Projeto de Iniciação Científica

Métodos de Otimização aplicados a redes neurais para detecção de anomalias em transações com cartão de crédito

Candidato: Frederico José Ribeiro Pelogia* **Orientador:** Luís Felipe Bueno†

Maio de 2019

Resumo

Neste projeto pretende-se estudar métodos de otimização, em especial algoritmos estocásticos, aplicados ao treinamento de redes neurais. Em um trabalho recente proposto por L. F. Bueno e J. M. Martínez, *On the complexity of solving feasibility problems*, 2018, foi apresentado um algoritmo de primeira ordem com bons resultados de complexidade para problemas de quadrados mínimos. Um dos principais pontos da pesquisa deste projeto será desenvolver uma versão estocástica desse algoritmo. Serão analisados os desempenhos dos algoritmos estudados quando aplicados à detecção de fraudes em operações de cartão de crédito utilizando a base de dados *Credit Card Fraud Detection* do Kaggle.

Palavras Chave: Otimização, Métodos Estocásticos, Redes Neurais, Deep Learning, Detecção de Anomalias, Detecção de Fraudes, Operações de Cartão de Crédito.

1 Introdução ao tema escolhido

Em várias situações práticas é importante que se identifique dados que estão em desacordo com o comportamento normal esperado. Estes dados geralmente são chamados de anomalias ou *outliers*, embora outras nomenclaturas possam ser utilizadas. A presença deste tipo de dados indica uma situação de interesse no processo pois pode representar desde um simples registro mal feito na base de dados até um comportamento que deve ser tratado de maneira diferente do procedimento padrão.

Uma revisão bem completa da literatura sobre a importância de se detectar anomalias é feita em [5]. Um ponto abordado, originalmente tratado em [1], é relativo a transações com cartão de crédito fora do histórico padrão de um cliente. Este tipo de anomalia pode indicar o uso indevido do cartão, resultante de fraude ou roubo, por exemplo. Algumas outras aplicações apresentadas incluem a atuação de hackers, detecção da presença de tumores e indicativo de falha em componentes de espaçonaves. Uma característica fundamental para se escolher a ferramenta a ser utilizada na

*Graduando do Bacharelado em Ciência e Tecnologia, Universidade Federal de São Paulo, Campus São José dos Campos
†Departamento de Ciência e Tecnologia, Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo, Campus São José dos Campos. Email: lfelipebueno@gmail.com

detecção de anomalias é a existência ou não de um conjunto de dados previamente rotulado onde cada instância é classificada como normal ou anômala.

Neste projeto esperamos focar nossa atenção em problemas de identificação de anomalias relacionadas ao sistema financeiro. Pretendemos estudar a base de dados *Credit Card Fraud Detection* do Kaggle. O Kaggle é uma plataforma usual da literatura, fundada em 2010 por Anthony Goldbloom e adquirida em 2017 pelo Google (Alphabet) [10]. A base de dados de interesse consiste em dados sobre transações de cartão de crédito rotuladas como fraudulentas ou não.

Técnicas de inteligência artificial têm sido utilizadas para para bases de dados rotuladas. Recentemente redes neurais profundas têm sido utilizadas com sucesso para classificar um grande volume de dados [7]. Matematicamente treinar uma rede neural pode ser formulada como o problema de otimização

$$\text{minimizar } f(x) \equiv \sum_{i=1}^{N_{dados}} \left(R_i(x)\right)^2, \tag{1}$$

onde $R_i(x)$ é a diferença entre o valor rotulado do dado i e o valor previsto pela rede neural para este mesmo dado, usando parâmetros x. Esta será a abordagem que será usada neste projeto para estudar a base de dados mencionada.

Geralmente o número de dados em (1) é muito grande e portanto avaliar a soma completa é computacionalmente muito custoso. Técnicas onde, em cada iteração, apenas uma parte amostral dos dados são utilizados, originam os chamados métodos estocásticos de otimização. Este tipo de abordagem tem sido comum neste ramo de aplicações, veja, por exemplo [7]. Além disso, em redes neurais profundas, o número de variáveis e de composições de funções é muito grande. Isso faz com que métodos de otimização que usem apenas informação de primeira ordem sejam mais eficientes para resolver o problema (1).

O estudo de redes neurais será feito com base no livro recém publicado pelo Prof. Weldon Lodwick (Colorado University) e coautores [4] e também pelo livro específico sobre *deep learning* [7]. O prof. Weldon Lodwick está como professor visitante da Universidade Federal de São Paulo e será um colaborador importante deste projeto. Dois artigos [8, 9], cujos estudos são utilizados no livro [4], são referentes à identificação de anomalias no sistema de saúde italiano. Estes artigos também serão estudados durante este projeto.

