

Review Article

A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data

Zena M. Hira and Duncan F. Gillies

Department of Computing, Imperial College London, London SW 7 2BZ, UK

Correspondence should be addressed to Zena M. Hira; zena.hira@gmail.com

Received 12 March 2016; Accepted 12 May 2016

Academic Editor: Huixiao Hong

Copyright © 2016 Z. M. Hira and D. F. Gillies. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We summarise various ways of performing dimensionality reduction on high-dimensional microarray data. Many different feature selection and feature extraction methods exist and they are being widely used. All these methods aim to remove redundant and irrelevant features so that classification of new instances will be more accurate. A popular source of data is microarrays, a biological platform for gathering gene expressions. Analysing microarrays can be difficult due to the size of the data they provide. In addition the complicated relations among the different genes make analysis more difficult and removing excess features can improve the quality of the results. We present some of the most popular methods for selecting significant features and provide a comparison between them. Their advantages and disadvantages are outlined in order to provide a clearer idea of when to use each one of them for saving computational time and resources.

1. Introduction

In machine learning as the dimensionality of the data rises, the amount of data required to provide a reliable analysis grows exponentially. Bellman referred to this phenomenon as the “curse of dimensionality” when considering problems in dynamic optimisation [1]. A popular approach to this problem of high-dimensional datasets is to search for a projection of the data onto a smaller number of variables (or features) which preserves the information as much as possible. Microarray data is typical of this type of small sample problem. Each data point (sample) can have up to 10,000 variables (gene probes) and processing a large number of data points involves high computational cost [2]. When the dimensionality of a dataset grows significantly there is an increasing difficulty in proving the result statistically significant due to the sparsity of the meaningful data in the dataset in question. Large datasets with the so-called “large small n ” problem (where n is the number of features and m is the number of samples) tend to be prone to overfitting. An overfitted model can mistake small fluctuations for important variance in the data which can lead to classification errors. This difficulty can also increase due to noisy features. Noise in a dataset is defined as “the error in the variance of a measured variable” which can result from errors in measurements or natural variation [3]. Machine learning algorithms tend to be affected by noisy data. Noise should be reduced as much as possible in order to avoid unnecessary complexity in the inferred models and improve the efficiency of the algorithm [4]. Common noise can be divided into two types [5]:

- (1) Attribute noise.
- (2) Class noise.

Attribute noise is caused by errors in the attribute values (wrongly measured variables, missing values) while class noise is caused by samples that are labelled to belong in more than one class and/or misclassifications.

As the dimensionality increases the computational cost also increases, usually exponentially. To overcome this problem it is necessary to find a way to reduce the number of features in consideration. Two techniques are often used:

- (1) Feature subset selection.
- (2) Feature extraction.

Cancer is among the leading causes of death worldwide accounting for more than 8 million deaths according to

