APRENDIZADO DE MÁQUINA: BREVE INTRODUÇÃO E APLICAÇÕES[1]

Ricardo Cerri[2]
André Carlos Ponce de Leon Ferreira de Carvalho[3]

RESUMO

Neste artigo são descritos alguns problemas complexos do mundo real, enfrentados na indústria e academia, para os quais são utilizados métodos de Aprendizado de Máquina. Também são apresentados métodos de classificação utilizados em diversos ramos de pesquisa, como, por exemplo, a Bioinformática. Além disso, foi feita uma revisão bibliográfica de métodos de Aprendizado de Máquina na agropecuária. Finalmente, são apresentados alguns grupos de pesquisa da Embrapa, da academia e da indústria que utilizam Aprendizado de Máquina.

Termos para indexação: agropecuária, big data, classificação de dados, inteligência artificial.

MACHINE LEARNING: BRIEF INTRODUCTION AND APPLICATIONS

ABSTRACT

In this article, some complex real world problems are described, which are faced by industry and academy, for which Machine Learning methods are used. This study also presents classification methods used in many research areas, such as Bioinformatics. In addition, a literature review of Machine Learning methods applied to agriculture and livestock was performed. Finally, the study presents some research groups of Embrapa, the academy and the industry that use Machine Learning method.

Index terms: agriculture and livestock, big data, data classification, artificial intelligence.

INTRODUÇÃO

A Inteligência Artificial (IA) tem sido cada vez mais aplicada para a resolução de problemas em diversos setores da economia e da pesquisa científica e tecnológica, também com várias aplicações bem-sucedidas na gestão pública e atividades que levem a benefícios sociais. Dentro da IA, uma área de estudo que merece destaque especial é a de Aprendizado de Máquina (AM).

[1] Trabalho derivado de pesquisas e estudos no treinamento doutoral do primeiro autor no Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP).

[2] Graduado em Ciência da Computação, doutor em Ciências da Computação, professor adjunto do Departamento de Computação da Universidade Federal de São Carlos, São Carlos, SP. cerri@dc.ufscar.br

[3] Graduado em Ciência da Computação, Ph.D. em Engenharia Eletrônica, professor titular do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, São Carlos, SP. andre@icmc.usp.br

Aprendizado de Máquina é uma área de pesquisa da Inteligência Artificial que visa ao desenvolvimento de programas de computador com a capacidade de aprender a executar uma dada tarefa com sua própria experiência (FACELI et al., 2011). Isso leva ao desenho de programas capazes de aprender por si sós, utilizando-se um conjunto de dados que representam experiências passadas. Trata-se de uma área de pesquisa multidisciplinar que engloba inteligência artificial, probabilidade e estatística, teoria da complexidade computacional, teoria da informação, filosofia, psicologia, neurobiologia, entre outros. Exemplos de tarefas de Aprendizado de Máquina são: classificação e agrupamento de dados, e previsão de séries temporais.

Como exemplo básico de AM, pode-se citar um programa de computador que deve executar uma tarefa simples, como distinguir entre três variedades diferentes de flor de uma mesma espécie. Em vez de codificar um programa utilizando todo o conhecimento acerca das variedades da flor em questão, características botânicas das três flores são apresentadas a um programa que implementa um algoritmo de AM, que, por meio de um processo de treinamento, vai aprender a caracterizar uma flor baseado em suas características. Assim como os seres humanos aprendem a diferenciar as variedades de flores observando suas características, o programa de AM também aprenderá a tarefa por meio dessas características.

Dentro do AM, existem tarefas descritivas e tarefas preditivas. Em tarefas descritivas, busca-se o desenvolvimento de algoritmos que descreverão os dados. Entre as tarefas descritivas, uma das principais é o agrupamento de dados (FORGY, 1965; HORTA; CAMPELLO, 2012; KOHONEN, 1990), que busca separar os dados de maneira que dados semelhantes fiquem em um mesmo grupo. De maneira geral, buscam-se grupos cujas distâncias entre seus membros sejam minimizadas, maximizando, ao mesmo tempo, a distância entre os grupos. Assim, as distâncias entre os dados de um mesmo grupo devem ser menores do que as distâncias entre dados em grupos diferentes. Um exemplo da aplicação de agrupamento de dados é o agrupamento de textos. Nesse caso, o algoritmo procura agrupar textos que abordem o mesmo assunto e separar em grupos diferentes os textos que abordam assuntos diferentes. Há aplicações de agrupamento de dados na Biologia (JASKOWIAK et al., 2014; SOLTO et al., 2015), Medicina (PAUL et al., 2010; VELOSO et al., 2014),

Marketing (ALSUWAIDAN et al., 2014; GULL et al., 2014), internet (HE et al., 2015; HROMIC, 2015), entre outras áreas.

As tarefas preditivas podem ser divididas em tarefas de classificação e tarefas de regressão. Em tarefas de classificação, busca-se atribuir categorias predefinidas a dados. Por exemplo, um banco pode desenvolver um sistema para a classificação de seus clientes em duas categorias para fornecimento de empréstimo: SIM e NÃO. Por meio do histórico de crédito dos clientes, e também de dados como salário e tempo de emprego (atributos de entrada), o sistema aprenderá a distinguir os clientes para os quais o banco deve (SIM) ou não deve (NÃO) fornecer um empréstimo. Assim, tem-se um sistema de recomendação de crédito, cujas categorias a serem preditas para um novo cliente são SIM e NÃO (atributo de saída). Como exemplos de classificação, podem-se citar predição de funções de proteínas (MENG et al., 2016; XU et al., 2016), classificação de documentos (LI et al., 2016; PHAN et al., 2016) e classificação de imagens (BARKER et al., 2016, HUANG et al., 2016).

