Efficient High Dimensional Data Classification

E

Hari Seetha
VIT University, Vellore, India

M.Narasimha Murty
Indian Institute of Science, Bangalore, India

INTRODUCTION

The present century is the century of big data. Recent advancements in technology have made huge amounts of data available. The trend today is towards not only collecting more patterns but rather to collect a larger number of variables that describe each pattern. The automatic and systematic collection of finer details of each pattern has led to high dimensional data. The classical classification methods are not designed to cope with this kind of explosive growth of the dimensionality of individual patterns. The demand for large number of patterns grows exponentially with the dimensionality of the feature space. To overcome this problem either we should reduce the dimensionality or increase the size of the training set by adding some artificially generated training patterns to the training set. Dimensionality reduction methods have been well studied. In this chapter we present the various methods employed in pattern synthesis and its effect on classification performance of both kNN (K-nearest neighbor) and SVM (support vector machine) classifiers.

BACKGROUND

Big data refers to the data that is large in size or large in dimensionality or both. It becomes difficult to store, analyze, search, share and visualize such a large data. Big data is being generated due to the advancements in sensor technology, wireless sensor networks, information gathering mobile devices, remote sensing devices, various data capturing tools and sophisticated cameras. Big data are of high volume (i.e. both size and dimensionality may be large), of high velocity (i.e. making time sensitive decisions by processing big data as it streams into the enterprise), and/or high variety information assets (i.e. structured and unstructured data). The traditional data processing applications and conventional decision making algorithms become impractical in handling big data of this size and volume.

A wide variety of sensors enable collecting a large number of observations (patterns). They also enable collecting larger number of features that describe each pattern leading to high dimensional data. The classification of the high dimensional data has become increasingly important in the fields of engineering, computational biology, genomics and pattern recognition. In high dimensional data, generally the number of features is greater than the number of patterns. One of the major challenges in classifying high dimensional data is high variance and overfitting (due to the noise). The other concern is the curse of dimensionality that makes the conventional classification impractical. The curse of dimensionality refers to the demand for more number of observations with an increase in the dimensionality in order to get good estimates. The increase in dimensionality increases the noise, over fitting and the complexity of any learning algorithm. These problems can be solved and the generalization performance of the classifier can be improved either by 1) reducing the dimensionality of the data 2) minimizing the

DOI: 10.4018/978-1-4666-5202-6.ch073

VC dimension of the classifier or 3) by providing a large number of training samples.

Dimensionality reduction involves feature extraction or feature selection. Feature extraction techniques such as Principal Component Analysis, Linear Discriminant Analysis, Random Projection, Independent Component Analysis etc., have been extensively explored (Van der Maaten Postma & Van den Herik, 2009; Subasi & Ismail Gursoy, 2010; Seetha, Murty & Saravanan, 2011a; 2011b; Deng et al., 2012). Feature selection techniques using wrappers and filters have also been well studied (Kohavi & John, 1997; Liu & Yu, 2005; Gao et al., 2011; Bermejo et al., 2012). The recent trend is towards pattern synthesis to combat the curse of dimensionality (Agrawal et al., 2005; Viswanath, Murty & Bhatnagar, 2006; Chen, Zhu & Nakagawa, 2011; Seetha, Saravanan & Murty, 2012).

MAIN FOCUS

Jain & Chandrasekharan (1982) have reported that the number of training samples per class should be at least 5-10 times the number of features. Practically it is expensive and difficult to obtain a large number of labeled samples. Further, the classification error increases with the dimensionality if the sample size is insufficient. Synthetic pattern generation (i.e. artificial pattern generation) becomes essential in such situations. Some of the methods employed to generate artificial patterns are discussed in the following sections.

Bootstrapping

The conventional bootstrap technique samples the dataset uniformly with replacement to form a training set. The artificial samples generated are called bootstrap samples. This method is used for error estimation of nearest neighbor classifier (Jain, Dubes & Chen, 1987). Their results have demonstrated that the error obtained using bootstrapping is less than that obtained using "leave one out" strategy. It has been reported in the literature that one of the drawbacks of bootstrapping is the discreteness of the empirical distribution function (Chernick, Murthy & Nealy, 1985). A method for comparing classifiers is to use the jackknife or bootstrap estimation method (Duda, Hart & Stork, 2005). In the application of the jackknife approach the accuracy of a given algorithm is estimated by training the classifier n separate times, each time deleting a single training pattern from the training set which is used for testing the resulting classifier. The jackknife estimate of accuracy is the mean of these leave-one-out accuracies. Although the computational complexity would be very high for large n, the advantage of the jackknife approach is that it can provide measures of confidence in the comparison between two classifier designs (Duda, Hart & Stork, 2005).

Hamamoto, Uchumira & Tomita (1997) have used various bootstrap techniques in designing nearest neighbor classifier. A novel bootstrap technique that creates bootstrap samples by locally combining the original training samples was proposed by Hamamoto, Uchumira & Tomita (1997). It was shown that on applying this new bootstrap technique the performance of 1NN classifier outperformed the conventional k-NN classifier. The other advantage of this bootstrapping technique is the removal of outliers which acts as a smoother of the distribution of the training samples. Therefore this bootstrap technique becomes useful in designing either a nearest neighbor or an SVM classifier in high dimensional spaces. The bootstrapping technique is used as follows to generate artificial training patterns: Let X be a training pattern and $X_1,...,X_r$ be its r nearest neighbors in its class. Then $X' = \frac{1}{r}\sum_{h=1}^{r} X_h$ is the artificial pattern generated for X.

For a given finite set of training data, the best generalization performance will be achieved if the right balance is struck between the accuracy obtained on that particular training set and the