

Evaluation of ensemble classifier (EC) machine learning methods for introduction of breast cancer genomic biomarkers

L. Mirsadeghi^{a,b*}, K. Kavousi^b, A. Banaei-Moghaddam^b, R. Hajhosseini^a

^a Payame Noor University, Tehran, Iran

^b Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

*1.mirsadeghi@gmail.com

Abstract: The prognosis and diagnosis of cancers are two of crucial issues in precision oncology. Researchers are investigating the possibility of using ensemble machine learning tools to help detect cancers including breast cancer. This study reviews ensemble classifier methods and their applications in breast cancer detection and compares their results. At first, we surveyed 14 ensemble classifier machine learning methods in related to breast cancer prognosis and diagnosis. Then, we selected five of these approaches that prioritize genes and introduce breast cancer genomic biomarkers. They are include of Multiple RFE (Recursive Feature Elimination) [1], DECORATE (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples) [2], HyDRA (Hybrid Distance-score Rank Aggregation) [3], GenEnsemble for four base classifiers (NBS, IB3, SVM and C4.5) [4], and the collective approach for four individual methods (correlation, color palette, color proportion and SVM) [5]. This study can help to researchers in using ensemble methods in field of molecular biology and breast cancer diagnosis to choose most suitable ensemble method according to their problem. The results show the best solution depends on kind of problem, the structure of the available data and prior knowledge about related algorithm. Finally, using machine learning is less costly than bio-molecular testing. However, we suggest in order to provide better evaluation of accuracy of the results of this approaches, bio-molecular techniques such as qRT-PCR and NGS are applied. In the light of this research, we hope to help for introducing of novel driver genes as genomic biomarkers for precision oncology to increase patients' life expectancy.

Keywords: Genomic biomarkers; Breast cancer; Ensemble machine learning; Prognosis and diagnosis; Precision oncology

References

- [1] R. Yang, B. J. Daigle, L. R. Petzold, and F. J. Doyle, "Core module biomarker identification with network exploration for breast cancer metastasis," *BMC Bioinformatics*, vol. 13, no. 1, p. 1, 2012.
- [2] Y. Liu, F. Tian, Z. Hu, and C. DeLisi, "Evaluation and integration of cancer gene classifiers: identification and ranking of plausible drivers," *Sci. Rep.*, vol. 5, 2015.
- [3] M. Kim, F. Farnoud, and O. Milenkovic, "HyDRA: gene prioritization via hybrid distance-score rank aggregation," *Bioinformatics*, vol. 31, no. 7, pp. 1034–1043, 2015.
- [4] M. Reboiro-jato, R. Laza, H. López-fernández, D. Glez-peña, F. Díaz, and F. Fdez-riverola, "Expert Systems with Applications genEnsemble : A new model for the combination of classifiers and integration of biological knowledge applied to genomic data," *Expert Syst. Appl.*, vol. 40, no. 1, pp. 52–63, 2013.
- [5] T. Les, T. Markiewicz, S. Osowski, W. Kozłowski, and M. Jesiotr, "Fusion of FISH image analysis methods of HER2 status determination in breast cancer," *Expert Syst. Appl.*, vol. 61, pp. 78–85, 2016.