A Bi-Stage IG - PCA Dimensionality Reduction Approach for the Diagnosis of Pandemic Diseases from Clinical Attributes

Preethi K.* and Ramakrishnan. M.

Department of Computer Applications School of Information Technology, Madurai Kamaraj University, Madurai, India

*Corresponding authors: mail2preethi06@gmail.com, ramkrishod@mkuniversity.org

ABSTRACT

Diagnosis of pandemic diseases is a challenging task due to high dimensional clinical data sets. Feature selection plays a vital role in extracting the features relevant to the diagnosis. To reduce the computational complexity and to enhance the accuracy of prediction results, a bi-stage IG-PCA based feature extraction technique was proposed in this research work. The Input COVID-19 patient records each containing 20 features is initially pre-processed by applying data cleaning, transformation and normalization operations. For feature selection, in stage I, the information gain value of each input feature is computed, ranked and only the features whose gain value greater than the threshold value alone are selected. In stage II, principal component analysis is applied to extract the optimal features. Among the 20 input features, 8 features were selected in stage I and an optimal feature set with 7 features was extracted in stage II. This featureset was further used to train the supervised Naïve Bayes machine learning classifier, the results were analysed and evaluated with various classification metrics and found that the proposed method outperformed with the prediction accuracy of 98.9%.

Keywords: Feature Selection, Information Gain, Naïve Bayes classifier

DOI: 10.21467/abstracts.158 ISBN: 978-81-965621-0-6 (eBook)