Nas tarefas de regressão, objetiva-se prever o valor de uma variável numérica (atributo de saída), dadas outras variáveis (atributos de entrada). Assim, em vez de encontrar uma classe associada, como na classificação, deve-se encontrar uma função que mapeie um exemplo para um número. Como exemplos de regressão, podem-se citar a predição de valores de ações (MACEDO et al., 2013, QIU et al., 2016), predição de volume de chuvas (DI et al. 2015; USHA et al., 2015), predição de preços de produtos (PARK; BAE. 2015; CHRISTINA; UMBARA, 2015) e predição de séries temporais (AK et al., 2016; WU; LEE, 2015). Todas essas aplicações evidenciam que tarefas de predição são muito relevantes e atuais, nas quais diversos métodos de AM vêm sendo empregados.

Neste artigo, inicialmente serão discutidos alguns problemas de classificação mais complexos, cada vez mais estudados na literatura. São problemas que podem ser enfrentados na indústria e na academia, seja em aplicações comerciais, seja em processos de investigação científica. Serão apresentados exemplos de aplicação de métodos de classificação em diferentes campos, como, por exemplo, a Bioinformática. Em seguida, apresenta-se uma revisão bibliográfica, com a literatura que trata da utilização de métodos de AM na agropecuária. Trabalhando os conteúdos apresentados, este texto

pode ser relevante para aqueles que trabalham e/ou têm interesse em conhecer a aplicação de métodos de AM. Muitos grupos que trabalham com análise de dados já aplicam métodos de AM. Como exemplos, podem ser mencionados: a Dow AgroSciences; o grupo de Bioinformática e Modelagem Matemática e Computacional de Biossistemas, da Embrapa Gado de Leite; o Grupo de Pesquisa de Inteligência Computacional, da Embrapa Informática Agropecuária; o Laboratório de Computação Bioinspirada, da Universidade de São Paulo (em São Carlos, SP); os grupos de Aprendizado de Máquina e Inteligência Artificial da Universidade Federal de São Carlos; além de diversos outros grupos em universidades e centros de pesquisa no Brasil e no exterior.

EMBASAMENTO TEÓRICO

Classificação de dados

Essa classificação consiste em atribuir categorias predefinidas a dados (FACELI et al., 2011; TAN et al., 2005). Formalmente, um problema de classificação pode ser definido da seguinte maneira. Dado um conjunto de exemplos de treinamento composto por pares (x_i, c_j), no qual x_i representa um vetor de atributos de entrada que descrevem um exemplo, e c_j sua classe associada, deve-se encontrar uma função que mapeie cada x_i para sua classe associada c_j, tal que $i = 1, 2, ..., n$, em que n é o número de exemplos de treinamento, e $j = 1, 2, ..., m$, em que m é o número de classes do problema.

Em muitos problemas de classificação, as classes do problema possuem uma relação. Por exemplo, um documento ou imagem pode pertencer a duas categorias. Nesse caso, o problema de classificação é mais complexo do que a classificação convencional, sendo denominado problema de classificação multirrótulo (TSOUMAKAS et al., 2010).

Classificação multirrótulo

Formalmente, pode-se definir um problema de classificação multirrótulo da seguinte maneira. Em uma classificação multirrótulo, cada exemplo pode ser associado a duas ou mais classes ao mesmo tempo. Um classificador multirrótulo pode ser definido como uma função $H : x \rightarrow 2^L$ que mapeia um exemplo x em

um conjunto de classes $C \in 2^L$, em que 2^L é o conjunto potência de L, ou seja, o conjunto formado por todos os subconjuntos de L.

No passado, o estudo de métodos de classificação multirrótulo era motivado principalmente pelas tarefas de classificação de textos e diagnósticos médicos. Em um problema de classificação de textos, cada documento pode pertencer simultaneamente a mais de uma classe (ou tópico). Um documento, por exemplo, pode ser classificado como pertencente à área de Ciência da Computação e à Física. Uma matéria de jornal que aborda as reações da Igreja Católica com o lançamento do filme *O Código da Vinci* pode ser classificada, ao mesmo tempo, nas categorias Sociedade/Religião e Artes/Cinema. Na área de diagnósticos médicos, um paciente pode sofrer de diabetes e câncer de próstata ao mesmo tempo (TSOUMAKAS; KATAKIS, 2007). A área de classificação de textos é a que tem maior aplicação de métodos de classificação multirrótulo (GONÇALVES; QUARESMA, 2003; LAUSER; HOTHO, 2003; LUO; ZINCIR-HEYWOOD, 2005; ZHANG et al., 2010). Entretanto, muitos trabalhos podem ser encontrados nas áreas de Bioinformática (CERRI et al., 2014; CLARE, 2003; ELISSEEFF; VENS et. al., 2008, SCHIETGAT et al., 2010; STOJANOVA et al., 2013; WESTON, 2001; ZHANG; ZHOU, 2005), diagnóstico médico (KARALIC; PIRNAT, 1991; NICOLAS et al., 2014; SHAO et al., 2013) e classificação de imagens e vídeo (BOUTELL, 2004; SHEN, 2003, ZHANG; ZHOU, 2007), classificação de música e emoções (LI et al., 2006; TROHIDIS et al., 2008) e marketing direcionado (AGRAWAL, 2013).

