Supervised Machine Learning in the Personalized Assessment of the Risk of Breast Cancer

Naphtali Rishe, Jeffrey Horstmyer, Mikhail Berman, Evgenia Cheremisina, Malek Adjouadi, Tajana Lucic
NSF I/UCRC-CAKE*

Aniket Bochare, Aryya Gangopadhyay, Yelena Yesha, Anupam Joshi, Yaacov Yesha, Michael Grasso
NSF I/UCRC-CHMPR*

Mary Brady
U.S. National Institute of Standards and Technology

Breast cancer is the most common form of cancer in women. Breast cancer comprises 22.9% of invasive cancers in women and 16% of all the female cancers. Currently, treatment decisions are based primarily on clinical parameters, with little use of genomic data. Our study takes into consideration the data of postmenopausal women of European descent and their single nucleotide polymorphism (SNP) information to assess the risk of developing breast cancer. We used various supervised machine learning and data mining techniques to generate a model for predicting risk of breast cancer using only genomic data. In this paper we propose an approach to select 9 best SNPs using various feature selection algorithms and evaluate binary classifiers performance. The machine learning model generated without the domain knowledge yields poor prediction results. We have evaluated the performance of a binary classifier by adding the domain knowledge of 11 SNPs into the training set and performing classification based on most informative features obtained from the feature selection technique. Our observations revealed that the machine learning model generated using both the domain knowledge and the feature selection technique performed slightly better compared to the naive approach of classification.

In this study we have used various data mining and supervised machine learning techniques for generating a prediction model capable of distinguishing between cases and controls for initial screening. We have statistically analyzed 3 different methods: *Naive SNP Selection Approach*, *Feature Selection Approach* and *Domain Knowledge Integration Approach.* We have demonstrated the benefit of the addition of domain knowledge of SNPs in machine learning procedures.

* This material is based in part upon work supported by the U.S. National Science Foundation (NSF) under Grant Nos. CNS-0821345, CNS-1126619, HRD-0833093, IIP-0829576, CNS-1057661, IIS-1052625, CNS-0959985, OISE-1157372, IIP-1237818, IIP-1215201, IIP-1230661, IIP-1026265, IIP-1058606, IIS-1213026, OISE-0730065, CCF-0938045, CNS-0747038, CNS-1018262, CCF-0937964 at the NSF Industry-University Cooperative Research Centers CAKE and CHMPR, http://cake.fiu.edu