Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 2289–2297
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Generative models for similarity-based classification

Luca Cazzanti[a], Maya R. Gupta[b,*], Anjali J. Koppal[c]

[a]*Applied Physics Lab, Seattle, WA, USA*
[b]*University of Washington, Seattle, WA, USA*
[c]*University of California, Berkeley, CA, USA*

Received 13 June 2007; received in revised form 26 October 2007; accepted 8 January 2008

Abstract

A maximum-entropy approach to generative similarity-based classifiers model is proposed. First, a descriptive set of similarity statistics is assumed to be sufficient for classification. Then the class-conditional distributions of these descriptive statistics are estimated as the maximum-entropy distributions subject to empirical moment constraints. The resulting exponential class-conditional distributions are used in a maximum a posteriori decision rule, forming the *similarity discriminant analysis* (SDA) classifier. Simulated and real data experiments compare performance to the k-nearest neighbor classifier, the nearest-centroid classifier, and the potential support vector machine (PSVM).

Keywords: Similarity; Maximum entropy; Discriminant analysis

1. Overview

Similarity-based classifiers classify a test sample x given only the pairwise similarities for the a test sample x and a set of training samples $\{x_i\}$, $i=1,\ldots,n$ [1–4]. The training samples' class labels are also given and denoted $\{y_i\}$ for $i=1,\ldots,n$. A similarity function s is a mapping that accepts two samples x, z from some sample space $x, z \in \mathscr{B}$, and returns a real number. That is, $s : \mathscr{B} \times \mathscr{B} \to \Omega$, where $\Omega \subset \mathbb{R}$. It is useful to think of the sample space $\mathscr{B}$ as an abstract space, such as "the space of all proteins," or "the space of all blogs." The similarity $s(x, z)$ is some judgement of how near samples x and z are, but similarities are not required to satisfy metric properties or any specific mathematical properties. The term "similarity-based classification" is also used when the given information is "dissimilarities," where a dissimilarity is a judgement of how far two samples are, but is not required to satisfy any specific mathematical properties.

Similarity-based learning is a useful approach when samples are described by categorical variables. For example, DNA is described as a sequence of bases, A, T, G, and G. Similarity-based learning is of course appropriate when the similarity or dissimilarity between samples is not a metric. For example, driving-times between any two given locations is not a metric, as it is often not symmetric and can violate the triangle inequality. Categorical variables and non-metric similarities/disssimilarities are common in fields such as bioinformatics, information retrieval, and natural language processing [1,4]. Also, similarity-based learning may be a better model than standard Euclidean-space learning for how humans classify, as psychologists have shown that metrics do not account for human judgements of similarity in complex situations [5–7]. Laub et al. have shown that non-metric similarities lead to information that can be useful for pattern recognition [8].

The simplest method for similarity-based classification is the nearest neighbor classifier, which determines the most-similar training sample to the test sample, and classifies the test sample as its most-similar neighbor's class. In fact, nearest neighbor classifiers using a tangent distortion [9] and a shape similarity metric [10] have both been shown to achieve lower error than metric k-NN for the MNIST character recognition task.

In this paper, we propose maximum-entropy generative similarity-based classifiers, which we term similarity discriminant analysis (SDA). We provide a review of the different

* Corresponding author.
E-mail addresses: luca@apl.washington.edu (L. Cazzanti), gupta@ee.washingtonedu (M.R. Gupta).

0031-3203/$30.00
doi:10.1016/j.patcog.2008.01.005

approaches to similarity-based classification, and discuss how the proposed generative architecture ties together many of these approaches. We compare the resulting log-linear SDA classifier to the state-of-the-art in similarity-based classification on benchmark data sets and on an illustrative simulated example.

Reviews of similarity functions relevant for pattern recognition can be found in Refs. [11,12]. Many similarity functions are information theoretic, including information content similarity [13], mutual information similarity [14,15], residual entropy similarity [16], and the similarity defined by the compressability of one sample given another [17,18].

