ESTUDO DE CARACTERÍSTICAS DE BASES DE DADOS PARA SELEÇÃO DE ALGORITMOS DE MODELAGEM

Artur Falci Antonon de Souza (IC) e Mario Olimpio de Menezes (Orientador)

Apoio:PIBIC Mackpesquisa

RESUMO

Este artigo tem como objetivo estudar as características de certas bases de dados para a seleção de algoritmos de modelagem através de medidas de complexidade propostas por Lorena et al. (2018). Com isso buscou-se a melhor compreensão de tais, fornecidas em R pela pesquisadora, para a aplicação em Python. Todas as bases usadas neste estudo foram retiradas do repositório UCI - Machine Learning Repository (DHEERU; TANISKIDOU, 2017). A análise dos resultados das medidas é indicativa da simplicidade ou complexidade dos conjuntos de dados. Além disso foram obtidos resultados da Raiz do Erro Médio Quadrático (RMSE) através dos algoritmos de modelagem, como XGBoost (CHEN; GUESTRIN, 2016) e Random Forest (BREIMAN, 2001), assim como de Regressão Linear Multivariada. Apesar de utilizarmos apenas quatro bases e três algoritmos de regressão, podemos perceber que as medidas de complexidade podem ajudar a prever o desempenho dos algoritmos. Por exemplo, as bases Airfoil e Concrete tiveram valores elevados em S2 (uma das medidas de Suavidade), que foi refletido no pior resultado do RMSE com regressão linear para estas bases. Os conjuntos que tiveram valores menores nessa medida apresentam melhor RMSE comparáveis entre os algoritmos. Isto indica também, que o algoritmo de Regressão Linear é o mais sensível a esta qualidade da base de dados.

Palavras-chave: Medidas de Complexidade; Algoritmos de Modelagem; Base de Dados.

ABSTRACT

This article aims the study the characteristics of certain datasets for the selection of prediction algorithms through complexity measures proposed in Lorena et al. (2018). Thereat sought to better comprehend such measures given in R code (R Core Team, 2018; ROSSUM, 1995), to later by applied in Python. Every single dataset used in this study was taken from the UCI - Machine Learning Repository (DHEERU; TANISKIDOU, 2017). The analyses of the measures is an indicative of the complexity of such datasets. Furthermore, the aim was to compare the results obtained by the complexity measures to those obtained by getting the Root Mean Square Error (RMSE) of each dataset through modeling and predictions algorithm's such as XGBoost (CHEN; GUESTRIN, 2016), Random Forest (BREIMAN, 2001) and multivariate linear regression. Although it was only used 4 datasets and 3 regression algorithms, we can realize that such complexity measures can help predict the quality of the algorithm. As an

example, both Airfoil and Concrete, did have greater values for S2 (Smoothness group), which lead to an even worse RMSE output for the Linear Regression to those datasets. Those who have had a smaller value of S2, had a better RMSE output. This is an indication that the Linear Regression algorithm is highly sensible to this measure

Keywords: Complexity Measures; Prediction Algorithms; Datasets

1. INTRODUÇÃO

1.1. Problema de Pesquisa

O problema de se selecionar um algoritmo efetivo ou o melhor algoritmo surge em um grande número de situações, dentre elas, na modelagem estatística, incluindo aqui as técnicas de aprendizado de máquina que visam criar modelos preditivos, de regressão, de classificação e agrupamento etc. (RICE, 1976; FERRARI; CASTRO, 2015)

A técnica utilizada para se realizar esta seleção e configuração de algoritmos é denominada “Metalearning” ou “Meta-Aprendizagem”, e tem sido foco de intensas pesquisas (BRAZDIL; GIRAUD-CARRIER, 2018; ABDULRAHMAN et al., 2018; LEMKE; BUDKA; GABRYS, 2015).

