Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 115 (2017) 209–217

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India

A Comparative Performance Evaluation of Supervised Feature Selection Algorithms on Microarray Datasets

C.ArunKumar[a]*, Sooraj M. P.[b], S.Ramakrishnan[c]

[a]*Dept. of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, Amrita University,India*
[b]*Dept. of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, Amrita University India*
[c]*Dept. of Information Technology, Dr.Mahalingam College of Engineering and Technology, Pollachi, India*

Abstract

The focus of this research paper is to compare the different filter, wrapper and fuzzy rough set based feature selection methods based on three parameters namely execution time, number of features selected in the reduced subset and classifier accuracy. The results are analyzed using the different feature selection methods on cancer microarray gene expression datasets. This research work finds KNN classifier to produce higher classifier accuracy compared to traditional classifiers available in literature. Also fuzzy rough set based feature selection approach is computationally faster and produces lesser number of genes in the reduced subset compared to correlation based filter.

Peer-review under responsibility of the scientific committee of the 7th International Conference on Advances in Computing & Communications.

Keywords: Feature Selection;Fuzzy;RoughSet;Microarray;Gene expression analysis;

* Corresponding author. Tel.: +91-99650-55500;
E-mail address: c_arunkumar@cb.amrita.edu

1877-0509
Peer-review under responsibility of the scientific committee of the 7th International Conference on Advances in Computing & Communications
10.1016/j.procs.2017.09.127

1. Introduction

Feature Selection (FS) is a crucial part of preprocessing step in the process of Knowledge Data Discovery (KDD) [1]. Attribute Selection, Instance Selection, Data Selection, Feature Construction, Variable Selection and Feature Extraction are some of the different names assigned to Feature Selection Algorithms (FSA). They are predominantly used for data reduction by removing irrelevant and redundant data. As mentioned in [2], feature selection improves the quality of the data and increases the accuracy of data mining algorithms by reducing the space and time complexity. Feature selection focuses on eliminating redundant and irrelevant data. Many FSAs are introduced in past decade but most of them do not perform well on high-dimensional datasets with a large number of redundant features [3].

Microarray chip helps the simultaneous analysis of gene expression profiles of a large number of genes in a single experiment. Understanding gene expression pattern can help to diagnose and distinguish different type of cancer [4]. Generally, Microarray datasets have a high number of features (ranges from 2000 to 30000) compared to the samples size (mostly less than 150) and this issue is called “curse of dimensionality” [5]. So, microarray analysis brings an exciting field of study for Machine Learning researchers. In addition to this, noise and variability of the data make this domain more exciting [6].

[7] says that several genes present in the microarray dataset are irrelevant and poses a threat for accurate classification of the problem. So feature selection plays a bigger role in removing irrelevant features. Thus machine learning algorithms focus only on the necessary features pertaining to build the model. There are three different types of feature selection methods named Filter, Embedded and Wrapper method. Apart from this, feature selection can be Univariate or Multivariate. When a Univariate method does not take into account the dependency among the features, a Multivariate method does it [8].

This work uses two well-known microarray binary datasets (Leukemia and Breast Cancer), which suffer from the problems of class imbalance and dataset shift. The datasets are evaluated using k-fold cross-validation technique since it is considered as the best choice as mentioned in [9-11]. Fig. 1 shows a unified model of feature selection [12]. Individual and Subset evaluation are the two general methods used to evaluate attributes. Our research study uses different well-known classifiers, such as Random Forest (RF), Decision Tree (J48), k-nearest neighbors (k-NN), Naive Bayes (NB) and Support Vector Machine (SVM) for validating the output of FSAs. Reduced feature subset is used to train the classifiers and thereby measure its classification ability.

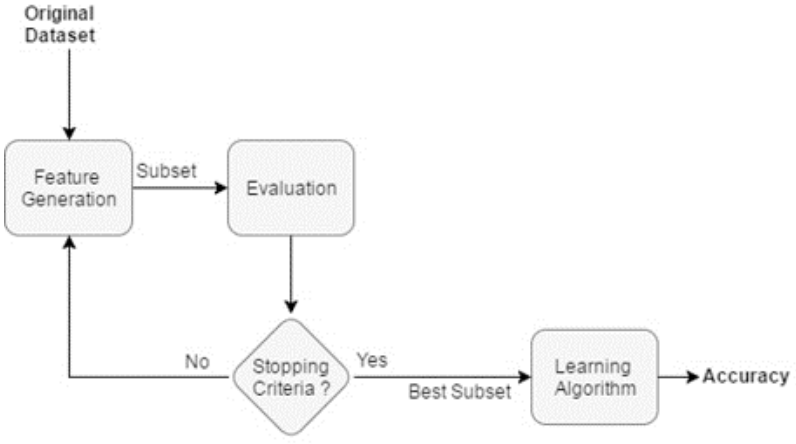

Fig. 1. A unified model of feature selection.