APLICAÇÃO DE TÉCNICAS DE MACHINE LEARNING PARA A IDENTIFICAÇÃO DE PADRÕES EM SEQUÊNCIAS DE DNA

PALMA, Wallace Pannace;[1] ZALEWSKI, Willian[2]

RESUMO

O estudo do sequenciamento de DNA possibilita o desenvolvimento de novos métodos para diagnosticar doenças e para a formulação de novos medicamentos. Os genes codificadores de proteínas constituem a parte funcional do DNA, onde estão as instruções biológicas básicas para a expressão gênica (produção de proteínas), são compostos por segmentos introns, sequências não codificadoras, e exons, sequências que serão traduzidas para proteínas. Os avanços na automação do sequenciamento de DNA possibilitaram a criação de grandes quantidades de dados, gerando uma alta demanda por métodos de análise mais eficientes. Técnicas computacionais como aprendizado de máquina (AM) oferecem a possibilidade de processar grandes quantidades de dados através de diversas formas de análise. Nesse contexto, neste trabalho temos por objetivo utilizar a técnica de Ensemble de classificadores para combinar diferentes preditores e um sistema de pesos para a identificação de regiões introns e exons de sequências genéticas. Para alcançar este objetivo empregamos o uso da linguagem de programação Python 3.0, por meio da qual utilizamos os algoritmos de AM disponibilizados gratuitamente pela biblioteca scikit-learn em conjunto com nossas próprias linhas de código. O método proposto neste trabalho consiste em combinar classificadores por meio da técnica votação que a partir de um conjunto de preditores de hipótese única. Cada preditor classifica uma dada sequência atribuindo um peso, baseado no desempenho do classificador no conjunto de treinamento Assim a sequência é classificada na classe que acumulou mais pontos entre os preditores. Para validar o método proposto trabalhamos com a base de dados Molecular Biology (Splice-junction Gene Sequences) Data-Set, que contém 3190 subsequências de DNA com 60 nucleotídeos cada, classificadas em que tipo de junção elas são intron-exon (IE) (767 sequências), exon-intron (EI) (768 sequências) ou nenhuma (NE) (1655 sequências). Como resultado da avaliação experimental, pudemos obter um aumento de precisão no desempenho do votador que construímos (com precisão nas classificações de 91% para classe EI, 92% para classe IE e 94% para a classe NE), em comparação ao classificador que não utiliza a ideia dos pesos personalizados disponibilizados pela biblioteca utilizada (com precisão nas classificações de 93% (EI), 90% (IE) e 88% (NE)). Considerando a avaliação experimental realizada neste estudo , observamos que nossa abordagem demonstrou resultados interessantes para o conjunto de dados estudado. A utilização de diferentes técnicas de predição combinadas para a classificação possibilitou maior desempenho em termos da taxa de precisão. Mais especificamente, ressalta-se que a combinação de preditores obteve melhores resultados que qualquer um dos preditores avaliados individualmente.

Palavras-chave: aprendizado de máquina, dna, programação, classificação, sequências.

REFERÊNCIAS

[1] DUA, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science

[2] DIETTERICH, Thomas G. (2002). Ensemble Learning. Corvallis, Oregon USA: Oregon State University, Department of Computer Science

[3] ZALEWSKI, Willian. Modelagem Simbólica de Padrões Morfológicos para a Classificação de Séries Temporais. Curitiba, PR, 2015.

AGRADECIMENTOS

Agradeço ao meu orientador Willian Zalewski que me auxiliou em todo o projeto e no meu aprendizado das ferramentas utilizadas, do domínio estudado e do método científico. Agradeço à PRPPG-UNILA pelo financiamento da bolsa que possibilitou maior dedicação ao projeto.

[1] Estudante do Curso de Engenharia Física, – ILACVN – UNILA; bolsista IC-UNILA. E-mail: wpd.palm.2016@aluno.unila.edu.br ;

[2] Docente da área de Ciência da Computação. – ILATIT – UNILA. Orientador de bolsista IC-UNILA. E-mail: willian.zalewski@unila.edu.br.