Meta-conhecimento é o que se aprende no processo de estudo do problema sobre o problema, assim dando melhor base para escolher o algoritmo para resolução do problema. Lemke et al. (2015) expande a definição de meta-conhecimento para incluir conhecimento prévio obtido com outros conjuntos de dados e também de diferentes domínios ou problemas.

Outro aspecto importante do problema de seleção de algoritmos é a constatação apresentada por Lorena et al. (2018) que, o aprendizado obtido pela técnica de meta aprendizagem voltado para problemas de classificação, não é apropriado para problemas de regressão. Nas tarefas de classificação, a meta-aprendizagem descreve os problemas através de uma variedade de características, incluindo medidas de complexidade, que permitem capturar a complexidade das fronteiras que separam as classes. Para problemas de regressão, entretanto, as medidas voltadas para a estimativa da complexidade da função de regressão são diferentes e obtidas por diferentes processos de meta-aprendizagem.

Uma das estratégias utilizadas no processo de meta-aprendizagem é a definição de meta-características que são utilizadas para se descrever o problema em questão. As meta-características adotadas no processo de meta-aprendizagem são, de modo simplista, classificadas em:

- Medidas simples, como medidas estatísticas e de informação, extraídas diretamente dos dados de aprendizado;
- Descritores indicativos, extraídos de modelos simples de aprendizado aplicados aos dados;
- Características baseadas no modelo, tais como tamanho de uma árvore de decisão induzida.

A especificação destas meta-características é um dos problemas encontrados na meta-aprendizagem. Para os problemas de regressão, algumas meta-características têm sido propostas, mas são fortemente baseadas nas meta-características utilizadas em problemas de classificação, reduzindo assim, sua efetividade (SMITH-MILES, 2009). Problemas de regressão estão na classe de Algoritmos de Aprendizado Supervisionado. Destes, o mais comum é a Regressão Linear de Mínimos Quadrados. Existe também os Algoritmos não Supervisionados, que por outro lado, nos permite abordar problemas com pouca ou nenhuma ideia do que os resultados devem ser aparentar. Podemos derivar estrutura de dados onde nós não necessariamente saberíamos o efeito das variáveis. A diferença entre aqueles supervisionados e não supervisionados é a relação que se estabelece entre os dados. Enquanto os algoritmos supervisionados utilizam das características presentes nos dados para estabelecer futuras relações (estimar os parâmetros do modelo), os não supervisionados aprendem características intrínsecas dos dados para classificá-los.

1.2 Justificativa

Diversos outros autores têm proposto outras medidas de complexidade que visam melhorar a seleção de algoritmos, com um foco quase que exclusivamente em problemas de classificação. Poucos estudos, entretanto, se voltam para problemas de regressão.

Desta forma, um estudo que explore diferentes conjuntos de dados para diferentes algoritmos de regressão, especialmente algoritmos baseados em conjuntos de árvores (tree ensemble), apresenta-se como uma proposta sólida para um trabalho de pesquisa de Iniciação Científica.

1.3 Objetivo

O objetivo do presente projeto foi avaliar as meta-características apresentadas por Lorena et al. (2018) para algoritmos de regressão baseados em conjuntos de árvores, tais como: XGBoost (CHEN; GUESTRIN, 2016), Random Forest (BREIMAN, 2001) e Regressão Linear Multivariada.
Como objetivos secundários foram:

- Melhor entendimento do processo de pesquisa científica;

- Exploração de diversas técnicas de aprendizado de máquina voltadas para regressão;

- Desenvolvimento de habilidade de análise de dados, incluindo modelagem estatística, aprendizado de máquina, etc., utilizando ferramentas como Python e R (R Core Team, 2018; ROSSUM, 1995).

2. REFERENCIAL TEÓRICO

A área de meta-aprendizagem, como já mencionado anteriormente, é uma área muito ativa de pesquisa. Diversas ações entre grandes centros de pesquisa têm colaborado para o seu avanço, como o projeto STATLOG (MICHIE et al., 1994), seguido pelo projeto METAL (GIRAUD-CARRIER, 2005) e também pelo e-LICO (2012), entre outros.