Outro campo de aplicação muito importante que envolve classificação multirrótulo é a produção de remédios. Para a indústria farmacêutica, a predição do perfil metabólico de uma droga é importante para prevenir interações entre diferentes drogas administradas simultaneamente. Quando dois remédios são administrados em conjunto, o efeito de um pode ser inibido pela presença do outro. Por exemplo, duas drogas podem ser metabolizadas pela mesma enzima, e a competição pelo sítio de ligação com a enzima pode resultar na inibição de uma das drogas (MICHIELAN, 2009; SU et al., 2010).

O número de classes envolvidas em problemas de classificação multirrótulo também pode chegar a centenas ou milhares, tornando ainda mais desafiadora a tarefa de classificação. Em problemas de classificação que envolvem páginas web, por exemplo, a categorização das páginas pode ajudar

no desenvolvimento de sistemas de recomendação de propaganda (propaganda direcionada), envolvendo milhares de classes (AGRAWAL, 2013; DEKEL; SHAMIR, 2010).

Classificação hierárquica

Problemas ainda mais complexos do que os problemas multirrótulo podem ser encontrados. Além de exemplos poderem ser classificados em mais de uma categoria ao mesmo tempo, podem-se estabelecer relacionamentos hierárquicos entre essas categorias. Assim, as categorias de um problema formam uma hierarquia, caracterizando um problema de classificação hierárquica multirrótulo (SILLA; FREITAS, 2010). Dada uma hierarquia, o programa deve ser capaz de classificar um exemplo em mais de um ramo dessa hierarquia, simultaneamente. Esse tipo de classificação é muito comum, por exemplo, nos campos da Bioinformática e classificação de textos.

Da mesma maneira que nos problemas multirrótulo, classificadores hierárquicos podem ser definidos como funções $H : \mathrm{x} \rightarrow 2^L$ que mapeiam um exemplo **x** em um conjunto de classes $C \in 2^L$, sendo 2^L o conjunto potência de L, ou seja, o conjunto formado por todos os subconjuntos de L. Adicionalmente, o classificador H deve respeitar a restrição imposta pela hierarquia de classes, ou seja, para todo $c \in C$, qualquer c' superclasse de c deve pertencer a C.

Como exemplo de aplicação no campo da Bioinformática, pode-se citar a proteômica, que visa à identificação de proteínas expressas pelo genoma e predição de suas funções. Muitas bases de dados são organizadas de maneira hierárquica, quando distribuídas em classes e subclasses, e alguns problemas hierárquicos são multirrótulo (proteínas podem desempenhar mais de uma função), dificultando ainda mais a tarefa de classificação. Apesar de existirem métodos específicos para a classificação funcional de proteínas, métodos de AM têm sido muito utilizados, pois são capazes de aprender a tarefa de classificação utilizando diversas características das proteínas, e ainda exploram os relacionamentos hierárquicos entre suas funções (CERRI et al., 2014; FACELI et al., 2011; SCHIETGAT et al., 2010; STOJANOVA et al., 2013; VENS et al., 2008).

Um exemplo de hierarquia de classes de proteínas é a EC, provida pela Enzyme Commission (WEBB et al., 1992), que organiza a classificação funcional de enzimas (proteínas que funcionam como catalizadores de reações químicas). Nessa classificação, as classes são identificadas por um esquema

numérico composto por quatro algarismos, em que o primeiro especifica a classe (função) mais geral, e o último, a classe (função) mais específica da proteína. Por exemplo, a enzima tripéptido aminopeptidase tem o código EC 3.4.11.4, em que EC 3 representa uma hidrolase (enzima que usa a água para catalisar algumas moléculas), EC 3.4 representa a hidrolase que atua sobre as ligações peptídicas, EC 3.4.11 representa aquelas hidrolases que atuam sobre o amino-terminal de aminoácidos de um polipeptídeo, e EC 3.4.11.4 são as hidrolases que atuam sobre o amino-terminal final de um tripeptídeo.

A Gene Ontology (ASHBURNER et al., 2000) é outro exemplo de hierarquia multirrótulo de classes. Ela está organizada em três ontologias: ontologia de componentes celulares, ontologia de processos biológicos e ontologia de funções moleculares. Nessas ontologias, genes e proteínas, por exemplo, estão organizados de maneira hierárquica, e podem apresentar mais de uma função ou característica. Desenvolver ferramentas para organizar e recuperar essas informações é muito importante para a correta predição da função de novas proteínas.

No campo da classificação de textos, com o aumento do número de documentos gerados, principalmente na internet, classificar documentos de maneira hierárquica tem se tornado uma tarefa cada vez mais frequente. Documentos podem ser simultaneamente classificados em mais de uma classe. A base de dados MedLine (MEDLINE DATABASE, 2008) é um exemplo de repositório hierárquico multirrótulo que contém milhões de registros científicos. Apesar de repositórios como o MedLine terem uma grande quantidade de informação, eles ainda carecem de bons mecanismos para recuperar essas informações. Sendo assim, métodos de AM podem desempenhar um importante papel na classificação de documentos. A Wikipédia é outro exemplo de repositório de dados no qual os textos são classificados em várias categorias, estruturadas hierarquicamente.

