Novel Features Ranking Method for High Dimensional Biological Data

Fatemeh Kargarfard[1]*, Esmaeil Ebrahimie[2, 3,4 5], Ashkan Sami[1]
[1]Department of Computer Science and Engineering, School of Electrical Engineering and Computer, Shiraz University, Shiraz, Iran.
[2]The University of Adelaide, School of Medicine, Adelaide, South Australia, Australia.
[3]Institute of Biotechnology, Shiraz University, Shiraz, Iran.
[4]School of Information Technology and Mathematical Sciences, Division of Information Technology Engineering & Environment, University of South Australia, Adelaide, Australia.
[5]School of Biological Sciences, Faculty of Science and Engineering, Flinders University, Adelaide, Australia.
*kargarfard870@gmail.com

Abstract. Undoubtedly, feature selection phase plays a signifcant role in the accurate result of any machine learning process in general and data classification in particular. Biological data often contain irrelevant and noisy features. So selecting a small subset out of thousands of features in biological dataset is a considerable step to identify biological variables and meaningful patterns based on DNA or protein sequences. Existing feature selection methods are computationally expensive and time consuming. Since nominal datasets are the major parts of data in biological world, in this paper, we propose a new feature ranking method on high dimensional nominal datasets. In order to increase the accuracy of classification, we first measure significance of each (feature) gene by (feature) gene ranking method via introducing a new metric which calculate the importance of all features based on the correlation between the feature and class label and then allocate a score to every features. Therefore our method can improve classification accuracy without interruption of redundant features. Finally, experiments on real influenza sequence data and the UCI datasets show that our method compares very well with other well known methods (including: Information gain, Chi Squared, Information gain ratio and Relief) in classification and ranking while being computationally efficient and scalable.
Keywords: Feature Selection; Feature Ranking; Nominal Dataset; High dimensional Dataset; Biological dataset