Chapter 7
Heuristic Principal Component Analysis-Based Unsupervised Feature Extraction and Its Application to Bioinformatics

Y-H. Taguchi
Chuo University, Japan

Hideaki Umeyama
Chuo University, Japan

Mitsuo Iwadate
Chuo University, Japan

Yoshiki Murakami
Osaka City University, Japan

Akira Okamoto
Aichi University of Education, Japan

ABSTRACT

Feature Extraction (FE) is a difficult task when the number of features is much larger than the number of samples, although that is a typical situation when biological (big) data is analyzed. This is especially true when FE is stable, independent of the samples considered (stable FE), and is often required. However, the stability of FE has not been considered seriously. In this chapter, the authors demonstrate that Principal Component Analysis (PCA)-based unsupervised FE functions as stable FE. Three bioinformatics applications of PCA-based unsupervised FE–detection of aberrant DNA methylation associated with diseases, biomarker identification using circulating microRNA, and proteomic analysis of bacterial culturing processes–are discussed.

INTRODUCTION

Feature extraction (FE) is a task that reduces the number of features (independent variables) for predicting/estimating (a dependent variable). For example, when performing face recognition as a computational task, the specific parts of facial photos, e.g., eye lines, colors of irises, or shapes of jaws, should be considered. Alternatively, to predict tomorrow's stock prices computationally, certain factors, e.g., today's prices, economical indices, or weather, should be taken into account. The problem is that increased numbers of features considered does not always result in better

DOI: 10.4018/978-1-4666-6611-5.ch007

performance. To perform better face recognition, considering hairstyle as a key feature is not useful, since it can vary. For stock price prediction, what your wife or husband cooks this morning is not useful as an important factor. However, neglecting the shape of the nose in face recognition or ignorance of this week's unemployment rate for stock price prediction may reduce performance of the task. Thus, clearly, there should be minimum number of critical features to achieve the best performance. The current problem is how to determine the specific set of such features.

The era of big data has added additional difficulty to this problem: a small number of samples (cases) versus too many features observed. For example, in facial recognition, it is not difficult to obtain many features from a facial photo that often consists of several millions of pixels, each of which has more than a million color grades. However, it is not a realistic requirement to collect a million facial photos, especially from a cost point of view, if payment is required for research use of individual photos. For stock market prediction, the situation appears better, because stock prices, even measured per second over months, can be collected and recorded. However, the use of many sample (case) numbers does not always resolve the "many features vs. small cases" difficulty, since these records are not always independent of each other. Stock price varies periodically over time, e.g., daily, weekly, or even seasonally. Thus, huge amounts of data are often simply replicates. What is required are samples taken under different economic situations. For example, if the market is lively, data should be obtained from when the market is bad. However, this kind of data is only available after the economic crash, thus stock price prediction based only on measurements taken under good economic situations naturally fails to predict price reduction (and money is therefore lost). Thus, "many features vs small samples" problems must be resolved to compete with massive data flowing into "prediction" systems.

In this chapter, we would like to propose an alternative solution to the difficulty of FE when the number of samples is much lower than the number of features. There have been many proposals to overcome this problem, e.g., stepwise feature extraction (Prasad, 2008), regression with regularization (Girosi, 1995) and Bayesian work frame (Chu, 2006). Despite these efforts, FE problems with small sample size and large number of features have not been solved completely.

Recently, unsupervised FE (De Backer, 1998) has gained the interest of many researchers. Unsupervised FE is robust and thus, it is expected to provide more stable FEs, i.e., unsupervised FEs have weaker sample dependence than other methods with optimization procedures in all senses.

Although many kinds of implementations of unsupervised FE are possible, we employed one of the simplest, principal component analysis (PCA) based FE (Murakami, 2012; Taguchi, 2013) in this chapter. Since PCA is a linear method, it is expected to have greater robustness than other more complicated methods such as those with kernel tricks (Scholkopf, 2001) or Bayesian statistics. Linear methods, such as PCA, are also expected to be less computationally challenging. Finally, it is easier to interpret the obtained results, since it is a linear combination of the original features. In contrast to these advantages, linear methods including PCA usually have poorer performances than other more complicated methods (Cao, 2003). In this chapter, we also discuss how these difficulties can be overcome when applying PCA to FEs.

The applications considered in this chapter are the detection of aberrant DNA methylation associated with diseases (Ishida, 2014; Kinoshita 2014), biomarker identification of circulating microRNA (miRNA) (Murakami, 2012; Taguchi, 2013) and proteomic analysis of bacterial culture (Taguchi, 2012). In the *background* section, we introduce FE methods and compare them to our methods. Then, we illustrate details of our proposed method and an illustrative (artificial) example. Each ap-