Dados do Projeto de Pesquisa	
Título do Projeto de Pesquisa:	Pacote Estatístico Computacional para Classificação
Grande Área/Área segundo o CNPq (https://goo.gl/JB3tAs):	Ciências Exatas e da Terra/Ciência da Computação
Grupo de Pesquisa vinculado ao projeto:	Modelagem Estatística, Simulação e Otimização de Risco
Linha de pesquisa do grupo de pesquisa vinculado ao projeto:	Análise Probabilística de Risco
Categoria do projeto:	( ) projeto em andamento, já cadastrado na PRPI ( ) projeto não iniciado, mas aprovado previamente (X) projeto novo, ainda não avaliado
Palavras-chave:	Classificação; Otimização; Inteligência Artifical; Computação Evolutiva

1. INTRODUÇÃO

A solução de um problema de classificação trata do agrupamento de objetos em classes de objetos similares. O grau de dificuldade para classificar um conjunto de dados em algum número de classes depende da variabilidade nos valores característicos. Para um certo objeto, basicamente, dois fatores podem ter efeito sobre essa variabilidade: a complexidade dos dados e o ruído [1, 2, 3, 4]. Se o conjunto de dados apresenta um baixo nível de ruído e complexidade com relações linearmente independentes para as características, então a tarefa de classificação é um problema linearmente separável e existe um hiperplano ótimo que conduz à uma classificação com taxa de reconhecimento de 100%. Por outro lado, se o problema de classificação é um problema não linearmente separável, essa taxa de reconhecimento é frequentemente impossível de ser alcançada. Em geral busca-se determinar a probabilidade de um dado objeto pertencer a cada uma das possíveis classes. Portanto, o problema linearmente separável é relativamente mais simples para resolver do que um problema não linearmente separável [5, 6].

O uso da classificação, permeia as diversas áreas do conhecimento. Sua aplicação pode ser vista em situações como diagnóstico de doenças, busca na internet, filtro de e-mails, identificação de pessoas, seleção automática de qualidade, perfil de clientes, análise de sequências de DNA, detecção de fraudes, marketing, entre muitas outras. Em geral, as técnicas de agrupamento/classificação são caracterizadas por um processo de aprendizagem não supervisionado. De modo que os dados são agrupados em *clusters* de acordo com alguma medida de similaridade. Assim, um sistema de clusterização pode definir classes (ou *clusters*) para um conjunto de dados usando um vetor característico. A ideia principal é especificar um objeto de um conjunto de dados para uma destas classes definidas.

Alguns métodos são usuais para a tarefa de classificação. Sejam não supervisionado como o *k-means* [7] e/ou supervisionado como *k-Nearest Neighbor* - kNN [8] e *Support Vector Machine*(SVM) [9]. Bem como, algumas variações destes métodos tem sido propostas [10, 11, 12, 13].

Recentemente, metodologias de clusterização/classificação tem sido apresentadas a partir da Inteligência Artificial (IA), um ramo da Ciência da Computação cujo objetivo é desenvolver sistemas que executam funções desempenhadas pelo ser humano usando co-

nhecimento e raciocínio. Estas metodologias são de grande importância e um ramo usual em algumas áreas como mineração de dados, estatística, engenharia, ciências da computação entre outras ciências que trabalham com análises de dados [1, 14, 15]. A IA é uma área que busca, através de técnicas inspiradas na natureza, o desenvolvimento de sistemas inteligentes que imitam aspectos do comportamento humano, como o aprendizado. Por exemplo, as Redes Neurais Artificiais (RNAs) [16, 17, 18] são modelos computacionais não lineares, inspirados na estrutura e operação do cérebro humano; Os Algoritmos Genéticos (AGs) [19] são algoritmos matemáticos inspirados nos mecanismos de evolução natural e recombinação genética.

Modelos baseados em RNAs [20, 21], têm sido cada vez mais aplicadas em problemas de difícil modelagem computacional ou em áreas em que um modelo matemático poderia ser muito complexo. Modelos usuais de RNAs para classificação incluem redes recorrentes, *multilayer perceptron* (MLP) e redes Kohonen. Os algoritmos genéticos tem sido usados como algoritmo de busca e otimização. Outro método de otimização muito usual é o *simulated annealing* (SA) [22]. Surgido no contexto da mecânica estatística, baseado no processo utilizado para fundir um metal, onde este é aquecido a uma temperatura elevada e em seguida é resfriado lentamente. Neste contexto, o processo de otimização é realizado por níveis, simulando os níveis de temperatura no resfriamento.

