Locally Adaptive Techniques for Pattern Classification

Carlotta Domeniconi
George Mason University, USA

Dimitrios Gunopulos
University of California, USA

INTRODUCTION

Pattern classification is a very general concept with numerous applications ranging from science, engineering, target marketing, medical diagnosis, and electronic commerce to weather forecast based on satellite imagery. A typical application of pattern classification is mass mailing for marketing. For example, credit card companies often mail solicitations to consumers. Naturally, they would like to target those consumers who are most likely to respond. Often, demographic information is available for those who have responded previously to such solicitations, and this information may be used in order to target the most likely respondents. Another application is electronic commerce of the new economy. E-commerce provides a rich environment to advance the state of the art in classification, because it demands effective means for text classification in order to make rapid product and market recommendations.

Recent developments in data mining have posed new challenges to pattern classification. Data mining is a knowledge-discovery process whose aim is to discover unknown relationships and/or patterns from a large set of data, from which it is possible to predict future outcomes. As such, pattern classification becomes one of the key steps in an attempt to uncover the hidden knowledge within the data. The primary goal is usually predictive accuracy, with secondary goals being speed, ease of use, and interpretability of the resulting predictive model.

While pattern classification has shown promise in many areas of practical significance, it faces difficult challenges posed by real-world problems, of which the most pronounced is Bellman's curse of dimensionality, which states that the sample size required to perform accurate prediction on problems with high dimensionality is beyond feasibility. This is because, in high dimensional spaces, data become extremely sparse and are apart from each other. As a result, severe bias that affects any estimation process can be introduced in a high-dimensional feature space with finite samples.

Learning tasks with data represented as a collection of a very large number of features abound. For example, microarrays contain an overwhelming number of genes relative to the number of samples. The Internet is a vast repository of disparate information growing at an exponential rate. Efficient and effective document retrieval and classification systems are required to turn the ocean of bits around us into useful information and, eventually, into knowledge. This is a challenging task, since a word level representation of documents easily leads 30,000 or more dimensions.

This paper discusses classification techniques to mitigate the curse of dimensionality and to reduce bias by estimating feature relevance and selecting features accordingly. This paper has both theoretical and practical relevance, since many applications can benefit from improvement in prediction performance.

BACKGROUND

In a classification problem, an observation is characterized by q feature measurements $\boldsymbol{x} = (x_1, \cdots, x_q) \in \Re^q$ and is presumed to be a member of one of J classes, L_j, $j = 1, \cdots, J$. The particular group is unknown, and the goal is to assign the given object to the correct group, using its measured features $\boldsymbol{x}$.

Feature relevance has a local nature. Therefore, any chosen fixed metric violates the assumption of locally constant class posterior probabilities, and fails to make correct predictions in different regions of the input space. In order to achieve accurate predictions, it becomes crucial to be able to estimate the different degrees of relevance that input features may have in various locations of the feature space.

Consider, for example, the rule that classifies a new data point with the label of its closest training point in the measurement space (1-Nearest Neighbor rule). Suppose

that each instance is described by 20 features, but only three of them are relevant to classifying a given instance. In this case, two points that have identical values for the three relevant features may, nevertheless, be distant from one another in the 20-dimensional input space. As a result, the similarity metric that uses all 20 features will be misleading, since the distance between neighbors will be dominated by the large number of irrelevant features. This shows the effect of the curse of dimensionality phenomenon; that is, in high dimensional spaces, distances between points within the same class or between different classes may be similar. This fact leads to highly-biased estimates. Nearest neighbor approaches (Ho, 1998; Lowe, 1995) are especially sensitive to this problem.

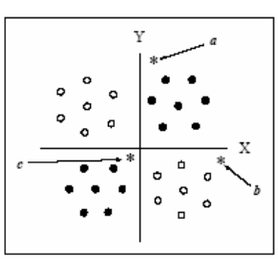
In many practical applications, things often are further complicated. In the previous example, the three relevant features for the classification task at hand may be dependent on the location of the query point (i.e., the point to be classified) in the feature space. Some features may be relevant within a specific region, while other features may be more relevant in a different region. Figure 1 illustrates a case in point, where class boundaries are parallel to the coordinate axes. For query *a*, dimension *X* is more relevant, because a slight move along the *X* axis may change the class label, while for query *b*, dimension *Y* is more relevant. For query *c*, however, both dimensions are equally relevant.

These observations have two important implications. Distance computation does not vary with equal strength or in the same proportion in all directions in the feature space emanating from the input query. Moreover, the value of such strength for a specific feature may vary from location to location in the feature space. Capturing such information, therefore, is of great importance to any classification procedure in high-dimensional settings.

MAIN THRUST

Severe bias can be introduced in pattern classification in a high dimensional input feature space with finite samples. In the following, we introduce adaptive metric techniques for distance computation capable of reducing the bias of the estimation.

Figure 1. Feature relevance varies with query locations



Friedman (1994) describes an adaptive approach (the Machete and Scythe algorithms) for classification that combines some of the best features of kNN learning and recursive partitioning. The resulting hybrid method inherits the flexibility of recursive partitioning to adapt the shape of the neighborhood $N(\mathbf{x}_0)$ of query $\mathbf{x}_0$, as well as the ability of nearest neighbor techniques to keep the points within $N(\mathbf{x}_0)$ close to the point being predicted. The method is capable of producing nearly continuous probability estimates with the region $N(\mathbf{x}_0)$ centered at $\mathbf{x}_0$ and the shape of the region separately customized for each individual prediction point.

The major limitation concerning the Machete/Scythe method is that, like recursive partitioning methods, it applies a greedy strategy. Since each split is conditioned on its ancestor split, minor changes in an early split, due to any variability in parameter estimates, can have a significant impact on later splits, thereby producing different terminal regions. This makes the predictions highly sensitive to the sampling fluctuations associated with the random nature of the process that produces the training data and, therefore, may lead to high variance predictions.

In Hastie and Tibshirani (1996), the authors propose a discriminant adaptive nearest neighbor classification method (DANN), based on linear discriminant analysis. Earlier related proposals appear in Myles and Hand (1990) and Short and Fukunaga (1981). The method in Hastie and Tibshirani (1996) computes a local distance metric as a product of weighted within and between the sum of squares matrices. The authors also describe a method of performing global dimensionality reduction by pooling the local dimension information over all points in the training set (Hastie & Tibshirani, 1996a, 1996b).

While sound in theory, DANN may be limited in practice. The main concern is that in high dimensions, one may never have sufficient data to fill in $q \times q$ (within and between sum of squares) matrices (where q is the dimensionality of the problem). Also, the fact that the distance metric computed by DANN approximates the weighted Chi-squared distance only when class densities are Gaussian and have the same covariance matrix may cause a performance degradation in situations where data do not follow Gaussian distributions or are corrupted by noise, which is often the case in practice.

A different adaptive nearest neighbor classification method (ADAMENN) has been introduced to try to minimize bias in high dimensions (Domeniconi, Peng & Gunopulos, 2002) and to overcome the previously mentioned limitations. ADAMENN performs a Chi-squared distance analysis to compute a flexible metric for produc-