A Novel Machine Learning Approach for Determining the Confounding Factors for Cancer Identification: An Integration of Neural Learning and Decision Tree

Ghosh, Shinjini (School: South Point High School)

Cancer is a heterogeneous disease comprising different subtypes. Its occurrence alters clinical parameters in human body. The exact cause being unknown, various internal factors, demographic and lifestyle related issues are found responsible. Identification of a set of confounding factors of cancer can be of immense help for early diagnosis. In this project, we try to automatically identify the major factors causing cancer using machine learning. However, paucity of cancerous samples is the chief bottleneck. We design a novel neural network and decision tree based machine learning algorithm that can identify the set of main causal factors of cancer yielding more than 95% detection accuracy. This helps predict occurrence of cancer in patients. Initially a generative restricted Boltzmann Machine (RBM), a supervised deep neural network, is used to model the bigger (non-cancerous) class. Functioning as one-class classifier, it confidently determines the non-cancerous samples (using the principle of free energy) to be removed from the dataset. We also devise a heuristic based on inter-intra class distance ratio to prune noisy samples. The resultant near-balanced set is used to find out the causal factors using decision tree. The built-in tree is later used for cancer detection. A multilayer perceptron based similar system is also designed. Experiments are carried out with some benchmark datasets. The system is compared with existing Tomek link based pruning and 1-NN classifier for handling imbalanced data. An average increase in accuracy of 5%-50% is noticed. Moreover, the identified factors corroborate to medical literature. To further elicit the robustness of the system, experiments are conducted on datasets of other diseases and the enhancement is 5%-25%.

Awards Won:
Fourth Award of $500