A comunidade científica que trabalha com Redes Neurais geralmente utiliza um teorema, que a princípio parece jocoso, para expressar uma constatação que aparentemente vai contra os esforços de meta-aprendizagem. É o teorema No Free Lunch (NFL), ou “Sem almoço grátis”. Este teorema introduz o conceito de que é impossível construir um algoritmo genérico de aprendizagem de máquina, que possa resolver quaisquer problemas. A comunidade de aprendizagem de máquina contextualizou o teorema NFL na forma de uma Lei de Conservação para a Generalização de Desempenho (LCG): quando tomada sobre todas as tarefas de aprendizado, o desempenho de generalização de qualquer algoritmo de aprendizado soma a 0 (SCHAFFER, 1994).

Efetivamente, o teorema significa que nunca será possível construir um algoritmo de aprendizado que seja único e o melhor: um que, na média, tenha um desempenho melhor do que todos os outros algoritmos em todas as possíveis distribuições de dados. Deve ser entendido aqui, que a premissa subjacente é que todas as possíveis distribuições de dados são igualmente prováveis. E esta é uma premissa muito forte: ela se refere a um universo no qual tudo é igualmente possível. Em algumas interpretações do teorema afirma-se que não é possível aprendizado sem algum conhecimento prévio, a priori, sem se fazer hipóteses sobre a distribuição dos dados (GIRAUD-CARRIER, 2008).

Entretanto, os algoritmos de aprendizado são heurísticos por natureza, de modo que é muito difícil de se provar teoricamente que um novo algoritmo é melhor do que outros, mesmo considerando-se apenas um subconjunto de problemas. A busca por uma forma de estabelecer-se o um relacionamento entre os conjuntos de dados e os algoritmos ainda é um campo de muita investigação científica.

Desta forma, o desempenho de um determinado algoritmo será, quase que sempre, avaliado empiricamente, implicando que o Aprendizado de Máquina, é, em grande parte, uma ciência empírica. Para se testar o desempenho de um novo algoritmo de aprendizado, ou para se determinar qual, de muitos algoritmos, é o mais apropriado para se abordar um

problema de modelagem ou de mineração de dados, os algoritmos são implementados e o seu desempenho é avaliado em conjuntos de dados reais.(VANSCHOREN, 2010)

A utilidade de qualquer algoritmo de aprendizado dependerá, em última instância, de quão bem a sua própria característica de aprendizado (seu viés) consegue capturar a estrutura dos dados em questão; e esta utilidade será medida, na prática, rodando-se este algoritmo nos dados.

Entender os relacionamentos existentes nos dados (metadados) é a finalidade da meta-aprendizagem; um desafio entre propriedades mensuráveis dos dados e o próprio viés do algoritmo de um lado e os dados de desempenho empírico de outro lado. A descoberta automática de padrões nestes metadados é o que auxilia a melhorar o entendimento do comportamento de um algoritmo em diferentes tipos de dados, e finalmente, a se especificar melhores algoritmos.

A definição destes metadados, ou meta-características, é um dos desafios da meta-aprendizagem, já que estas características devem ser informativas e discriminativas o suficiente para identificar diferentes aspectos dos problemas de aprendizado que podem produzir viés no desempenho do algoritmo.

Para problemas de regressão, a complexidade do problema pode ser atribuída a vários fatores, alguns dos quais são similares aos problemas de classificação. A complexidade do problema pode ser fruto de limitações nas características disponíveis, que não são informativas o suficiente. Outro fator que pode contribuir para a complexidade do problema é a esparsidade dos dados. Aspectos como a distribuição da variável alvo também podem indicar se o problema de regressão é simples ou não. E finalmente, a própria função objetivo, que relaciona as entradas à variável alvo, tem sua complexidade intrínseca, que pode ser avaliada.

