Chapter 5
Feature Selection Algorithms for Classification and Clustering in Bioinformatics

Sujata Dash
Gandhi Institute for Technology, India

Bichitrananda Patra
KMBB College of Engineering and Technology, India

ABSTRACT

This chapter discusses some important issues such as pre-processing of gene expression data, curse of dimensionality, feature extraction/selection, and measuring or estimating classifier performance. Although these concepts are relatively well understood among the technical people such as statisticians, electrical engineers, and computer scientists, they are relatively new to biologists and bioinformaticians. As such, it was observed that there are still some misconceptions about the use of classification methods. For instance, in most classifier design strategies, the gene or feature selection is an integral part of the classifier, and as such, it must be a part of the cross-validation process that is used to estimate the classifier prediction performance. Simon (2003) discussed several studies that appeared in prestigious journals where this important issue is overlooked, and optimistically biased prediction performances were reported. Furthermore, the authors have also discuss important properties such as generalizability or sensitivity to overtraining, built-in feature selection, ability to report prediction strength, and transparency of different approaches to provide a quick and concise reference. The classifier design and clustering methods are relatively well established; however, the complexity of the problems rooted in the microarray technology hinders the applicability of the classification methods as diagnostic and prognostic predictors or class-discovery tools in medicine.

DOI: 10.4018/978-1-4666-4936-1.ch005

1. INTRODUCTION

As computer and database technologies advance rapidly, data accumulates in a speed unmatchable to the human's capacity of data processing. Data mining (Sanjay Chawla, 2010; J. Han and M. Kamber, 2001) as a multidisciplinary from databases, machine learning and statistics, is efficient in transforming the mountains of data into nuggets. Researchers and practitioners realize that, to use effectively data mining tools, data pre-processing is highly essential (M.A. Hall, 2000). Feature selection or dimensionality reduction is one of the important and frequently used techniques in data pre-processing for data mining and bio-informatics applications.

In contrast to other dimensionality reduction techniques like those based on projection (e.g. principal component analysis) or compression, feature selection techniques do not alter the original features of the variables, but merely selects a subset of them. Thus, they preserve the original semantics of the variables, hence offering the advantage of interpretability by a domain expert. It reduces the number of features, removes irrelevant, redundant, or noisy data and brings immediate effects for applications such as improving execution time of a data mining algorithm, improving mining performance such as classification accuracy and result comprehensibility.

Feature selection has been a fertile field of research and development since 1970s in statistical pattern recognition (Michael D Swartz, et al., 2008; P. Mitra et al., 2002), machine learning (Jennifer G. Dy and Carla E. Brodley, 2004; Jianqing Fan et al., 2009A. L. Blum and P. Langley, 1997; G. H. John et al, 1994) and data mining (M. Dash et al, 2002) and widely applied to many fields such as text categorization (E. Leopold and J. Kindermann, 2002) image retrieval (Y. Rui et al., 1999), customer relationship management (K. S. Ng and H. Liu, 2000), intrusion detection (W. Lee et al., 2000) and genomic analysis (E. Xing et al., 2001).

The main aim of this chapter is to make researchers aware of the benefits, and in some cases even the necessity of applying feature selection techniques in Bioinformatics domain, highlighting the efforts given by the bioinformatics community in developing novel and adapted procedures. This chapter is organized into six sections. Section 2 describes the basic steps associated with feature selection techniques. Section 3 demonstrates the different feature selection algorithms considering the evaluation criteria involved. Section 4 demonstrates the classifier performance and section 5 demonstrates the unsupervised classification methods. Section 6 concludes the chapter with discussion on current trends and future direction.

2. FEATURE SELECTION TECHNIQUES

As many pattern recognition techniques were originally not designed to cope with large amounts of irrelevant features, combining them with FS (Feature Selection) techniques has become a necessity in many applications (H. Liu and L. Liu, 2005). The objectives of feature selection are manifold, the most important ones being:

1. To avoid over fitting and improve model performance, i.e. prediction performance in the case of supervised classification and better cluster detection in the case of clustering,
2. To provide faster and more cost-effective models, and
3. To gain a deeper insight into the underlying processes that generated the data.

However, the advantages of feature selection techniques come at a certain price, as the search for a subset of relevant features introduces an additional layer of complexity in the modelling task. Instead of just optimizing the parameters of the model for the full feature subset, we now need to find the optimal model parameters for the