2. Review of similarity-based classifiers

Similarity-based classifiers make decisions based on the outputs of a pairwise similarity function s and an explicit description of the sample space $\mathscr{B}$ is not required. That is, the similarity function s can be treated as a black box by the classifier. If in fact the sample space $\mathscr{B}$ is a known set of categorical features, then naive Bayes, neural nets, and decision trees can also be applied.

2.1. Nearest neighbors

Experiments have shown that nearest neighbors can perform well on practical similarity-based classification tasks [2,9,10,19]. Condensed near-neighbor strategies replace the set of training samples for each class with a set of prototypes for that class. Usually the prototype set is an edited set of the original training samples (also called edited nearest neighbors), but the prototypes do not need to be from the original training set. Many authors have considered strategies for condensing near-neighbors for similarity-based classification to increase classification speed, decrease the required memory, and possibly attain better performance [3,20–23].

2.2. Nearest centroid

An extreme form of condensed near-neighbors is to replace each class's training samples by one prototypical sample, often called a *centroid*. The resulting "nearest-centroid" classifier can be considered a simple parametric model [20], but lacks a probabilistic structure. The nearest-centroid approach classifies x as the class

$$\hat{y} = \arg \max_{h=1,\ldots,G} s(x, \mu_h), \tag{1}$$

where μ_h is the representative centroid for the class h. A standard definition for the centroid of a set of training samples is the training sample that has the maximum total similarity to all the training samples of the same class [3,20]:

$$\mu_h = \arg \max_{\mu \in \mathscr{X}_h} \sum_{z \in \mathscr{X}_h} s(z, \mu), \tag{2}$$

where $\mathscr{X}_h$ is the set of training samples from class h.

The nearest-centroid classifier is analogous to the nearest-mean classifier in Euclidean space, which is the optimal Euclidean-based classifier if one assumes Gaussian class-conditional distributions and that each class covariance is the identity matrix.

2.3. Embed in Euclidean space

One can embed the training and test samples in an Euclidean space using multi-dimensional scaling [24], and then use standard statistical learning methods in the Euclidean feature space. More generally, the data can be embedded in a pseudo-Euclidean space for classification [2,25]. The embedding approach can also be used for clustering, for example [26] embed samples based on pairwise similarities in a low-dimensional Euclidean space by computing a multi-dimensional scaling solution subject to an entropy constraint. This results in an Euclidean embedding that maximizes the separation between clusters in a data set, while maintaining as much as possible the original pairwise similarity structure of the data. For most nonlinear embedding methods, classifying a new test sample requires re-computing the metric space embedding for all the data. If the underlying similarity relationships are not well represented by a metric distance, the embedding may be relatively high-dimensional, invoking the curse of dimensionality. On the other hand, the Procrustes approach of embedding the training samples in a low-dimensional Euclidean space may fail to sufficiently capture the similarity relationships between the samples [5,6,23,27].

2.4. Use the similarities to training samples as features

Similarity-based classification problems can be turned into standard Euclidean-based learning problems by treating the $n \times 1$ vector of similarities between a test sample and the n training samples as a feature vector [2,28,29]. Graepel et al. [28] propose a separating hyperplane classifier using this approach. Duin et al. [2,29] consider various standard learning techniques for this approach, including a regularized Fisher linear discriminant classifier for this space.

An issue with using the vector of similarities as a feature vector is that the feature vector size is equal to the number of training samples, causing Bellman's curse-of-dimensionality difficulties for learning [30]. As investigated by Pekalska et al. [2], one way to mitigate the problem that the dimension of the feature space is equal to the number of training samples is to regularize the covariance matrix when applying linear discriminant analysis (LDA) or quadratic discriminant analysis (QDA). Another approach they suggest for solving the dimensionality problem is to use only a subset of the training samples to define the feature vector. The results of Pekalska et al. show that, on average over their different experiments, linear classifiers built on the similarity vectors achieve similar errors as 1-nearest neighbor, except in cases of severe noise, where the 1-nearest neighbor has high error. Also, their similarity-based linear classifiers generally perform slightly better than first embedding the training samples in an metric space and then applying a linear classifier.