EXEMPLOS DE APLICAÇÃO DE APRENDIZADO DE MÁQUINA NA AGROPECUÁRIA

Métodos de AM vêm sendo aplicados na agropecuária há muitos anos. No trabalho de Garner et al. (1995), um framework muito utilizado pela comunidade de AM (HALL et al., 2009) foi utilizado para gerar "árvores de decisão" (QUINLAN, 1993). Dessas árvores, foram extraídas regras de classificação

para auxiliar fazendeiros na tomada de decisão com relação ao abate de gado. Uma árvore de decisão pode ser definida como uma estrutura que particiona um conjunto de dados em subconjuntos, até que os conjuntos obtidos com esse particionamento contenham dados de apenas um tipo. Nós internos da árvore representam atributos dos dados, e nós folhas representam as classes. Cada particionamento é obtido por meio de um nó interno, comparando, para cada exemplo, seu valor correspondente ao nó interno (atributo) com o valor contido no nó. Assim, a classificação de um exemplo é dada passando-se por todos os nós internos da árvore, até que um nó folha seja alcançado.

Na agricultura de precisão, o AM tem sido utilizado de muitas maneiras. Dados coletados durante a colheita, plantio e fertilização, por exemplo, podem ser analisados com o propósito de obter vantagem competitiva e, consequentemente, vantagem econômica. No trabalho de Russ et al. (2009), por exemplo, redes neurais conhecidas como Mapas Auto-Organizáveis (KOHONEN, 1990) foram utilizadas para visualização de dados agrícolas. Foram extraídas características de diferentes áreas de plantio, após a utilização de diferentes estratégias de fertilização. A utilização das redes neurais permitiu que os dados das áreas fossem analisados visualmente, tendo permitido que fossem encontradas correlações e, eventualmente, feitas previsões utilizando dados passados. Redes neurais podem ser definidas como modelos computacionais baseados no funcionamento do cérebro. Assim como o cérebro possui neurônios, as redes neurais possuem neurônios artificiais organizados em camadas. Cada neurônio computa uma função matemática com os sinais que recebe como entrada. Esses sinais podem ser os dados propriamente ditos, ou podem ser provenientes de outros neurônios. O conhecimento que a rede aprende é armazenado nos pesos das conexões entre os neurônios. Esses pesos vão sendo ajustados durante o aprendizado da rede.

A tarefa de previsão de exportações de produtos também já foi investigada. Em Sujja-Viriyasup e Pitiruek (2013), Máquinas de Vetores de Suporte (do inglês *Support Vector Machines* – SVMs) (VAPNIK, 1999) foram utilizadas para previsão de exportações de produtos como camarão e frango. As SVMs são utilizadas em problemas de classificação e regressão, e funcionam maximizando a margem de separação entre exemplos em problemas binários. São algoritmos poderosos, pois podem mapear exemplos para um espaço de dimensão superior, no qual estes podem ser facilmente classificados.

Em Dimitriadis e Goumopoulos (2008), regras de decisão foram descobertas utilizando AM. As regras foram construídas utilizando informações extraídas do monitoramento do solo, colheita e clima. Assim, foram descobertas regras para dar suporte a decisões relacionadas a partes específicas do solo, como, por exemplo, que tipo de irrigação utilizar, ou o tipo de fertilização mais adequado.

Redes neurais artificiais (HAYKIN, 1999) foram utilizadas no trabalho de Satizabal et al. (2012) com o objetivo de previsão de produção de determinados tipos de frutas. As informações utilizadas para o treinamento das redes neurais foram retiradas de várias observações de diferentes colheitas das frutas em um mesmo local. Diferentes locais também foram investigados.

No trabalho de Kessler et al. (2015), Máquinas de Vetores de Suporte foram utilizadas para classificar grãos de trigo como provenientes de plantações orgânicas e não orgânicas. A justificativa do trabalho é que, com o aumento do consumo de produtos orgânicos, há a necessidade de comprovar a autenticidade de tais produtos.

A previsão dos preços das commodities na agricultura, utilizando AM, também já foi investigada. No trabalho de Ticlavilca et al. (2010), foi utilizado Aprendizado Bayesiano (BAYES, 1763). O estudo investigou a previsão dos preços do milho, gado e porco. Foram feitas previsões dos preços para um, dois e três meses. O aprendizado Bayesiano utiliza o teorema de Bayes para classificação de dados. Trata-se de um aprendizado probabilístico, que considera que os atributos dos exemplos são independentes entre si.

No trabalho de Coopersmith et al. (2014), a previsão da umidade do solo foi feita utilizando árvores de decisão, algoritmo dos K-vizinhos mais próximos (AHA et al., 1991) e redes neurais. Trata-se de uma tarefa importante, pois o não conhecimento das condições do solo pode levar ao atolamento de máquinas pesadas. Consequentemente, ocorrem atrasos que aumentam os custos relacionados. O algoritmo dos K-vizinhos mais próximos faz parte do paradigma de aprendizado baseado em exemplos. Nesse algoritmo, exemplos são armazenados em memória. Um novo exemplo é então comparado com todos os exemplos armazenados, e classificado na classe majoritária dos seus k exemplos mais próximos.