Nesta mesma linha o orientador escreveu dois artigos científicos [2, 6], em fase de revisão em congressos, sobre métodos estocásticos de otimização aplicados em redes neurais. O primeiro destes artigos é em coautoria com Kléber A. Benatti e Tiago S. de Nazaré e o segundo com Luiz Felipe S. dos Santos, todos funcionários do Itaú-Unibanco. Estes tipos de métodos devem ser estudados pelo aluno durante o projeto, bem como os métodos estocásticos mais utilizados na literatura segundo [7].

Recentemente o orientador, em conjunto com J. M. Martínez (UNICAMP), disponibilizou um relatório técnico, que pode ser encontrado em [3], sobre métodos de primeira ordem para resolver problemas do tipo (1). Nesse trabalho é mostrado que o método proposto tem complexidade $O(\epsilon^{-2})$, que é o resultado ótimo para algoritmos para otimização sem restrições. Além disso, é mostrado um resultado inédito de complexidade $O(-\log(\epsilon))$ para problemas em que o jacobiano de $R(x)$ tem posto completo. Testes computacionais mais robustos estão sendo elaborados pelo prof. E. Birgin (USP), para que o trabalho seja submetido a alguma revista de boa reputação internacional. Seria interessante pensar quais dos resultados poderiam ser adaptados a problemas de redes neurais onde, obviamente, a hipótese de posto completo não é satisfeita. A verificação do desempenho de uma versão

estocástica do método apresentado em [3] seria interessante e será a principal contribuição científica do candidato ao longo do projeto.

2 Justificativa

Os pontos de pesquisa deste projeto são de alto interesse direto da comunidade científica de otimização, tendo sido abordada nos últimos anos por diversos pesquisadores de renome internacional. Além disso, a própria aplicação de identificação de fraudes em cartões de crédito é muito relevante para várias instituições financeiras do país, que podem estender o estudo para suas bases de dados reais e confidenciais. Mais ainda, os métodos estudados e desenvolvidos podem ser utilizados em diversas outras situações, tais como determinação linhas de crédito, decisões em transações cambiais, identificação de ataques cibernéticos, etc... Esperamos contribuir com este processo sobretudo em parceria com o Sr. Mateus Polizeli, gerente de soluções analíticas na área de gestão de riscos do Itaú-Unibanco e aluno de mestrado do Programa de Pesquisa Operacional da UNIFESP-ITA, que tem trabalhado com técnicas estatísticas de detecção de anomalias em conjunto com professor orientador deste projeto. Desta forma o projeto pretende alcançar impactos científicos interessantes, bem como contribuir para o retorno mais direto para a sociedade da pesquisa científica.

3 Objetivos

O objetivo principal deste projeto é a capacitação humana em uma área em que o Brasil tem pouquíssimas pessoas qualificadas. Para isso esperamos obter implementações eficientes de métodos de otimização aplicados a redes neurais profundas. Pretendemos que os algoritmos implementados consigam identificar satisfatoriamente os dados fraudulentos da base de dados *Credit Card Fraud Detection* do Kaggle. Por fim, temos como objetivo específico central propor versões estocásticas de algoritmos que, baseados em [3], sejam competitivos contra as principais referências da literatura. Possivelmente obteremos algum resultado de convergência em probabilidade e/ou de complexidade neste caso, embora isso seja bastante ambicioso para o primeiro ano do projeto. Mesmo que isso não seja possível, o objetivo de ganhar a familiaridade com este tópico seria muito importante para pesquisas futuras mais profundas.

4 Cronograma e plano de trabalho das atividades

O cronograma das atividades a serem realizadas pelo aluno está listado a seguir.

1. Agosto de 2019: Revisão bibliográfica sobre detecção de anomalias tendo como base o artigo [5].
2. Setembro e outubro de 2019: Estudo de redes neurais profundas utilizando os livros [4] e [7].
3. Setembro a dezembro de 2019: Estudo e implementação dos métodos estocásticos mais utilizados na literatura segundo [7].
4. Novembro e dezembro de 2019: Estudo dos artigos [2, 6], sobre métodos estocástico aplicados a redes neurais e [8, 9], sobre redes neurais aplicadas à detecção de fraudes.

Metas	Meses											
	2019 Ago	2019 Set	2019 Out	2019 Nov	2019 Dez	2020 Jan	2020 Fev	2020 Mar	2020 Abr	2020 Mai	2020 Jun	2020 Jul
1	X											
2		X	X									
3		X	X	X	X							
4				X	X							
5					X	X	X					
6							X					
7							X	X	X			
8									X			
9										X	X	X
10											X	
11												X

Tabela 1: Tabela com o cronograma proposto

5. Dezembro de 2019 a fevereiro de 2020: Familiarização com a base de dados *Credit Card Fraud Detection* do Kaggle e testes numéricos para detecção de fraudes usando os algoritmos estudados e implementados nas etapas anteriores.
6. Fevereiro de 2020: Produção e envio do 1º Relatório Científico de Progresso para submissão à FAPESP.
7. Fevereiro a abril de 2020: Início do estudo do artigo [3], desenvolvimento de métodos estocásticos associados e testes numéricos referentes aos mesmos.
8. Abril de 2020: Sintetização dos resultados obtidos e redação do resumo para o XXVIII Congresso de Iniciação Científica da UNIFESP.
9. Maio a julho de 2020 : Conclusão do desenvolvimento e testes numéricos de métodos estocásticos associados ao artigo [3].
10. Junho de 2020: Preparação e apresentação do projeto no XXVIII Congresso de Iniciação Científica da UNIFESP.
11. Julho de 2020: Redação do Relatório Científico Final a ser enviado à FAPESP.

