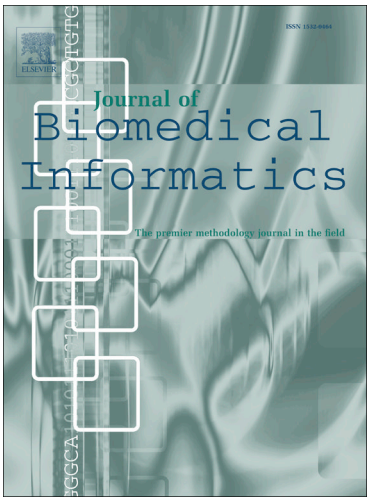

Benchmarking Relief-Based Feature Selection Methods for Bioinformatics Data Mining

Ryan J. Urbanowicz, Randal S. Olson, Peter Schmitt, Melissa Meeker, Jason H. Moore

PII: S1532-0464(18)30141-2
DOI: https://doi.org/10.1016/j.jbi.2018.07.015
Reference: YJBIN 3020

To appear in: *Journal of Biomedical Informatics*

Received Date: 15 January 2018
Revised Date: 30 June 2018
Accepted Date: 14 July 2018

Please cite this article as: Urbanowicz, R.J., Olson, R.S., Schmitt, P., Meeker, M., Moore, J.H., Benchmarking Relief-Based Feature Selection Methods for Bioinformatics Data Mining, *Journal of Biomedical Informatics* (2018), doi: https://doi.org/10.1016/j.jbi.2018.07.015

Benchmarking Relief-Based Feature Selection Methods for Bioinformatics Data Mining

Ryan J. Urbanowicz[a,*], Randal S. Olson[a], Peter Schmitt[a], Melissa Meeker[b], Jason H. Moore[a]

[a]*Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA*
[b]*Ursinus College, Collegeville, PA, 19426, USA*

Abstract

Modern biomedical data mining requires feature selection methods that can (1) be applied to large scale feature spaces (e.g. 'omics' data), (2) function in noisy problems, (3) detect complex patterns of association (e.g. gene-gene interactions), (4) be flexibly adapted to various problem domains and data types (e.g. genetic variants, gene expression, and clinical data) and (5) are computationally tractable. To that end, this work examines a set of filter-style feature selection algorithms inspired by the 'Relief' algorithm, i.e. Relief-Based algorithms (RBAs). We implement and expand these RBAs in an open source framework called ReBATE (Relief-Based Algorithm Training Environment). We apply a comprehensive genetic simulation study comparing existing RBAs, a proposed RBA called MultiSURF, and other established feature selection methods, over a variety of problems. The results of this study (1) support the assertion that RBAs are particularly flexible, efficient, and powerful feature selection methods that differentiate relevant features having univariate, multivariate, epistatic, or heterogeneous associations, (2) confirm the efficacy of expansions for classification vs. regression, discrete vs. continuous features, missing data, multiple classes, or class imbalance, (3) identify previously unknown limitations of specific RBAs, and (4) suggest that while MultiSURF* performs best for explicitly identifying pure 2-way interactions, MultiSURF yields the most reliable feature selection performance across a wide range of problem types.

Keywords: Feature Selection, ReliefF, Epistasis, Genetic Heterogeneity, Classification, Regression