A delineação de zonas de manejo também é um problema da agricultura que já foi investigado com o uso de AM. Nesse problema, objetiva-se descobrir a existência de partes de uma área que podem ser tratadas de maneira similar, por exemplo, para fins de fertilização. Em AM, tal problema é tratado como um problema de agrupamento de dados. Outros exemplos da investigação de tal problema podem ser encontrados na literatura (RUSS et al., 2009; SHIRATSUCHI et al., 2005).

Os trabalhos apresentados até o momento são exemplos de algumas aplicações dentro da agropecuária nas quais o AM mostra-se útil. Muitas outras aplicações e trabalhos podem ser encontrados na literatura, que é vasta no assunto. O leitor interessado pode buscar, por exemplo, pelo tema *Big Data*, e encontrará diversos artigos e reportagens que tratam da aplicação desse tema na pesquisa agropecuária e em outras áreas.

O termo *Big Data* refere-se à análise, utilizando AM, de grandes volumes de dados, provenientes muitas vezes de fontes heterogêneas. Trata-se de uma "avalanche de dados" que precisa ser analisada para a extração de conhecimento que venha a ser útil para, por exemplo, aumentar a produtividade, número de vendas, entre outras aplicações.

Um exemplo do uso da *Big Data* no setor agropecuário pode ser encontrado na agricultura de precisão. Cada vez mais, dados coletados por tratores e outras máquinas agrícolas, além de dados coletados por satélites, serão analisados utilizando métodos computacionais, com a finalidade de obter aumento de produtividade.

Empresas como Monsanto, entre outras, já apostam em tecnologias para orientar o agricultor no momento do plantio. É o chamado "plantio sob receita", que visa à realização de pequenos ajustes no plantio de acordo com os dados obtidos do local no qual o plantio será realizado. Esses dados incluem desde informações sobre o solo até padrões climáticos e históricos dos rendimentos das culturas daquele local. Veja por exemplo reportagem do *The Wall Street Journal* sobre o tema (BUNGE, 2014).

CONSIDERAÇÕES FINAIS

Neste trabalho, buscou-se fazer uma breve introdução sobre os conceitos básicos de Aprendizado de Máquina, área da Ciência da Computação que busca

a construção de métodos computacionais capazes de aprender, baseados em experiência (treinamento utilizando-se conjuntos de dados), a executar tarefas. Foram apresentados alguns conceitos de problemas avançados de classificação de dados, e também apresentados exemplos de aplicações do Aprendizado de Máquina em uma das mais importantes áreas para o desenvolvimento econômico do País, a agropecuária.

Com o crescente aumento do volume de dados gerados, não só por pesquisas científicas, mas também por qualquer outra fonte, como, por exemplo, redes sociais e comércio eletrônico, uma conclusão torna-se inevitável: empresas terão que investir em equipes de analistas de dados se quiserem permanecer competitivas. Por trabalharem em uma área multidisciplinar, esses profissionais devem ter o domínio de métodos computacionais e estatísticos, além de conhecimentos de diversas outras áreas, como neurociência e biologia molecular. Todo esse domínio deve ser utilizado para extrair conhecimento útil dos dados disponibilizados.

REFERÊNCIAS

AGRAWAL, R.; GUPTA A.; PRABHU, Y.; VARMA, M. Multi-label learning with millions of labels: recommending advertiser bid phrases for web pages. 2013. Disponível em: <http://manikvarma.org/pubs/agrawal13.pdf>. Acesso em: 24 jan. 2018.

AHA, D. W.; KIBLER, D.; E ALBERT, M. K. Instance-based learning algorithms. **Machine Learning**, v. 6, n. 1, p. 37-66, 1991.

AK, R.; FINK, O.; ZIO, E. Two Machine Learning Approaches for Short-Term Wind Speed Time-Series Prediction. **IEEE Transactions on Neural Networks and Learning Systems**, v. 27, n. 8, p. 1734-47, 2016. DOI: 10.1109/TNNLS.2015.2418739.

ALSUWAIDAN, L.; YKHLEF, M.; ALNUEM, M. A. A novel spreading framework using incremental clustering for viral marketing. In: INTERNATIONAL CONFERENCE ON COMPUTER SYSTEMS AND APPLICATIONS, 11th, 2014, Doha. **Abstract…** Doha, 2014. p. 78-83.

ASHBURNER, M.; BALL, C. A.; BLAKE, J. A.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J. M.; DAVIS, A. P.; DOLINSKI, K.; DWIGHT, S. S.; EPPIG, J. T.; HARRIS, M. A.; HILL, D. P.; ISSEL-TARVER, L.; KASARSKIS, A.; LEWIS, S.; MATESE, J. C.; RICHARDSON, J. E.; RINGWALD, M.; RUBIN, G. M.; SHERLOCK, G. Gene ontology: tool for the unification of biology. **Nature Genetics**, v. 25, n. 1, p. 25-29, May 2000.

BARKER, J.; HOOGI, A.; DEPEURSINGE, A.; RUBIN, D. L. Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. **Medical Image Analysis**, v. 30, p. 60-71, May, 2016.

BAYES, T. An essay towards solving a problem in doctrine of chances. **Philosophical Transactions**, v. 53, p. 293-315, 1763.

BOUTELL, M. R.; LUO, J.; SHEN, X.; BROWN, C. M. Learning multi-label scene classification. **Pattern Recognition**, v. 37, n. 9, p. 1757-1771, 2004.