Uma vez que os problemas não linearmente separáveis são relativamente mais complexos de resolver, pode-se buscar um método alternativo baseado no Teorema de Cover [23]. O qual postula que expandindo a dimensionalidade do espaço de representação a probabilidade assintótica de classificações ambíguas irá decrescer. Assim, se esta dimensionalidade cresce o suficiente, o problema não linearmente separável, com alta probabilidade, pode tornar-se um problema linearmente separável. Partindo desse pressuposto, Sousa et al. [24] propuseram um método de preprocessamento de dados, em que, após esse processo, qualquer método de classificação utilizado poderá fornecer altas taxas de assertividade.

Uma vez que cada conjunto de dados tem características próprias, como dimensionalidade, número de classes, apresenta ou não sobreposição/não linearidade, o objetivo principal desta proposta é a criação de um pacote estatístico computacional para classificação, que considere tais caracterísiticas. A princípio, uma arquitetura de RNA que forneça a melhor taxa de acertos poderá ser obtida por meio de um algoritmo de busca, como AGs ou SA. A seguir, um algoritmo de busca e otimização será usado para apontar o método (RNA, *k-means*, kNN ou SVM) que melhor se adeque para classificação de tal conjunto de dados. E para problemas não linearmente separáveis, o preprocessamento dos dados será realizado antes da aplicação das metodologias de classificação mencionadas. Esperando, com isso, atingir altas taxas de acerto na classificação de qualquer conjunto de dados, das diversas áreas do conhecimento.

2. OBJETIVOS

Geral:

Criar e desenvolver um pacote estatístico computacional para classificação, com registro junto ao Instituto Nacional da Propriedade Industrial (INPI).

Específicos:

1. Estudar e aplicar os métodos de classificação para conjuntos de dados convencionais;

2. Estudar e aplicar algoritmos de busca e otimização;

3. Obter uma arquitetura de rede neural que forneça uma classificação com alta taxa de assertividade;

4. Aplicar algoritmos de busca e otimização para apontar o método mais adequado na classificação de cada conjunto de dados;

5. Tornar lineamermente separáveis, conjuntos de dados não linearmente separáveis;

6. Implementar o pacote estatístico computacional para classificação, em R.

3. METODOLOGIA

Para criação e desenvolvimento do pacote estatístico computacional para classificação de dados convencionais, iniciaremos o trabalho a partir de revisões da literatura associada ao formalismo de classificação. A revisão terá como base a bibliografia disponível a partir da CAPES, além de livros como [1, 17, 16, 19].

A fim de atingir os objetivos específicos, a metodologia é a seguinte:

Para o objetivo 1, serão estudados e aplicados os métodos *k-means*, kNN e SVM a conjuntos de dados obtidos de UCI *Machine Learning Repository* [25], bem como, conjuntos de dados disponíveis na base de dados do R-Program [26]. E para um melhor entendimento das RNAs, a classifciação a partir desse método também será realizada.

Para atingir o objetivo 2, serão estudados os algoritmos de busca e otimização *simulated annealing* e algoritmos genéticos.

O objetivo 3 será alcançado a partir da aplicação dos algortimos SA e AGs. Uma sugestão dos parâmetros a serem usados na RNA será os parâmetros dos outros métodos. Por exemplo, o kNN fornece como melhor taxa de assertividade um número k de vizinhos. Pensando numa MLP, esse número poderá ser usado como número de neurônios na camada escondida, ou número de camadas escondidas, ou outro parâmetro da RNA.

Ainda usando algoritmos de busca e otimização, o ojetivo 4 será alcançado quando, a partir da taxa de acertos, um método seja apresentado como mais adequado para o conjunto de dados em estudo.

No método proposto por [24] a dimensionalidade do conjunto de dados é dobrada. Em que cada número do conjunto de dados a ser classificado, é considerado como o módulo de um número complexo, gerando assim um par ordenado de números. A obtenção desse par ordenado tem sido por meio de um AG. O método SA também será implementado, a fim de comparar em termos de custo e desempenho.

Finalmente, a implementação do pacote será iniciada. Até esse momento, muito já terá sido desenvolvido e implementado no programa estatístico gratuito, R-Program [26].

4. PRINCIPAIS CONTRIBUIÇÕES CIENTÍFICAS, TECNOLÓGICAS OU DE INOVAÇÃO DO PROJETO

Pretende-se, ao longo da execução da proposta, a participação em pelo menos três eventos nacionais; Criação e desenvolvimento do pacote estatístico para classificação no programa R; Registro junto ao INPI; Utilização do pacote desenvolvido, para solução de problemas locais e regionais; E, publicação de pelo menos três artigos em periódicos de grande circulação.

