Neural Networks 33 (2012) 247–256

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A new approach to a maximum a posteriori-based kernel classification method

Nopriadi [a,b,*], Yukihiko Yamashita [a]

[a] Department of International Development Engineering, Graduate School of Science and Engineering, Tokyo Institute of Technology, South 6th building, Ookayama, Meguro-ku, Tokyo, 152-8552, Japan

[b] Department of Engineering Physics, Faculty of Engineering, Gadjah Mada University, Jl. Grafika 2, Yogyakarta, 55281, Indonesia

ARTICLE INFO

Article history:
Received 8 September 2011
Received in revised form 6 May 2012
Accepted 18 May 2012

Keywords:
Maximum a posteriori
Objective function
Linear programming
Kernel function

ABSTRACT

This paper presents a new approach to a maximum a posteriori (MAP)-based classification, specifically, MAP-based kernel classification trained by linear programming (MAPLP). Unlike traditional MAP-based classifiers, MAPLP does not directly estimate a posterior probability for classification. Instead, it introduces a kernelized function to an objective function that behaves similarly to a MAP-based classifier.

To evaluate the performance of MAPLP, a binary classification experiment was performed with 13 datasets. The results of this experiment are compared with those coming from conventional MAP-based kernel classifiers and also from other state-of-the-art classification methods. It shows that MAPLP performs promisingly against the other classification methods.

It is argued that the proposed approach makes a significant contribution to MAP-based classification research; the approach widens the freedom to choose an objective function, it is not constrained to the strict sense Bayesian, and can be solved by linear programming. A substantial advantage of our proposed approach is that the objective function is undemanding, having only a single parameter. This simplicity, thus, allows for further research development in the future.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

In the pattern recognition field, there are commonly three different basic approaches to design a classifier: the concept of similarity, the probabilistic approach, and the geometric approach (Jain, Duin, & Mao, 2000). In the probabilistic approach, maximum *a posteriori* probability (MAP) estimation is usually used as a technique to assign unknown patterns to their appropriate class (Duda, Hart, & Stork, 2000; Jain et al., 2000). Research on MAP based classification is available in Gauvain and Lee (1994), Sueiro, Arribas, Munoz, and Vidal (1999), Xu, Huang, Zhu, King, and Lyu (2007) and Xu, Huang, Zhu, King, and Lyu (2009). Furthermore, successful application of the method has been reported for a number of patterns, including breast-cancer detection (Arribas, Sueiro, & Lopez, 2007), DNA classification (Loewenster, Berman, & Hirsh, 1998), face image recognition (Xu et al., 2009), image watermark identification (Ng & Garg, 2009), and speech recognition (Huy, Takeda, & Itakura, 2005).

Traditionally, in a MAP-based classification *a posteriori* probability of each class must be calculated and compared, and from this comparison one is chosen in order to get the most probable class for unknown patterns. On the basis of Bayes rule, the prior probability and the class conditional density must also be estimated or derived from data in advance. However, direct calculation of *a posteriori* probability from data is a nontrivial task. For instance, for neural networks, *a posteriori* probability must be estimated under certain assumptions: an infinite amount of training patterns are available, the network is sufficiently complex, and the training error is able to achieve the global minimum (Richard & Lippmann, 1991).

Moreover, freedom to select an objective function to provide *a posteriori* probability is limited, since such an objective function should be strict sense Bayesian (SSB) (Arribas, Sueiro, Adali, & Vidal, 1999; Sueiro et al., 1999). From previous studies, it is clear that the SSB objective function required to estimate a *posteriori* probability is a nonlinear (Ng & Garg, 2009) and, therefore, it can be solved by only nonlinear optimization (Bertsekas, 2004).

Another well-known method to estimate a posterior probability in statistic and machine learning is logistic regression (de Souza, Queiroz, & de A Cysneiros, 2011; Lee, Lee, Abbeel, & Ng, 2006). The objective of the method is to find the best fitting model that describes a relationship between a class and samples. The model parameters or regression coefficients are determined by maximum likelihood estimation in which the optimal values are searched by using an iterative numerical method such as Newton–Raphson and Fisher algorithms. Beside the choice of optimization technique, the determination of parameter initialization is crucial factor in finding the optimal parameter value. Moreover, the parameter estimation relies heavily on having an adequate number of samples, where

* Correspondence to: Department of International Development Engineering, Graduate School of Science and Engineering, Tokyo Institute of Technology, South 6th building, Ookayama, Meguro-ku, Tokyo, 152-8552, Japan. Tel.: +81 3 5734 3497; fax: +81 3 5734 3497.

E-mail addresses: nopriadi@yahoo.com (Nopriadi), yamasita@ide.titech.ac.jp (Y. Yamashita).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.05.007

a small sample size can lead to a widely inaccurate estimate of the parameter. Another problem in using logistic regression is the assumption that the sample data is linearity in the logit and additivity, which makes the method poorly suited to large datasets.

Another proposed approach to MAP-based classification is using the kernel method; specifically, the kernel-based maximum *a posteriori* (KMAP) classification method (Xu et al., 2007, 2009). The common assumption in traditional MAP techniques is that data points satisfy a multivariate normal distribution in the input space. However, KMAP operates in a high-dimensional feature space and assumes that mapped data in the feature space follows such a distribution (Xu et al., 2009). A considerable benefit of operating in a high-dimensional feature space is that it is straightforward to discriminate data having complex distributions in the input space. Conversely, the primary drawback of the approach is that a large number of parameters have to be tuned. In practice, it is challenging to tune numerous parameters for generalization of a model.

