

Universidade Federal de Pernambuco

Graduação em Ciência da Computação

Centro de Informática

2015.1

Análise de dados públicos de saúde com redes SOM **Proposta de Trabalho de Graduação**

Aluno: Vítor Hugo Antero de Melo (vham@cin.ufpe.br)

Orientador: Germano Crispim Vasconcelos (gcv@cin.ufpe.br)

Contexto

Nos últimos anos vimos a disponibilização de grandes quantidades de dados para o público, até mesmo por parte de governos. Esta oferta cada vez maior de dados impulsiona o uso de técnicas de extração de conhecimento, mineração de dados e *Business Intelligence*. Apesar dos avanços nessas áreas, ainda presenciamos graves ineficiências em serviços públicos, sonegações de impostos, fraudes em planos de saúde, cartões de crédito e seguros, para citar alguns ([1], [2]).

Aderindo à tendência, todas as esferas governamentais do Brasil disponibilizam dados abertos de vários departamentos, ministérios, secretarias e programas. Grande parte dos dados abertos do Brasil são disponibilizados no Portal Brasileiro de Dados Abertos [3].

Dados abertos raramente são classificados de alguma forma, o que implica que qualquer técnica a ser utilizada neles deve ser própria para problemas que envolvem aprendizagem não supervisionada. Abordagens não supervisionadas, por exemplo, podem ser utilizadas para detecção de *outliers* e análise de clusters. Aplicar e estudar tais abordagens em bases de dados ricas é importante tanto para fins técnicos quanto sociais.

Objetivo

Acima foi visto a necessidade de aplicar e analisar técnicas de análise de dados em bases públicas. Em [4] os autores relatam casos de uso de abordagens não supervisionadas em análise de fraudes em várias áreas. Neste trabalho, os autores citam vários outros que lançaram mão de *Self-Organizing Maps* (SOM) [5] para realizar a segmentação dos dados. A rede SOM provê uma ótima maneira de reduzir a dimensionalidade dos dados, enquanto mantém as relações topológicas no conjunto de treinamento.

O objetivo deste trabalho é aplicar e analisar a rede SOM para realizar uma clusterização da base de dados de Autorização de Internações Hospitalares, disponibilizada pelo DATASUS em [3]. O trabalho seguirá o *workflow* ditado pelo modelo de processo em mineração de dados CRISP-DM, o mais seguido na área [6].

Inicialmente será feita uma revisão na literatura da área, em busca de utilização de redes SOM e técnicas de tratamento de dados. Posteriormente, será feito o tratamento da base e a aplicação das técnicas nesta base. Por fim, os resultados serão analisados seguindo indicadores de eficiência indicados por organizações de saúde e/ou especialistas.

Cronograma

[illegible]

Possíveis Avaliadores

Paulo Jorge Leitão Adeodato

Frederico Luiz Gonçalves de Freitas

Referências

- [1] Ng, K.S. et al, "Detecting Non-compliant Consumers in Spatio-Temporal Health Data: A Case Study from Medicare Australia," Data Mining Workshops (ICDMW), 2010 IEEE International Conference on (2010), 613-622.
- [2] Raj, S.B.E.; Portia, A.A., "Analysis on credit card fraud detection methods," Computer, Communication and Electrical Technology (ICCCET), 2011 International Conference on (2011), 152-156.
- [3] Portal Brasileiro de Dados Abertos. Disponível em <http://dados.gov.br/>. Acessado em abril de 2015.
- [4] C. Phua, V. Lee, K. Smith, R. Gayler, "A comprehensive survey of data mining-based fraud detection research," Artificial Intelligence Review (2005) 1–14.
- [5] Kohonen, T., "The self-organizing map," Proceedings of the IEEE (1990), 1464-1480.
- [6] CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Disponível em <http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>. Acessado em abril de 2015.

Assinaturas

Recife, Pernambuco. Brasil

23 de Abril de 2015

Germano Crispim Vasconcelos

(orientador)

Vítor Hugo Antero de Melo

(proponente)