube. Disponível em: <https://www.youtube.com/watch?v=hEFFCKxYbKM>. Acesso em: 21 abr 2018.

LUIZ, Giampaolo, OSHIRO, Rodrigo Mithuhiro, NETTO, Antônio Valério, CARVALHO, André Ponce de L. F. de, OLIVEIRA, Maria Cristina F. de, **Técnicas de Aprendizado de Máquina para análise de imagens oftalmológicas.** [S.l: s.n.], 2003. Instituto de Matemáticas e Computação –ICMC – Universidade de São Paulo - USP. ResearchGate.

MAÇADA, Antônio C. G.; BRINKHUES, Rafael A.; JÚNIOR, José C. F**. Big data e as capacidades de gestão da informação**. Disponível em: <http://www.comciencia.br/comciencia/handler.php?section=8&edicao=115&id=1388&tipo=1>. Acesso em: 26 abr 2018.

MITCHELL, Tom M. Does Machine Learning ReallyWork? **AI Magazine**, v. 18, n. 3, p. 11, 15 Set 1997.

MONARD, Maria Carolina, BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. sistemas inteligentes fundamentos e aplicações. 1 ed. Barueri-SP: Manole Ltda, 2003. p. 89--114.

O que é Machinelearning ou Aprendizagem de Máquina? Disponível em: <https://news.sap.com/brazil/2017/10/11/o-que-e-machine-learning-ou-aprendizagem-de-maquina/>. Acesso em: 20 abr 2018.

SANTANA, Felipe. Afinal, o que é Big Data e Mineração de dados? - Aprenda definitivamente. Aprenda Data Science e Alavanque a sua Carreira. [S.l: s.n.]. Disponível em: <http://minerandodados.com.br/index.php/2017/02/08/oque-big-data-mineracao-de-dados/>. Acesso em: 1 mai 2018. 8 Fev 2017

SATURNO, Ares. Tecnologia da Google combina aprendizado de máquina e RA para detectar câncer - Saúde. Disponível em: <https://canaltech.com.br/saude/tecnologia-da-google-combina-aprendizado-de-maquina-e-ra-para-detectar-cancer-111978/>. Acesso em: 21 abr 2018.

SOCZEK, Felipe Cebulski; ORLOVSKI, Regiane. Mineração de Dados: Conceitos e aplicação de algoritmos em uma Base de Dados na área da saúde p. 25, 2014. Semana Acadêmica – Revista Científica.

VIEIRA, Marcos Rodrigues; MAIMONE, Josiel; VIEBRANTZ, Álvaro Fellipe Mendes. Bancos de Dados NoSQL: Conceitos, Ferramentas, Linguagens e Estudos de Casos no Contexto de Big Data. p. 30, 2012. Simpósio Brasileiro de Bancos de Dados - SBBD 2012. São Paulo – SP.