To address the problems in current MAP-based methods, we propose a new approach to MAP-based classification. As opposed to estimating the *a posteriori* probability, we feed a surrogate function into an objective function that behaves similarly to the MAP classifier. The surrogate function is expressed in the kernel form, specifically in the Gaussian radial basis function (RBF) kernel.

We argue that our new approach has three advantages. Firstly, the method widens the freedom to choose an objective function, and is not restricted to the SSB objective function. As a consequence, a second benefit is that we can optimize the objective function by linear programming, which has many robust solvers, rather than by the nonlinear optimization that is used for any SSB objective function. Lastly, our model is less demanding than KMAP. In contrast to KMAP, which has four parameters to be tuned, we need only one parameter. As a result of the advantage, it is considered that selecting a model in an experiment is easier and that the technique opens up possibilities for further research development.

Since our new model can be optimized by using linear programming and it also uses a kernel function, we call the method "maximum a posteriori based kernel classifier trained by linear programming" (MAPLP).

The remainder of the paper is organized as follows. Section 2 explains conventional maximum *a posteriori* probability estimation. Following this explanation, the MAPLP method is then presented in Section 3. Section 4 reports on experiments and provides a discussion on these. Finally, Section 5 concludes this study and explores future work.

2. Maximum a posteriori (MAP) estimation

This section elucidates basic probability formulae and presents MAP-based classifiers that operate in an input space in addition to a high-dimensional feature space.

2.1. Basic probability formulae

Let x be data in an original input space which are generated from an unknown distribution and can be assigned into the category $y \in \{1, 2, \ldots, L\}$, where L is the number of classes. For binary classification, category y is a member of the set $\{-1, +1\}$. Furthermore, let $P(x), P(y), P(x|y)$, and $P(y|x)$ denote, respectively, a prior probability density function (pdf) of x, a prior probability of y, a conditional pdf of x given y, and an *a posteriori* of y.

Then, expectation of a function, $f(x)$, by $P(x)$ is defined as

$$E\{f\} = \int f(x)P(x)dx. \tag{1}$$

From Bayes' theorem, the joint probability of pattern, x, and class, y, is given by

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x). \tag{2}$$

By applying the sum rule of probability to $P(x)$ in Eq. (2), we can calculate the posterior probability of each class y from

$$P(y|x) = \frac{P(x|y)P(y)}{\sum_y P(x|y)P(y)}. \tag{3}$$

Finally, the MAP estimation to classify the data to an estimated class, $\hat{y}$, is defined as

$$\hat{y} = \arg\max_y \; P(y|x). \tag{4}$$

A MAP-based classification system is designed to estimate the most likely class, $\hat{y}$, for a new pattern x. As shown in Eq. (4), estimation of $P(y|x)$ plays a central role in MAP-based classification, and operates either in the input space or in the feature space.

2.2. MAP-based classifiers

Next, we summarize two instances of MAP-based classification methods; one that operates in the input space and one that operates in the high-dimensional feature space. At the end of the explanation for each method, the main problem that arises in applying the method is emphasized.

Objective functions that operate in the input space, $C(h, d)$, to estimate *a posteriori* probabilities have been explored in Sueiro et al. (1999). If a vector of functions is substituted into h and a vector that expresses a category is substituted into d, then the criterion $C(h, d)$ for a binary classification problem after Sueiro et al. (1999) can be rewritten as

$$\sum_{y\in\{+1,-1\}} E_x P(y|x) C\left(h(x), \begin{pmatrix}\delta_{y,+1}\\ \delta_{y,-1}\end{pmatrix}\right),$$

where $h(x)$ is a two-dimensional vector of functions in x that is to be optimized and δ is Kronecker's delta.

The critical issue in this approach is how to design $C(h, d)$ in order that h becomes an *a posteriori* probability;

$$h(x) = \begin{pmatrix}P(+1|x)\\ P(-1|x)\end{pmatrix}.$$

A study of general conditions for objective functions providing *a posteriori* probability estimates is found in Sueiro and Arribas (1998), and have discussed specifically for the binary class in Miller, Goodman, and Smyth (1991) and Pearlmutter and Hampshire (1990) and for the M-ary class in Sueiro et al. (1999). These studies show that such objective functions must be SSB objective functions. Moreover, any symmetric and separable SSB objective function, $C(h, d)$, can be written in the form Sueiro et al. (1999)

$$C\left(\begin{pmatrix}h_1\\ h_2\end{pmatrix}, \begin{pmatrix}d_1\\ d_2\end{pmatrix}\right) = \sum_{i=1}^{2} \int_{d_i}^{h_i} g_i(\alpha)(\alpha - d_i)d\alpha + r(d),$$

where h_i is a posterior probability of category i to be optimized, d_i is a decision vector ($d_i = \delta_{y,i}$), $g_i(\alpha)$ is any positive function ($g_i(\alpha) > 0, 0 \le \alpha \le 1$) that is independent of d_i, and $r(d)$ is an arbitrary function that is independent of h.

Thus, it is important to note that the primary problem for MAP-based classifiers that operate in the input space is that objective functions are limited to SSB objective functions; there is no other freedom of choice. Furthermore, since $\int_{d_i}^{h_i} g_i(\alpha)(\alpha-d_i)d\alpha+r(d)$ is a nonlinear function, it can be solved by only nonlinear optimization.

Download English Version:

https://daneshyari.com/en/article/405517

Download Persian Version:

https://daneshyari.com/article/405517

Daneshyari.com