SELEÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA DIAGNÓSTICO DE CÂNCER

Aluna: Ariel A. dos Santos Nascimento

Orientador: Prof. André C. Ponce de Leon F. de Carvalho

Universidade de São Paulo

arielanny@usp.br

Objetivos

Uma questão emergente dentro do campo da bioinformática é: “Como podemos fazer uso do aprendizado de máquina para compreender melhor doenças como o câncer? ”[1][2][3][5]. Estudos prognósticos[2][3] vêm ganhando força nas últimas duas décadas para tentar detectar a doença, ou verificar a possibilidade da doença[2] antes que ela ocorra.

Infelizmente não é o caso para todos, por isso o objetivo desse projeto é de classificar o tipo do câncer para ajudar no diagnóstico dos pacientes. Por exemplo, o primeiro câncer estudado neste trabalho é o câncer de mama, e dado um microarray dos genes das células, queremos descobrir dentre 'basal', 'luminal B', 'HER', 'luminal A' ou 'cell_line', qual deles é o que acometeu o paciente. A meta principal é achar através de experimentos qual a melhor seleção de algoritmos de aprendizado de máquina para a classificação de câncer.

Métodos e Procedimentos

Para o projeto, foi feito uso da linguagem de programação Python, principalmente em cima das bibliotecas numpy, pandas e scikit-learn[7]. E, como parte da bolsa concedida pela Intel®, serão adicionadas ferramentas de otimização da Intel® AI Analytics Toolkit dentro do espaço do Jupyterlab.
Até, foram usados apenas dados de Câncer de Mama fornecidos pela 'Curated Microarray Database' (CuMiDa)[6]. Em etapas finais, os modelos feitos serão aplicados em dados de Leucemia e Câncer Cerebral, também de mesma origem.
A primeira parte do projeto é aplicar modelos padrão de ‘Random Forests’, ‘K-Nearest Neighbours’ (KNN) e ‘Support Vector Machines’ (SVM) em dados de microarray para ver sua performance, depois estudar a possibilidade de tuning desses algoritmos por ‘GridSearchCV’, para aumentar sua acurácia e testes com o método de validação ‘Cross Validation’.
O seguinte passo aplicará técnicas de redução de dimensionalidade aos dados por meio de ‘Principal Component Analysis’ (PCA), os dados de câncer de mama, por exemplo, têm 54677 atributos para 151 amostras. Na conclusão do trabalho serão feitas comparações da eficiência do código atual com as otimizações do Intel® AI Analytics Toolkit

Resultados

Os resultados dos experimentos até então estão de acordo com a seguinte tabela

Câncer de Mama			
Modelo	Tipo de Experimento	Accuracy	F1 Score
Random Forests	Padrão	90,32	89,93
Random Forests	Cross Validation	95,35 (média)	-
K-Nearest Neighbours	Padrão	83,87	83,22
K-Nearest Neighbours	Cross Validation	86 (média)	-
Support Vector Machines	Padrão	70,97	66,76
Support Vector Machines	Cross Validation	74,86 (média)	-
Random Forests	GridSearchCV	98,33 (melhor pontuação)	-
K-Nearest Neighbours	GridSearchCV	91,666 (melhor pontuação)	-
Support Vector Machines	GridSearchCV	96,667 (melhor pontuação)	-

Figura 1: Resultados iniciais

Agradecimentos

Agradeço à Intel® e a FAFQ pela oportunidade que me concederam de uma bolsa. Também ao meu orientador por ter me guiado pelo projeto quando precisei. Sobretudo, gostaria de agradecer minha família e meu parceiro queestiveram ao meu lado sempre me apoiando.

Conclusões

Baseado nos resultados iniciais, temos o algoritmo Random Forests vencendo em acurácia em todos os experimentos. Porém, sua pontuação alta pode sugerir um caso de 'overfitting' que será investigado com o avanço do trabalho. Os outros testes também são fortes, apesar da pontuação menor, podem estar sendo mais gerais que o Random Forests, que é uma das características que buscamos quando temos um modelo de predição como os estudados. Ainda é muito cedo para apontar se a combinação X ou Y é a melhor, esperamos ter uma direção concreta ao final do trabalho.

Referências Bibliográficas

[1] Support vector machines combines with feature selection for breast cancer. Akay, M.F.. 2009. Cukurova University.
[2] Applications of Machine Learning in Cancer Prediction and Prognosis. Cruz, J.A.; Wishart, D. S.. 2006. University of Alberta Edmonton.
[3] Machine learning applications in cancer prognosis and prediction. Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M. V.; Fotiadis, D.I. 2015. University of Ioannina.
[4] Multi-Objective Parameter Selection for Classifiers. Mussel, C.; Lausser, L; Maucher, M; Kestler, H.A.. 2021. University of Ulm.
[5] Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. Yue, W.; Wang, Z.; Chen, H.; Payne, A.; Liu X.. 2018. School of Mathematics, Southeast University, Nanjing 210096, China.
[6] Feltes, B.C.; Chandelier, E.B.; Grisci, B.I.; Dorn, M. CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research. Journal of Computational Biology, 2019.
[7] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.