BUNGE, J.; **Tecnologia do 'big data' chega à lavoura e semeia desconfiança**. 2014. Disponível em: <http://blogdocaminhoneiro.com/2014/03/tecnologia-do-big-data-chega-a-lavoura-e-semeia-desconfianca/>. Aceso em: 5 mar. 2014.

CERRI, R.; BARROS, R. C.; CARVALHO, A. C. P. L. F. Hierarchical Multi-label Classification using Local Neural Networks. **Journal of Computer and System Sciences**, v. 80, p. 39-56, 2014.

CHRISTINA, C.; UMBARA, R. F. Gold price prediction using type-2 neuro-fuzzy modeling and ARIMA. In: INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGY, 3rd, 2015, [Gedung Teknik]. [**Abstracts…**] [Gedung Teknik], 2015. p. 272-277. DOI: ieeexplore.ieee.org/document/7231435/.

CLARE, A. **Machine learning and data mining for yeast functional genomics**. 2003. 204 f. Thesis (Doctor of Philosophy) – University of Wales, Aberystwyth.

COOPERSMITH, E. J.; MINSKER, B. S.; WENZEL, C. E.; GILMORE, B. J. Machine learning assessments of soil drying for agricultural planning. **Computers and Electronics in Agriculture**, v. 104, p. 93-104, 2014.

DEKEL, O.; SHAMIR, O. Multiclass-multilabel classification with more classes than examples. **Journal of Machine Learning Research**, v. 9, p. 137-144, 2010.

DI, Y.; DING, W.; MU, Y.; SMALL, D. L.; ISLAM, S.; CHANG, N. B. Developing machine learning tools for long-lead heavy precipitation prediction with multi-sensor data. In: INTERNATIONAL CONFERENCE ON NETWORKING, SENSING AND CONTROL, 12th, 2015, Taipei. **Proceedings…** Taipei: IEEE, 2015. p. 63-68. DOI: http://ieeexplore.ieee.org/document/7116011/.

DIMITRIADIS, S.; GOUMOPOULOS, C. Applying machine learning to extract new knowledge in precision agriculture applications. In: PANHELLENIC CONFERENCE ON INFORMATICS, 12th, 2008, Washington, DC. **Proceedings…** Washington, DC, 2008. p. 100-104.

ELISSEEFF, A.; WESTON, J. Kernel methods for multi-labelled classification and categorical regression problems. **Advances in Neural Information Processing Systems**, v. 14, p. 681-687, 2001.

FACELI, K.; LORENA, A. C.; GAMA, J. ; CARVALHO, A. C. P. L. F. de. **Inteligência artificial:** uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

FORGY, E. W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. **Biometrics**, v. 21, n. 3, p. 768-769, 1965.

GARNER, S. R.; CUNNINGHAM, S. J.; HOLMES, G.; NEVILL-MANNING, C. G.; WITTEN, I. H. Applying a machine learning workbench: experience with agricultural

databases. MACHINE LEARNING IN PRACTICE WORKSHOP, 1995. Tahoe City. **Conference…** Tahoe City: 1995. p. 14-21.

GONÇALVES, T.; QUARESMA, P. **A preliminary approach to the multilabel classification problem of portuguese juridical documents**. Heidelberg, 2003. DOI:O rg/10.1007/978-3-540-24580-3_50.

GULL, K. C.; ANGADI, A. B.; SEEMA, C. G.; KANAKARADDI, S. G. A clustering technique to rise up the marketing tactics by looking out the key users taking Facebook as a case study, In: IEEE INTERNATIONAL ADVANCE COMPUTING CONFERENCE, 3rd, 2014, Gurgaon. **Proceedings…** Gurgaon: IACC, 2014. p. 579-585.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA data mining software: an update. **SIGKDD Explorations Newsletter**, v. 11, n. 1, p. 10-18, 2009.

HAYKIN, S. **Neural networks:** a comprehensive foundation. 2nd ed. Upper Saddle River: Prentice Hall PTR, 1999.

HE, J.; YANG, T.; WEI A.; DONG, W. Research on degree of video completion of Internet videos with clustering algorithms. In: INTERNATIONAL CONFERENCE ON CONTROL, AUTOMATION AND INFORMATION SCIENCES, 20th, 2015, Changshu. **Proceedings…** Changshu, 2015. p. 89-95.

HORTA, D.; CAMPELLO, R. J. G. B. Automatic aspect discrimination in data clustering. **Pattern Recognition**, v. 45, p. 4370-4388, 2012.

HROMIC, H. Real time analysis of sensor data for the Internet of things by means of clustering and event processing. In: IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS, 2015, Kansas City. **Proceedings…**Kansas City, 2015. p. 685-691.

HUANG, K.; LI, S.; KANG, X.; FANG, L. Spectral-spatial hyperspectral image classification based on KNN. **Sensing and Imaging**, v. 17, n. 1, n. 1, p. 1-13, 2016.

JASKOWIAK, P. A.; CAMPELLO, R. J. G. B. ; COSTA, I. G. On the selection of appropriate distances for gene expression data clustering. **BMC Bioinformatics**, v. 15, p. S2, 2014.

KARALIC, A.; PIRNAT, V. Significance level based multiple tree classification. **Informatica**, 1991, v. 15, n. 5. p. 12, 1991.