Lorena et al. (2018) propõe algumas medidas que visam estimar a complexidade dos dados para problemas de regressão a partir de diferentes perspectivas, divididas nas seguintes categorias:

- Medidas de Correlação de Variáveis (grupo **C**): buscam capturar o relacionamento entre os valores das variáveis e os resultados (**y**);

- Medidas de Linearidade (grupo **L**): visam estimar se uma função linear pode ajustar os dados, isto é, se eles são distribuídos linearmente;

- Medidas de suavidade (grupo **S**): visam estimar a suavidade das funções que devem ser ajustadas aos dados;

- Medidas de densidade, geometria e topologia (grupo **T**): buscam capturar a distribuição espacial e a estrutura dos dados.

Em cada categoria, diversas medidas são estabelecidas na forma de funções matemáticas e ou algoritmos, e são então calculadas para o conjunto de dados que está sendo analisado.

3. METODOLOGIA

3.1. Seleção dos conjuntos de dados

Essa foi a etapa onde foram selecionados os conjuntos de dados, importados dos repositórios especializados, tais como do UCI - Machine Learning Repository (DHEERU; TANISKIDOU, 2017). Tais conjuntos de dados foram utilizados para tarefas de regressão.

3.2. Determinação das medidas de complexidade nos conjuntos de dados

Foram obtidos os resultados gerados pelas medidas de complexidade propostas por Lorena et al. (2018). Estas são medidas que visam estimar a complexidade dos conjuntos de dados voltados para problemas de regressão a partir de diferentes perspectivas, divididas nas quatro categorias citadas anteriormente: Medidas de Correlação de Variáveis; Medidas de Linearidade; Medidas de Suavidade, e Medidas de Densidade, Geometria e Topologia. Estas medidas foram obtidas a partir de uma implementação na linguagem R.

3.3. Geração, Organização e Análise dos resultados

Após obtidos, os resultados das medidas de complexidade sobre as bases foram organizados em uma tabela comparativa. Importante ressaltar que todas as bases foram normalizadas (média = 0 e desvio padrão = 1) para que os resultados do RMSE (indicativo da qualidade do ajuste) pudessem ser comparados entre os algoritmos de regressão. .

Os algoritmos de regressão comparados foram Random Forest (BREIMAN, 2001) e XGBoost (CHEN; GUESTRIN, 2016) e além da Regressão Linear Multivariada, também empregado por Lorena et al. (2018), para podermos estabelecer uma base de comparação dos resultados.

4. RESULTADO E DISCUSSÃO

Para solucionar um problema que envolva análise de dados, é necessário entender a base de dados para que se possa escolher o algoritmo que poderá ter o melhor desempenho. Foram utilizadas, nesse artigo, as medidas de complexidade propostas por Lorena et al. (2018) para, justamente, entender melhor as bases de dados. Dentro de todo o

conjunto de medidas, o foco maior foi o grupo das medidas de Correlação de Variáveis (C) e de Suavidade (S).

Foram selecionadas quatro bases de dados reais do repositório UCI, voltadas para problemas de regressão, e portanto próprias para o tratamento das medidas propostas e aplicação dos algoritmos selecionados, Random Forest (BREIMAN, 2001), XGBoost (CHEN; GUESTRIN, 2016) e de Regressão Linear Multivariada.

Para a base Autos, foram removidas duas variáveis (“symboling” e “normalized losses”) pois ambas variáveis continham valores que não eram inteiros e não poderiam ser normalizados. Também foram removidas todas as observações que continham dados faltantes, resultando numa base de dados, como mostrado na Tabela 1, com 193 observações. As demais bases foram utilizadas como foram baixadas do repositório UCI, indicadas na Tabela 1 – os números entre parênteses são o total de observações de cada base. Os tempos de processamento dos algoritmos foi negligível, nenhuma das bases utilizadas era grande o suficiente para oferecer dificuldades de processamento.