5 Materiais e métodos

A metodologia a ser utilizada neste projeto é usual da área. O aluno apresentará semanalmente ao orientador os assuntos estudados, os testes numéricos realizados e uma análise parcial dos resultados obtidos. Reuniões com o Prof. Dr. Weldon Lodwick devem ocorrer a cada 2 meses para adaptarmos o andamento do projeto às suas sugestões. Relatórios parciais serão compartilhados com o Sr. Mateus Polizeli, gerente de soluções analíticas na área de gestão de riscos do Itaú-Unibanco, para que tenhamos um retorno sobre a pertinência da pesquisa para o setor produtivo. Como usual, os dados do *Credit Card Fraud Detection* do Kaggle são divididos em um conjunto de testes e um conjunto de validação. Os temas serão estudados em livros e artigos da área e os testes computacionais serão implementados em Python.

6 Forma de Análise dos Resultados

Os resultados do estudo teórico do aluno serão analisados nas reuniões semanais com o professor orientador e com seminários para nossos parceiros, Prof. Weldon Lodowick e Sr. Mateus Polizeli. Os resultados computacionais dos métodos clássicos de otimização estocástica serão validados em problemas tradicionais da literatura. As versões estocásticas de algoritmos que serão propostas com base em [3] vão ser comparadas com métodos clássicos de otimização estocástica tanto em relação à qualidade da solução encontrada quanto ao desempenho computacional do algoritmo. A principal análise de interesse é quantificar o grau de precisão ao identificar os dados fraudulentos da base de dados *Credit Card Fraud Detection* do Kaggle.

Referências

[1] Emin Aleskerov, Bernd Freisleben e Bharat Rao. "CARDWATCH: A neural network based database mining system for credit card fraud detection". Em: *Proceedings of IEEE/IAFE Conference on Computational Intelligence for Financial Engineering*. 1997, pp. 220 –226. DOI: 10.1109/CIFER.1997.618940.

[2] K. A. Benatti, T. S. Nazaré e L. F. Bueno. *O método de Levenberg-Marquardt estocástico aplicado à redes neurais artificiais*. Artigo submetido. 2019.

[3] L. F. Bueno e J. M. Martínez. *On the complexity of solving feasibility problems*. Rel. técn. 2018. URL: http://www.optimization-online.org/DB_HTML/2018/11/6929.html.

[4] M. Buscema, W. A. Lodwick, M. Breda, M. Guilia, F. Newman e M. Asadi. *Artificial Adaptive Systems Using Auto Contractive Maps: Theory, Applications and Extensions*. New York: Springer-Verlag, 2018. DOI: https://doi.org/10.1007/978-3-319-75049-1.

[5] Varun Chandola, Arindam Banerjee e Vipin Kumar. "Anomaly Detection: A Survey". Em: *ACM Comput. Surv.* 41.3 (2009), 15:1–15:58. DOI: 10.1145/1541880.1541882. URL: http://doi.acm.org/10.1145/1541880.1541882.

[6] L. F. Dos Santos e L. F. Bueno. *Stochastic Variance Reduction with Adaptive Optimization Methods*. Artigo submetido. 2019.

[7] I. Goodfellow, Y. Bengio e A. Courville. *Deep Learning*. Massachussets: MIT press, 2016.

[8] F. S. Mennini, L. Gittoa, Russoa S., A. Cicchetti, M. Ruggeri, S. Coretti, G. Maurelli e P. M. Buscema. "Artificial neural networks and their potentialities in analyzing budget health data: an application for Italy of what-if theory". Em: *Quality & Quantity: International Journal of Methodology* 51.3 (2016), pp. 1261–1276.

[9] F. S. Mennini, L. Gittoa, Russoa S., A. Cicchetti, M. Ruggeri, S. Coretti, G. Maurelli e P. M. Buscema. "Does regional belonging explain the similarities in the expenditure determinants of Italian healthcare deliveries? An approach based on Artificial Neural Networks". Em: *Economic Analysis and Policy* (2017).

[10] Thiago Ribeiro. *Kaggle - O que é? Como funciona?* 2018. URL: http://tirandolicoesdetudo.com.br/kaggle-o-que-e-como-funciona.