KESSLER, N.; BONTE, A.; ALBAUM, S. P.; MÄDER, P.; MESSMER, M.; GOESMANN, A.; NIEHAUS, K.; LANGENKÄMPER, G.; NATTKEMPER, T. W. Learning to classify organic and conventional wheat - a machine- learning driven approach using the meltdb 2.0 metabolomics analysis platform. **Frontiers in Bioengineering and Biotechnology**, v. 24, n. 3, p. 35, Mar. 2015.

KOHONEN, T. K. The self-organizing map. **Proceedings of the IEEE**, v. 78, n. 9, p. 1464-1480, Sept. 1990.

LAUSER, B.; HOTHO, A. Automatic multi-label subject indexing in a multilingual environment. 7th European In: CONFERENCE IN RESEARCH AND ADVANCED

TECHNOLOGY FOR DIGITAL LIBRARIES, 7th, 2003, Trondheim. **Proceedings...** Trondheim: Springer, 2003. p. 17-22.

LI, T.; ZHANG, C.; E ZHU, S. Empirical studies on multi-label classification. In: IEEE INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE, 13-15, nov. 2006, Arlington. **Proceedings…** Arlington: IEEE, 2006. p. 86-92.

LI, X.; OUYANG, J.; ZHOU, X. Labelset topic model for multi-label document classification. **Journal of Intelligent Information Systems**, v. 46, n. 1, p. 83-97, 2016.

LUO, X.; ZINCIR-HEYWOOD, N. A. Evaluation of two systems on multi-class multi-label document classification. In: INTERNATIONAL SYMPOSIUM ON METHODOLOGIES FOR INTELLIGENT SYSTEMS, 15th 2005, Saratoga Spring. **Proceedings…**Saratoga Pring: ISMS, 2005. p. 161-169.

MACEDO, J. A.; CAMARGO, L. T. O.; OLIVEIRA, H. C. B.; SILVA, L. E.; MENEZES. R. S. An intelligent decision support system to investment in the stock market. **Revista IEEE América Latina**, v. 11, p. 812-819, 2013.

MEDLINE DATABASE. **MedLine database**. 2008. Disponíel em: <https://www.ebscohost.com/nursing/products/medline-databases>. Acesso em: 25 jan. 2018.

MENG, J.; WEKESA, J.-S.; SHI, G.-L LUAN, Y.-S. Protein function prediction based on data fusion and functional interrelationship, **Mathematical Biosciences**, v. 274, p. 25-32, Apr. 2016.

MICHIELAN, L.; TERFLOTH, L.; GASTEIGER, J.; MORO, S. Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome p450 substrates. **Journal of Chemical Information and Modeling**, v. 49, n. 11, p. 2588-2605, 2009.

NICOLAS, R.; FORNELLS, A.; GOLOBARDES, E.; CORRAL, G.; PUIG, S.; MALVEHY, J. DERMA: a melanoma diagnosis platform based on collaborative multilabel analog reasoning. **The Scientific World Journal**, v. 4, n. 1, 2014.

PARK, B.; BAE, J. K. Using machine learning algorithms for housing price prediction: the case of Fairfax County, Virginia housing data. **Expert Systems with Applications**, v. 42, v. 6, n. 15, p. 2928-2934, 2015.

PHAN, T. V.; NAKAGAWA, M. Combination of global and local contexts for text/non-text classification in heterogeneous online handwritten documents, **Pattern Recognition**, v. 51, Mar. p. 112-124, 2016.

PAUL, R.; HOQUE, A. S. M. L. Clustering medical data to predict the likelihood of diseases. INTERNATIONAL CONFERENCE ON DIGITAL INFORMATION MANAGEMENT, 5th, 2010, Thunder Bay. **Proceedings…** Thunder Bay: ICDIM, 2010. p. 44-49.

QIU, M.; SONG, Y.; AKAGI, F. Application of artificial neural network for the prediction of stock market returns: the case of the Japanese stock market. **Chaos, Solitons & Fractals**, v. 85, p. 1-7, 2016.

QUINLAN, J. R. **C4.5:** programs for machine learning. San Francisco: Morgan Kaufmann, 1993.

RUSS, G.; KRUSE, R.; SCHNEIDER, M.; WAGNER, P. Visualization of agriculture data using self-organizing maps. In: ALLEN, T.; ELLIS, R.; PETRIDIS, M. (Ed.). **Applications and Innovations in Intelligent Systems**. London: Springer, 2009. p. 47-60.

SATIZÁBAL, H.; BARRETO-SANZ, M.; JIMÉNEZ, D.; PÉREZ-URIBE, A.; COCK, J. Enhancing decision-making processes of small farmers in tropical crops by means of machine learning models. In: BOLAY, J.-C.; SCHMID, A.; TEJADA, G.; HAZBOUN, E. (Ed.). **Technologies and innovations for development**. Paris: Springer, 2012. p. 265-277.

SHAO, H.; LI, G.; LIU, G.; WANG, Y. Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine. **Science China Information Sciences**, v. 56, n. 5, p 1-13, 2013.

SCHIETGAT, L.; VENS, C.; STRUYF, J.; BLOCKEEL, H.; KOCEV, D.; DZEROSKI, S. Predicting gene function using hierarchical multi-label decision tree ensembles. **BMC Bioinformatics**, v. 11, n. 2, 2010.

SHEN, X.; BOUTELL, M.; LUO, J.; BROWN, C. Multilabel machine learning and its application to semantic scene classification. In: STORAGE AND RETRIEVAL METHODS AND APPLICATIONS FOR MULTIMEDIA, 2003, San Jose. **Proceedings**... San Jose: The International Society for Optics and Photonics, 2003. v. 5307, p. 188-199. DOI: 10.1117/12.523428.

