Nome do autor

Características inerentes a medidas de centralidade e uso de algoritmos de aprendizado de máquina para classificação de *bridging nodes*

Uberlândia

2018

Nome do autor

Características inerentes a medidas de centralidade e uso de algoritmos de aprendizado de máquina para classificação de *bridging nodes*

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Uberlândia, como requisito parcial para obtenção do grau de Mestrado em Ciência da Computação.
Área de concentração: Computação.
Linha de pesquisa: Inteligência Artificial.

Universidade Federal de Uberlândia

Programa de Pós-Graduação em Ciência da Computação

Orientador: Prof. Dr. Anderson Rodrigues dos Santos

Uberlândia

2018

Resumo

As redes de interação proteína-proteína (PPI), com frequência, comportam números expressivos de nós (proteínas) e arestas (interações) na ordem dos milhares, possivelmente milhões. Os nós promissores em redes PPI de grande porte, passíveis a serem utilizados na produção de fármacos, podem ser identificadosatravés de métodos exatos como bridging centrality, no entanto, isto pode se tornar um problema computacional a ser superado devido à complexidade destas redes. Como alternativa, se sugere o uso de Inteligência Artificial, sendo o objetivo desta pesquisa analisar algoritmos de aprendizado de máquina (ML) aplicadas ao problema de determinação de bridging centrality em redes PPI, modeladas por meio de Rede Complexa, e identificar o algoritmo que ofereça resultados próximos ao gerado pelo algoritmo exato tido como referência, mas com esforço computacional menor. Foram selecionadas redes PPI de nove diferentes bactérias, sobre as quais foi gerado um conjunto de métricas estruturais usando o software Gephi. Em seguida, cada arquivo de PPI contendo as métricas geradas foi submetido à análise de 15 algoritmos selecionados para a ML, disponíveis no software WEKA. Os arquivos de métricas de predição foram submetidos ao melhor modelo preditivo identificado e, a seguir, os nós foram classificados em weak ou strong. Por fim, houve a avaliação do desempenho do classificador, utilizando-se o software R e o pacote ROCR, obtendo-se a curva ROC, o valor Area Under the Curve (AUC), a acurácia e o threshold correspondentes. Os melhores modelos de aprendizagem identificados foram gerados pelos algoritmos *Bagging* e *Random Forest*, e os piores foram gerados pelos algoritmos *Naïve-Bayes* e *OneR*. Em termos gerais, a acurácia média da predição foi de $74,38\% \pm 5,84\%$, com limiar médio de 96%, e AUC médio de $65,09\% \pm 4,48\%$. Os nós preditos corretamente pelo classificador foram, em média, $24,33\%$ sendo $2,75\%$ verdadeiros positivos e $21,58\%$ verdadeiros negativos. Por outro lado, $75,66\%$ foram incorretamente classificados, sendo $23,67\%$ falsos positivos e $51,99\%$ falsos negativos. Comparando os $2,75\%$ verdadeiros positivos com os identificados pelo algoritmo exato, obteve-se uma taxa de acerto médio de 77,16% $\pm 20,23\%$.$Oresultadopreditivogeradopeloprocessode MLaproximou-sedoobtidopeloalgoritmoexato, apresentandoeficcia, noentanto, comconsiderveltaxadeerro.Dessaf$

Palavras-chave: PPI. Bridging Centrality. Aprendizado de Máquina. Redes Complexas

Abstract

Protein-protein interaction (PPI) networks often carry expressive numbers of proteins and interactions in the order of thousands, possibly millions. The promising nodes in large PPI networks, which can be used in drug production, can be identified through exact methods such as bridging centrality, however, this can become computationally expensive to be overcome due to the complexity of these networks. As an alternative, the use of machine learning is suggested. The objective of this study was analyzed machine learning algorithms (ML) applied to the problem of determining bridging centrality in PPI networks, modeled by Complex Network, and identify the algorithm which offers results closest to generated by the exact algorithm taken as a reference, but with less computational effort. PPI networks were selected from nine different bacteria, on which a set of structural metrics were generated using Gephi. Then, each PPI file containing the generated metrics was submitted to the analysis of 15 algorithms selected for ML available in the WEKA. The prediction metrics files were submitted to the best predictive model identified and then the nodes were classified as weak or strong. Finally, the performance of the classifier was evaluated using the R software and the ROCR package, obtaining the ROC curve, the area under ROC curve (AUC), the corresponding accuracy and threshold. The best learning models identified correspond to the algorithms Bagging and Random Forest and the worst were the NaiveBayes and OneR. In general terms, the mean prediction accuracy of the nodes of the PPIs was $74.38\% \pm 5.84\%$, with a mean threshold of 96%, and mean AUC of $65.09\% \pm 4.48\%$. The nodes correctly predicted by the classifier were, on average, 24.33%, with 2.75% true positive and 21.58% true negative. On the other hand, 75.66% were incorrectly classified, being $23,67\%$ false positive and 51.99% false negative. Comparing the 2.75% true positive with those identified by the exact algorithm, an average hit rate of $77.16\% \pm 20.23\%$ was obtained. The predictive result generated by the ML process approached that obtained by the exact algorithm, presenting efficacy, however, with a considerable error rate. Thus, our results corroborate the literature knowledge about the use of ML in complex networks, that is, ML algorithms applied to complex centrality measures such as bridging centrality are not effective. The plugin implemented as one of the products of this work for GEPHI software, version 0.9.1 or higher, is available on sourceforge.net under the name of BridgingCentralityPlugin.