Tabela 1 - Bases de Dados

Base de Dados	
Autos (193)	Airfoil (1503)
Power (9568)	Concrete (1030)

Para fazer as análises desejadas, tentamos utilizar a ferramenta Colaboratory da Google Research, para desenvolver quatro notebooks separados para cada conjunto. Neles fizemos o tratamento das bases e logo após passando-as pelas medidas de complexidade e em seguida pelos algoritmos de regressão. A ideia era implementar as medidas em R, e usando a biblioteca rpy2, chamá-las nos notebooks em Python. Porem isso se mostrou ineficiente e com resultados incoerentes, o que nos obrigou a fazer todas as análises diretamente no R, utilizando a implementação dos algoritmos desenvolvidos por Lorena et al. em R. Os resultados obtidos para as medidas de complexidade estão apresentados na Tabela 2.

Tabela 2 - Medidas de Complexidade

Medida	Autos	Airfoil	Power	Concrete
C1	0,921701	0,340783	0,944412	0,596020
C2	0,524760	0,224128	0,690071	0,291690
C3	0,000000	0,435130	0,000000	0,431068
C4	0,186528	0,241517	0,034594	0,609709
L1	0,059882	0,099144	0,048016	0,102344
L2	0,006723	0,016286	0,003643	0,016640
L3	0,005650	0,014111	0,002642	0,013892
S1	0,060041	0,055209	0,041118	0,080054
S2	0,483878	0,813669	0,316745	0,718033
S3	0,008784	0,004637	0,003297	0,011314
S4	0,003857	0,009927	0,003532	0,007233
T2	19,3	300,60	2392	128,75

É importante notar, na Tabela 2, que as medidas estão divididas em quatro grupos citados no referencial teórico deste artigo. A Tabela 3 apresenta um guia para a interpretação das medidas de complexidade (Lorena et al. 2018).

Tabela 3 - Resumo das Medidas de Complexidade (Fonte: Lorena et al (2018), traduzido pelo autor)

Categoria	Acrônimo	Min.	Max.	Custo Assintótico	Complexidade
Correlacão de variáveis	C1	0	1	$O(d \cdot n \cdot \log n)$	↓
	C2	0	1	$O(d \cdot n \cdot \log n)$	↓
	C3	0	1	$O(d \cdot n^2)$	↑
	C4	0	1	$O(d \cdot (d + n \cdot \log n))$	↑
Linearidade	L1	0	-	$O(n \cdot d^2)$	↑
	L2	0	-	$O(n \cdot d^2)$	↑
Suavidade	S1	0	-	$O(d \cdot n^2)$	↑
	S2	0	-	$O(n \cdot (d + \log n))$	↑
	S3	0	-	$O(d \cdot n^2)$	↑
Densidade e topologia	L3	0	-	$O(n \cdot (d^2 + \log n))$	↑
	S4	0	-	$O(n \cdot d \cdot \log n)$	↑
	T2	≈ 0	-	$O(n + d)$	↓

Todo o grupo de Correlação de Variáveis (grupo C), tem a característica de conter seu resultado entre 0 e 1, e valores maiores de C1 e C2 indicam problemas mais simples para esta categoria. As medidas C3 e C4, com valores também entre 0 e 1, tem comportamento inverso, ou seja, maiores valores destas medidas indicam problemas mais complexos. Analisando os valores obtidos das medidas do grupo C, observamos que nos casos onde C3 é zero (0.0), tais bases apresentam os maiores valores de C1, isto é, as variáveis explicativas destas bases tem as maiores correlações com a respectiva variável

alvo. Igualmente para a media C4, observamos que as mesmas bases que tiveram valores zero em C3, tem os menores valores em C4. Estas duas medidas tem o mesmo sentido de medida, isto é, quanto maiores os valores, mais complexas as bases. As outras bases apresentam valores diferentes de zero, e isto se reflete nas outras medidas, indicando, por exemplo, valores menores de C1, ou seja, correlações menores entre variáveis explicativas e alvo.