SHIRATSUCHI, L. S.; QUEIROS, L. R.; FACCIONI, G. C. **Classificação não-supervisionada no delineamento de zonas de manejo**. Brasília, DF: Embrapa Cerrados, 2005. (Embrapa Cerrados. Boletim de Pesquisa e Desenvolvimento, 150).

SILLA, C.; FREITAS, A. A survey of hierarchical classification across different application domains. **Data Mining and Knowledge Discovery**, v. 22, p. 31-72, 2010.

SOUTO, M. C. P. de; JASKOWIAK, P. A.; COSTA, I. G. Impact of missing data imputation methods on gene expression clustering and classification. **BMC Bioinformatics**, v. 16, p. 15, Feb. 2015.

STOJANOVA, D.; CECI, M.; MALERBA, D.; DZEROSKI, S. Using PPI network autocorrelation in hierarchical multi-label classification trees for gene function prediction. **BMC Bioinformatics**, v. 14, n. 1, p. 1-18, 2013.

SU, H.; HEINONEN, M.; ROUSU, J. Multilabel classification of drug-like molecules via max-margin conditional random fields. **European Workshop on Probabilistic Graphical Models**. [S.l.]: HIIT Publications, 2010. p. 265–272.

SUJJA-VIRIYASUP, T.; PITIRUEK, K. Agricultural product forecasting using machine learning approach. **International Journal of Mathemathical Analysis**, v. 7, p. 1869-1875, 2013.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. Boston: Addison-Wesley Longman, 2005.

TICLAVILCA, A. M.; FEUZ, D.; MCKEE, M. Forecasting agricultural commodity prices using multivariate bayesian machine learning regression. In: CONFERENCE ON APPLIED COMMODITY PRICE ANALYSIS, 2010, St. Louis. **Proceedings…** St Louis, 2010. p. 1-20.

TSOUMAKAS, G.; KATAKIS, I.; VLAHAVAS, I. P. Mining multi-label data. In: MAIMON, O.; ROKACH, L. (Ed.). **Data mining and knowledge discovery handbook**. 2nd ed. New York: Springer, 2010. p. 667-685. DOI: 10.1007/978-0-387-09823-4_34.

TSOUMAKAS, G.; KATAKIS, I. Multi Label Classification: an overview. **International Journal of Data Warehousing and Mining**, v. 3, n. 3, p. 1-13, 2007.

TROHIDIS, K.; TSOUMAKAS, G.; KALLIRIS, G.; VLAHAVAS, I. Multilabel classification of music into emotions. In: INTERNATIONAL CONFERENCE ON MUSIC INFORMATION RETRIEVAL, 2008, Philadelphia. **Proceedings…** Philadelphia, 2008. p. 325-330.

USHA R RANI, T.K.RAMA KRISHNA RAO AND KIRAN KUMAR R REDDY. An Efficient Machine Learning Regression Model for Rainfall Prediction. **International Journal of Computer Applications**, v. 115, n. 23, p. 24-30, 2015.

VAPNIK, V. N. **The nature of statistical learning theory**. 2nd. New York: Springer-Verlag, 1999. (Information Science and Statistics).

VELOSO, R.; PORTELA, F.; SANTOS, M. F.; SILVA, A.; RUA, F.; ABELHA, A.; MACHADO, J. A clustering approach for predicting readmissions in intensive medicine. **Procedia Technology**, v. 16, p. 1307-1316, 2016.

VENS, C.; STRUYF, J.; SCHIETGAT, L.; DZEROSKI, S.; BLOCKEEL, H. VENS, C.; STRUYF, J.; SCHIETGAT, L.; DZEROSKI, S.; BLOCKEEL, H. Decision trees for hierarchical multi-label classification. **Machine Learning**, v. 73, n. 2, p. 185-214, 2008.

XU, Y.; MIN, H.; SONG, H.; WU, Q. Multi-instance multi-label distance metric learning for genome-wide protein function prediction. **Computational Biology and Chemistry**, v. 63, p. 30-40, Aug. 2016.

WEBB, E. C. **Enzyme nomenclature 1992**: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes. San Diego: Academic Press, 1992. 862 p.

WU, S-F.; LEE, S-J. Employing local modeling in machine learning based methods for time-series prediction, **Expert Systems with Applications**, v. 42, n. 1, p. 341-354, 2015.

ZHANG, M.-L.; ZHOU, Z.-H. A k-nearest neighbor based algorithm for multi-label classification. In: IEEE INTERNATIONAL CONFERENCE ON GRANULAR COMPUTING, 2005. **Proceedings…**Beijing: IEEE, 2005. p. 718-721.

ZHANG, M.-L.; ZHOU, Z. Multi-Label learning by instance differentiation. In: CONFERENCE ON ARTIFICIAL INTELLIGENCE, 22nd, 2007, Vancouver. **Proceedings...** Vancouver: AAAI, 2007. p. 669-674.

ZHANG, M.-L.; ZHANG, K. Multi-label learning by exploiting label dependency. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 16th, 2010, Washington, DC. **Proceedings…** Washington, DC, 2010. p. 999-1008.

Trabalho recebido em 22 de maio de 2015 e aceito em 8 de abril de 2016.