Keywords: PPI. Bridging Centrality. Machine Learning. Complex Networks

Lista de abreviaturas e siglas

ABNT	Associação Brasileira de Normas Técnicas

Sumário

1 Introdução

1.1 Motivação

1.2 Objetivos

1.2.1 Objetivo geral

Apresentação do objetivo geral.

1.2.2 Objetivos específicos

- objetivo 1;
- objetivo 2;
- objetivo 3.

2 Como utilizar recursos do limarka

Consulte o wiki do projeto: https://github.com/abntex/limarka/wiki

Cada capítulo inicia automaticamente em página ímpar (em conformidade com as Normas). Por isso que existem várias páginas em branco nesse documento.

2.1 Como citar e referenciar

O arquivo de referências é configurado em "configuracao.pdf", utilize-o para gerenciar suas referências.

Veja um exemplo de citação direta e referenciação a seguir:

> A 'norma' 6023:2000 (2) é complicada e cheia de inconsistências. Jamais será possível gerar um estilo bibtex totalmente consistente com a 'norma', até porque nem a 'norma' é compatível com ela mesma. Um bom estilo bibliográfico deve ter uma linha lógica para formatação de referências. Assim, com alguns poucos exemplos, qualquer pessoa poderia deduzir os casos omissos. Nesse sentido, a 'norma' 6023 trafega pela contra-mão. É quase impossível deduzir sua linha lógica. O problema mais grave, no entanto, fica pela maneira de organizar nomes. A ABNT quebrou o sobrenome em duas partes. Normalmente se fala apenas em "*last name*", mas agora temos o "*last last name*" graças à ABNT. (ARAUJO, 2015, p. 5).

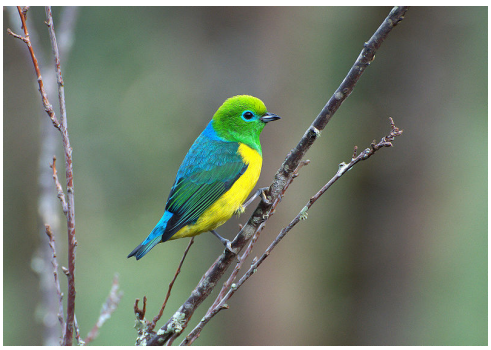
Consulte o documento Araujo (2015) para conhecer como referenciar os conteúdos.

2.2 Como inserir imagens

Por exemplo, a Figura 1 mostra um pássaro que possui as cores da bandeira do Brasil.

As imagens são inseridas o mais próximo possível do texto que as referenciam.

Figura 1 – Pássaro com as cores da bandeira do Brasil



Fonte: Alexandre (2016)

Referências

ALEXANDRE, E. de S. M. *Limarka*. 2016. Página do projeto. Disponível em: <https://github.com/abntex/limarka>. Acesso em: 5 de set 2016. Citado na página 14.

ARAUJO, L. C. *O pacote abntex2cite*: Estilos bibliográficos compatíveis com a abnt nbr 6023. [S.l.], 2015. Disponível em: <http://www.abntex.net.br/>. Citado na página 13.

Apêndices

APÊNDICE A – Primeiro apêndice

Quisque facilisis auctor sapien. Pellentesque gravida hendrerit lectus. Mauris rutrum sodales sapien. Fusce hendrerit sem vel lorem. Integer pellentesque massa vel augue. Integer elit tortor, feugiat quis, sagittis et, ornare non, lacus. Vestibulum posuere pellentesque eros. Quisque venenatis ipsum dictum nulla. Aliquam quis quam non metus eleifend interdum. Nam eget sapien ac mauris malesuada adipiscing. Etiam eleifend neque sed quam. Nulla facilisi. Proin a ligula. Sed id dui eu nibh egestas tincidunt. Suspendisse arcu.

APÊNDICE B – Segundo apêndice

Nunc velit. Nullam elit sapien, eleifend eu, commodo nec, semper sit amet, elit. Nulla lectus risus, condimentum ut, laoreet eget, viverra nec, odio. Proin lobortis. Curabitur dictum arcu vel wisi. Cras id nulla venenatis tortor congue ultrices. Pellentesque eget pede. Sed eleifend sagittis elit. Nam sed tellus sit amet lectus ullamcorper tristique. Mauris enim sem, tristique eu, accumsan at, scelerisque vulputate, neque. Quisque lacus. Donec et ipsum sit amet elit nonummy aliquet. Sed viverra nisl at sem. Nam diam. Mauris ut dolor. Curabitur ornare tortor cursus velit.

Morbi tincidunt posuere arcu. Cras venenatis est vitae dolor. Vivamus scelerisque semper mi. Donec ipsum arcu, consequat scelerisque, viverra id, dictum at, metus. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut pede sem, tempus ut, porttitor bibendum, molestie eu, elit. Suspendisse potenti. Sed id lectus sit amet purus faucibus vehicula. Praesent sed sem non dui pharetra interdum. Nam viverra ultrices magna.

Aenean laoreet aliquam orci. Nunc interdum elementum urna. Quisque erat. Nullam tempor neque. Maecenas velit nibh, scelerisque a, consequat ut, viverra in, enim. Duis magna. Donec odio neque, tristique et, tincidunt eu, rhoncus ac, nunc. Mauris malesuada malesuada elit. Etiam lacus mauris, pretium vel, blandit in, ultricies id, libero. Phasellus bibendum erat ut diam. In congue imperdiet lectus.