Classifiers for patients with breast cancer according to the neoadjuvant chemotherapy sensitivity

Pedro Kássio R. M. L. Carvalho, Thiago de Souza Rodrigues

CEFET-MG, CEFET-MG

The breast cancer is the most common tumor on women, about 508000 died in 2011 due to this disease. The treatment is usually done with neoadjuvant chemotherapy, followed by the operation to remove the tumor and after, the adjuvant chemotherapy. The neoadjuvant chemotherapy may show Complete Pathological Response (PCR), when the disease is completely eliminated, or, on the other hand, Residual Disease (RD). This project uses information about the molecular subtypes of breast cancer in order to classify the patients according to the chemotherapy sensitivity. Among the subtypes, the basal-like was not used, because of its difficult in classification problems. A dataset composed by gene expression of the patients, extracted from Gene Expression Omnibus repository, was used to create classifiers based on machine learning techniques, computational intelligence and evolutionary computation. Feature selection methods were applied in order to select the best characteristics to create the classifiers. From univariate feature selection method Volcano Plot, we selected 31 genes. From the multivariate feature selection method stepwise regression we selected 110 genes. And from the regression based on the Generalized Linear Model we selected 186 genes. Classifiers were created using different algorithms and the filtered data base. Six using neural networks with different types of training algorithms, one with particle swarm optimization with clustering and one with extreme learning machine algorithms. The neural network classifiers presented an average result of 52% accuracy. With the particle swarm optimization the best result was 62% of accuracy, using the 186 genic expressions. The best classifier was obtained using the Extreme Learning Machine algorithm, which has a very small runtime and 80% of accuracy on average, indicating a good result, which must also be adjusted to improve the hit rate. The genic expressions that showed this result were the 186. We can see that the extreme learning machine appears to be the most appropriate algorithm found for this problem and has the best runtime and results. It can still be improved to get a better result using fewer genic expressions, but has already shown a good initial result. This work is supported by CAPES, CNPq and FAPEMIG.