No grupo de Linearidade (L), quanto mais perto de zero o valor, mais simples é o problema. Algo que se precisa notar, é que o resultado da medida L2 é indicativo para problemas lineares somente, enquanto o L1 é indicativo para problemas lineares multivariados. O grupo das medidas de Linearidade, que visam estimar se uma função linear pode ajustar os dados, teve valores próximos de zero. Em todas as bases, as medidas deste grupo apresentam valores próximos de zero, o que era esperado, já que foram bases selecionadas do repositório UCI, da categoria Regressão, ou seja, próprias para abordagens deste tipo.

Por fim, o Grupo de Suavidade (S), assim como o grupo L, quanto mais baixo seu valor, mais simples o problema. Como esperado até então, todas as medidas do grupo S tiveram valores próximos de zero, com exceção da medida S2 que calcula a distância Euclidiana entre um exemplo e seus vizinhos, ou seja, S2 complementa a medida S1 por calcular o quão similar é o espaço da entrada de dados com a saída de dados.

Dos resultados das medidas de complexidade apresentados na Tabela 2, olhando para o grupo C, as bases Airfoil e Concrete apresentam, comparativamente, maior complexidade para os problemas de regressão. Este resultado é corroborado quando olhamos para o grupo S, onde estas duas bases apresentam os maiores valores de S2.

Após o cálculo das medidas de complexidade, foram obtidos os modelos de regressão para os algoritmos já mencionados. Na Tabela 4, temos os valores do RMSE (Raiz do Erro Médio Quadrático), que é uma métrica usual para avaliar a qualidade de modelos de regressão. Para permitir a comparação entre os diversos algoritmos entre as diversas bases, estas foram normalizadas como dito anteriormente. Desta forma, quanto mais próximos de zero os valores de RMSE, melhor o ajuste da regressão.

Tabela 4 - Valores de RMSE

Base	XGBOOST	RF	LM
AUTOS	0.3406538	**0.3034746**	0.4478332
AIRFOIL	**0.255173**	0.2567475	0.7471107
POWER	0.1904268	**0.1851376**	0.2619305
CONCRETE	**0.2548045**	0.2787932	0.619808

O pior resultado entre os três algoritmos foi o modelo linear multivariado (LM) para a base Airfoil, enquanto ambos os XGBoost e Random Forest tiveram uma performance parecida e boa. Em todas as bases o modelo linear multivariado (LM), se mostrou como o pior algoritmo. Essa diferença eventualmente possa ser atribuída as premissas do método dos mínimos quadrados, utilizados no LM, enquanto o XGBoost e o RF têm princípios totalmente diferentes baseados em conjuntos de arvores.

Ambas as bases, Airfoil e Concrete, tiveram valores elevados na medida S2, o que se mostrou com eles tendo o pior resultado dos RSME com o algoritmo LM, de 0.747 e 0.620, respectivamente. Já os conjuntos, Autos e Power, obtiveram valores das medidas do grupo C que apontam para bases de baixa complexidade, nas quais até um algoritmo simples como o LM, tem bons resultados de RMSE. Já os outros algoritmos, que são baseados em princípios totalmente diferentes, conseguiram resultados ainda melhores de RMSE para estas duas bases.

Olhando para todas as bases e os três algoritmos utilizados, vemos que tanto o XGBoost como o RF tem desempenhos muito semelhantes e muito bons. Isto se deve aos seus princípios de funcionamento, que compartilham muitas partes, e indica que são apropriados para diversos tipos de problema (bases) de regressão.

É importante esclarecer as barreiras encontradas durante a realização deste trabalho. Uma das maiores sendo a dificuldade de trabalhar com duas linguagens no mesmo notebook. Várias destas pequenas barreiras não tornaram possível que se pudesse achar uma relação extremamente coerente entre as mediadas de complexidade e os resultados do RMSE de cada algoritmo para cada base de dados. No entanto as medidas estão sim coerentes entre si e isso se mostra claro ao compará-las com as apresentadas em Lorena et al. (2018).