5. CRONOGRAMA DE EXECUÇÃO DO PROJETO

Atividade	Trimestre											
	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º
Revisão de Literatura	X				X				X			
Estudar e aplicar métodos de classificação	X	X	X	X								
Estudar e aplicar algoritmos de busca e otimização					X	X	X	X				
Criação e registro do pacote estatístico computacional para classificação									X	X	X	X
Escrever artigo científico				X				X				X
Participar de eventos				X				X				X

- Revisão de Literatura: Para observância de novos métodos desenvolvidos para classificação de dados.
- Estudar e aplicar métodos de classificação: No primeiro ano de execução do projeto, a proposta é estudar e aplicar os métodos de classificação k-means, kNN, SVM e RNAs (MLP e Kohonen).
- Estudar e aplicar algoritmos de busca e otimização: Para o segundo ano, a atenção será voltada ao estudo dos algoritmos de busca e otimização. A princípio, serão abordados os métodos *simulated annealing* e os algoritmos genéticos. Entretanto, outros métodos poderão vir a ser explorados.
- Criação e registro do pacote estatístico computacional para classificação: Uma vez que o conhecimento dos métodos de classificação e otimização seja admitido, no terceiro ano, propõe-se criar e registrar o pacote estatístico computacional para classificação.
- Escrever artigo científico: Ao final de cada ano de execução do projeto, objetiva-se ter pelo menos um artigo científico escrito/submetido para periódicos de grande circulação.
- Participar de eventos: A participação em eventos será importante para o enriquecimento intelectual do bolsista, bem como, divulgação da pesquisa desenvolvida ou em andamento.

Referências

[1] DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification*. Second. New York: John Wiley & Sons, 2000.

[2] WU, X.; ZHU, X. Mining with noise knowledge: error-aware data mining. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 2008.

[3] SAEZ, J. A.; LUENGO, J.; HERRERA, F. Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification. *Pattern Recognition*, 2013.

[4] CANO, J.-R. Analysis of data complexity measures for classification. *Expert Systems with Applications*, 2013.

[5] VAPNIK, V. N. *Statistical Learning Theory*. New York: John Wiley & Sons, 1998.

[6] ELIZONDO, D. The linear separability problem: Some testing methods. *IEEE Transactions on Neural Networks*, 2006.

[7] MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 1967. p. 281–297.

[8] COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21–27, 1967. ISSN 0018-9448.

[9] CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, 1995.

[10] PRABHU, P.; ANBAZHAGAN, N. Improving the performance of k-means clustering for high dimensional data set. *International Journal on Computer Science and Engineering (IJCSE)*, v. 3, n. 6, p. 2317–2322, 2011.

[11] ZHANG, C.; FANG, Z. An improved k-means clustering algorithm. *Journal of Information & Computational Science*, v. 10, p. 193–199, 2013.

[12] HUANG, X. et al. Dskmeans: A new kmeans-type approach to discriminative subspace clustering. *Knowledge-Based Systems*, v. 70, p. 293–300, 2014.

[13] HUANG, Z.; ZHOU, Z.; HE, T. Associative classification with knn. *Journal of Theoretical and Applied Information Technology*, v. 49, 2013.

[14] XU, R.; II, D. W. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 2005.

[15] WU, X. et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 2008.

[16] BRAGA, A. P.; LUDERMIR, T. B.; CARVALHO, A. C. P. L. F. *Redes Neurais Artificiais Teoria e aplicações*. First. Rio de Janeiro: LTC, 2000.

[17] HAYKIN, S. *Redes Neurais Princípios e Prática*. Second. Porto Alegre: Bookman, 2001.

[18] RUMELHART, D. E.; HINTON, G.; WILLIAMS, R. Learning representations by back-propagating errors. *Nature*, v. 323, n. 11, p. 533–536, 1986.

[19] HOLLAND, J. H. *Adaptation in natural and artificial systems*. Ann Arbor: University of Michigan Press, 1975.

[20] SOUZA, R. B. et al. *An Intelligent Agent to Classify Countries Based on Financial Indices*. 2013.

[21] HUSKEN, M.; STAGGE, P. Recurrent neural networks for time series classi cation. *Neurocomputing*, v. 50, p. 223–235, 2003.

[22] KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by simulated annealing. *Science*, American Association for the Advancement of Science, v. 220, n. 4598, p. 671–680, 1983. ISSN 0036-8075. Disponível em: <https://science.sciencemag.org/content/220/4598/671>.

[23] COVER, T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Eletronic Computers*, EC-14, n. 3, p. 326–334, June 1965.

[24] SOUSA, R. B. d. et al. A proposal of quantum data representation to improve the discrimination power. *Natural Computing*, Feb 2019. ISSN 1572-9796. Disponível em: <https://doi.org/10.1007/s11047-019-09734-w>.

[25] FRANK, A.; ASUNCION, A. *UCI Machine Learning Repository*. 2013. Irvine, CA: University of California, School of Information and Computer Science. Disponível em: <http://archive.ics.uci.edu/ml>.

[26] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2015. Disponível em: <https://www.R-project.org/>.