High-Dimensional Statistical and Data Mining Techniques

H

Gokmen Zararsiz
Hacettepe University Ankara, Turkey

INTRODUCTION

Because of the rapid growth of applications in computational biology, computer and information sciences has resulted in large volumes of high dimensional data (HDD) becoming available. In a DNA microarray dataset, there exist hundreds or thousands of dimensions, each corresponding to the gene expression of biological samples. In textual data, there exists high number of dimensions where each dimension corresponds to a word. In customer purchase behavior data, there also exists a high number of dimensions, corresponding to merchandises. Similar situation is present in market basket data where dimensions are the products of a transaction (Dziuda, 2009; Wang & Yang, 2005; Bekkerman & Allan, 2003; Tan, Steinbach, & Kumar, 2006).

Analyzing these datasets is more difficult than other datasets because of the dimension problem and because traditional statistical techniques are no longer applicable. This is because, most of the traditional approaches were based on the assumption that the observation number should be more than the variable number ($n>p$). Recent technologies provide an opportunity for researchers to measure lots of variables however their highcost limits the number of observations. The number of variables exceeds the number of observations ($n<p$) and the classical statistical techniques fail. Thus, more adapted techniques are proposed to analyze these types of data (Donoho, 2000; Verleysen, 2003).

In this chapter, we will give an overview of these adapted approaches and a brief review of the most popular statistical and data mining techniques in the analysis of HDD.

BACKGROUND

HDD refers to data whose number of dimension is at least larger than the dimensions considered in classical multivariate analysis in statistical theory. In many fields, it is probable to work with data in which the number of variables is more than the number of observations. Richard Bellman mentioned the difficulties of function optimization by exhaustive enumeration in the function domain and referred to the problem as the "curse of dimensionality" (Bellman, 1961). If we consider 10 data points simulated from uniform distribution, then let d indicate the number of dimensions. The data points look agglomerated with each other in one dimension ($d=1$). The data look more dispersed in two dimensions ($d=2$) and even more so in three dimensions ($d=3$). In a statistical problem, 10^d evaluations of a function are needed and this may involve huge computational cost, even for a reasonable value of d (Verleysen, 2003).

This curse of dimensionality makes classical statistical algorithms inapplicable due to their shortcomings. For instance, linear discriminant analysis is one of the powerful statistical method for predicting categorical outcomes. However, this method is very flexible in the presence of high number of correlated variables and overfits in classifying this kind of data. Conversely, it is very rigid and underfits the data when the class boundaries are nonlinear and complex (Hastie, Buja, & Tibshirani, 1995). These problems are arised from the covariance structure of the data and unpredictability of this structure makes this method inapplicable to HDD. Most of the multivariate statistical methods have similar problems based on the covariance structure of data. Also,

DOI: 10.4018/978-1-4666-5202-6.ch102

meeting the assumptions of multivariate statistical techniques in HDD, such as multivariate normality and homogeneity of covariance matrices is troublesome.

Many techniques have been proposed as a modified version of the classical techniques to analyze HDD. For instance, the *S* test (Tusher, Tibshirani, & Chu, 2001) is a modified form of *t* test for class comparison analysis of HDD due to the difficulty of error variance estimation. The penalized logistic regression technique (Zhu & Hastie, 2004) is a modified version of the ordinary logistic regression technique for HDD classification. A similar relation exists between penalized discriminant analysis (Hastie, Buja, & Tibshirani, 1995) and Fisher's linear discriminant analysis.

Apart from these techniques, novel algorithms have been developed to analyze HDD by taking into account the curse of dimensionality problem. In this respect, support vector machines (SVM) were proposed and applied in classification and regression analysis of HDD. Random forests were introduced as a combination of the prediction results of many decision trees and reported to be one of the most accurate algorithms in classification and regression (Tan, Steinbach, & Kumar, 2006). Self-organizing maps (SOM), which are a form of artificial neural networks, were developed to represent HDD in lower dimensions for cluster analysis.

The adapted and newly discovered methods have been applied in many fields involving HDD. For high dimensional textual data, Dhillon et al. successfully applied clustering techniques to create a vector-space model (Dhillon, Guan, & Kogan, 2002), Joachims adapted support vector machine algorithm for text categorization (Joachims, 1998), Bingham et al. described the use of dimensionality reduction techniques on this kind of data (Bingham & Mannila, 2001). For image data, Jain et al. showed the assessment and application of feature selection methods (Jain & Zongker, 1997), Tatiraju et al. applied some clustering algorithms for image segmentation (Tatiraju & Mehta, n.d.), Prastawa et al. implemented an outlier detection algorithm for brain tumor segmentation with MR images (Prastawa, Bullitt, & Gerig, 2004). For gene expression data, Golub et al. described how to apply clustering and classification algorithms for molecular classification of cancer (Golub et al., 1999), Cui et al. demonstrated the use of class comparison techniques (Cui & Churchill, 2003), Oba et al. developed a Bayesian approach for missing value analysis (Oba et al., 2003). For customer purchase behavior data, Kim et al. combined a number of data mining methods using genetic algorithm for prediction (Kim, Kim, & Lee, 2002). More empirical studies can be found in (Buhlmann & van de Geer, 2011; Cai & Shen, 2011; Kogan, 2007; Tan, Steinbach, & Kumar, 2006).

MAIN FOCUS

The data can be presented by a matrix with *n* observations, *p* variables and *c* classes. Observations are the samples used in the research, variables represent these samples' characteristics (or features) and classes represent the observation groups or clusters depending on their similarities to each other. A data matrix is demonstrated in Table 1. This data matrix contains the expression levels of 2000 genes belonging to 62 observations (22 control, 40 colon cancer patients) (Alon et al., 1999).

Outlier Detection

Outlier detection is a crucial pre-processing part of data analysis. The objective in outlier detection is to detect observations which are different from others based on their variable values. The techniques used in outlier detection can be classified into four groups: model based, proximity based, density based and clustering based approaches (Aggarwal & Yu, 2001; Tan, Steinbach, & Kumar, 2006).

Model based approaches are statistical approaches that are based on predefined probabil-