5. CONSIDERAÇÕES FINAIS

Tornou-se claro que é necessário mais estudo sobre das características das medidas de complexidade e de um estudo que contenha mais bases para que se possa aprofundar na relação que existe entre as medidas e os resultados dos RMSE dos algoritmos de regressão e modelagem, especificamente XGBoost e RF, já que para LM, os resultados, foram coerentes. Apesar de termos obtidos todas as medidas de complexidade, e elas estarem coerentes entre si, não foi possível, pelos motivos descritos acima, estabelecer uma relação intrínseca com os resultados dos algoritmos. Para estudos futuros, mostra-se muito

interessante adicionar outras métricas de avaliação dos algoritmos e mais conjuntos de dados, sintéticos e reais.

6. AGRADECIMENTOS

O autor Artur Falci Antonom de Souza agradece ao MackPesquisa (Universidade Presbiteriana Mackenzie) pela concessão da bolsa de Iniciação Científica, de Agosto/2018 a Julho/2019, para a realização deste trabalho.

7. REFERÊNCIAS

ABDULRAHMAN, S. M. et al. Speeding up algorithm selection using average ranking and active testing by introducing runtime. Machine Learning, v. 107, n. 1, p. 79–108, Jan 2018. ISSN 1573-0565. Disponível em: <https://doi.org/10.1007/s10994-017-5687-8>.

BRAZDIL, P.; GIRAUD-CARRIER, C. Metalearning and algorithm selection: progress, state of the art and introduction to the 2018 special issue. Machine Learning, v. 107, n. 1, p. 1–14, Jan 2018. ISSN 1573-0565. Disponível em: <https://doi.org/10.1007/s10994-017-5692-y>.

BREIMAN, L. Random forests. Machine Learning, v. 45, n. 1, p. 5–32, Oct 2001. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1010933404324>.

CAVALCANTI, G. D. C.; REN, T. I.; VALE, B. A. Data complexity measures and nearest neighbor classifiers: A practical analysis for meta-learning. In: 2012 IEEE 24th International Conference on Tools with Artificial Intelligence. [S.l.: s.n.], 2012. v. 1, p. 1065–1069. ISSN 1082-3409.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2016. (KDD '16), p. 785–794. ISBN 978-1-4503-4232-2. Disponível em: <http://doi.acm.org/10.1145/2939672.2939785>.

DHEERU, D.; TANISKIDOU, E. K. UCI Machine Learning Repository. 2017. Disponível em: <http://archive.ics.uci.edu/ml>.

e-LICO. e-LICO: An e-Laboratory for Interdisciplinary Collaborative Research in Data Mining and Data-Intensive Science. 2012. Disponível em: <http://www.e-lico.eu/>.

FERRARI, D. G.; CASTRO, L. N. de. Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods.

Information Sciences, v. 301, p. 181 – 194, 2015. ISSN 0020-0255. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0020025514011967>.

GARCIA, L. P.; CARVALHO, A. C. de; LORENA, A. C. Noise detection in the meta-learning level. Neurocomputing, v. 176, p. 14 – 25, 2016. ISSN 0925-2312. Recent Advancements in Hybrid Artificial Intelligence Systems and its Application to Real-World Problems. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231215005482>.

GIRAUD-CARRIER, C. The data mining advisor: Meta-learning at the service of practitioners. In: Proceedings of the Fourth International Conference on Machine Learning and Applications. Washington, DC, USA: IEEE Computer Society, 2005. (ICMLA '05), p. 113–119. ISBN 0-7695-2495-8. Disponível em: <http://dx.doi.org/10.1109/ICMLA.2005.65>.

GIRAUD-CARRIER, C. Metalearning-a tutorial. In: Tutorial at the 2008 International Conference on Machine Learning and Applications, ICMLA., 2008. p. 11–13.

