Classification performance improvement using random subset feature selection algorithm for data mining

Lakshmi Padmaja Dhyaram, B. Vishnuvardhan

PII: S2214-5796(17)30179-X
DOI: https://doi.org/10.1016/j.bdr.2018.02.007
Reference: BDR 94

To appear in: *Big Data Research*

Received date: 24 July 2017
Revised date: 21 January 2018
Accepted date: 23 February 2018

Please cite this article in press as: L.P. Dhyaram, B. Vishnuvardhan, Classification performance improvement using random subset feature selection algorithm for data mining, *Big Data Res.* (2018), https://doi.org/10.1016/j.bdr.2018.02.007

Classification Performance Improvement Using Random Subset Feature Selection Algorithm for Data Mining

Lakshmi Padmaja Dhyaram,Department Of IT,Associate Professor,Anurag Group Of Institutions,Hyderabad :email *lakshmipadmajait@cvsr.ac.in*

Dr.B Vishnuvardhan,Professor,JNTUH,Hyderabad ,email MAILVISHNUVARDHAN@GMAIL.COM

Abstract

This study focuses on feature subset selection from high dimensionality databases and presents modification to the existing Random Subset Feature Selection(RSFS) algorithm for the random selection of feature subsets and for improving stability. A standard k-nearest-neighbor (kNN) classifier is used for classification. The RSFS algorithm is used for reducing the dimensionality of a data set by selecting useful novel features. It is based on the random forest algorithm. The current implementation suffers from poor dimensionality reduction and low stability when the database is very large. In this study, an attempt is made to improve the existing algorithm's performance for dimensionality reduction and increase its stability. The proposed algorithm was applied to scientific data to test its performance. With 10 fold cross-validation and modifying the algorithm classification accuracy is improved. The applications of the improved algorithm are presented and discussed in detail. From the results it is concluded that the improved algorithm is superior in reducing the dimensionality and improving the classification accuracy when used with a simple kNN classifier. The data sets are selected from public repository. The datasets are scientific in nature and mostly used in cancer detection. From the results it is concluded that the algorithm is highly recommended for dimensionality reduction while extracting relevant data from scientific datasets.

Keywords: Random Forest, Subset Feature Selection, Dimensionality Reduction, Scientific Data, Stability

1. Introduction

Data mining, the extraction of useful hidden features from large databases, is an effective new innovation with incredible potential to help organizations, focus on developing business strategies. The tools, developed for mining data, anticipate future patterns and practices, permitting organizations to make proactive, learning-driven choices. Many data mining tools can address business challenges more effectively than can traditional query or report-based tools. The performance of traditional tool's is very poor because of the large quantities of data involved. However, large quantities