LEMKE, C.; BUDKA, M.; GABRYS, B. Metalearning: a survey of trends and technologies. Artificial Intelligence Review, v. 44, n. 1, p. 117–130, Jun 2015. ISSN 1573-7462. Disponível em: <https://doi.org/10.1007/s10462-013-9406-y>.

LEYVA, E.; GONZáLEZ, A.; PéREZ, R. A set of complexity measures designed for applying meta-learning to instance selection. IEEE Transactions on Knowledge and Data Engineering, v. 27, n. 2, p. 354–367, Feb 2015. ISSN 1041-4347.

LORENA, A. C. et al. Data complexity meta-features for regression problems. Machine Learning, v. 107, n. 1, p. 209–246, Jan 2018. ISSN 1573-0565. Disponível em: <https://doi.org/10.1007/s10994-017-5681-1>.

MICHIE, D. et al. (Ed.). Machine Learning, Neural and Statistical Classification. Upper Saddle River, NJ, USA: Ellis Horwood, 1994. ISBN 0-13-106360-X.

MIRANDA, P. B. et al. A hybrid meta-learning architecture for multi-objective optimization of svm parameters. Neurocomputing, v. 143, p. 27 – 43, 2014. ISSN 0925-2312. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231214007693>.

MORÁN-FERNÁNDEZ, L.; BOLÓN-CANEDO, V.; ALONSO-BETANZOS, A. Can classification performance be predicted by complexity measures? a study using microarray data. Knowledge and Information Systems, v. 51, n. 3, p. 1067–1090, Jun 2017. ISSN 0219-3116. Disponível em: <https://doi.org/10.1007/s10115-016-1003-3>.

PYTHON Programming Language. Disponível em: <http://www.python.org>. Acesso em: Março/2018.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>.

RICE, J. R. The algorithm selection problem**this work was partially supported by the national science foundation through grant gp-32940x. this chapter was presented as the george e. forsythe memorial lecture at the computer science conference, february 19, 1975, washington, d. c. In: RUBINOFF, M.; YOVITS, M. C. (Ed.). Elsevier, 1976, (Advances in Computers, v. 15). p. 65 – 118. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0065245808605203>.

ROSSUM, G. Python Reference Manual. Amsterdam, The Netherlands, The Netherlands, 1995.

SCHAFFER, C. A conservation law for generalization performance. In: COHEN, W. W.; HIRSH, H. (Ed.). Machine Learning, Proceedings of the Eleventh International Conference, Rutgers University, New Brunswick, NJ, USA, July 10-13, 1994. [S.l.]: Morgan Kaufmann, 1994. p. 259–265. ISBN 1-55860-335-2.

SMITH-MILES, K. A. Cross-disciplinary perspectives on meta-learning for algorithm selection. ACM Comput. Surv., ACM, New York, NY, USA, v. 41, n. 1, p. 6:1–6:25, jan. 2009. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/1456650.1456656>.

UCI Machine Learning Repository: Airfoil Self-Noise Data Set. Disponível em: <https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise>. Acesso em: 19 jul. 2019.

UCI Machine Learning Repository: Automobile Data Set. Disponível em: <https://archive.ics.uci.edu /ml/datasets/automobile>. Acesso em: 20 jul. 2019.

UCI Machine Learning Repository: Combined Cycle Power Plant Data Set. Disponível em: <https://archive.ics.uci.edu/ml/datasets/combined+cycle+power+plant>. Acesso em: 21 jul. 2019.

UCI Machine Learning Repository: Concrete Compressive Strength Data Set. Disponível em: <http://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength>. Acesso em: 1 ago. 2019.

VANSCHOREN, J. Understanding Machine Learning Performance with Experiment Databases. Tese (Doutorado) — Katholieke Universiteit Leuven – Faculty of Engineering, Leuven, Belgium, 2010.

Contatos: arturantonon@gmail.com e mario.menezes@mackenzie.br