Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 2665–2673
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

A 'No Panacea Theorem' for classifier combination

Roland Hu, R.I. Damper*

ISIS Research Group, School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK

Received 9 May 2007; received in revised form 28 November 2007; accepted 31 January 2008

Abstract

We introduce the 'No Panacea Theorem' (NPT) for multiple classifier combination, previously proved only in the case of two classifiers and two classes. In this paper, we extend the NPT to cases of multiple classifiers and multiple classes. We prove that if the combination function is continuous and diverse, there exists a situation in which the combination algorithm will give very bad performance. The proof relies on constructing 'pathological' probability density distributions that have high densities in particular areas such that the combination functions give incorrect classification. Thus, there is no optimal combination algorithm that is suitable in all situations. It can be seen from this theorem that the probability density functions (pdfs) play an important role in the performance of combination algorithms, so studying the pdfs becomes the first step of finding a good combination algorithm. Although devised for classifier combination, the NPT is also relevant to all supervised classification problems.

Keywords: Probability density functions; Gaussian mixtures; 'No Free Lunch' theorems

1. Introduction

For almost any pattern recognition problem, there exist many classifiers that provide potential solutions to it. It is well established that combinations of these may provide more accurate recognition performance than any individual classifier. There is, however, little general agreement upon the underlying theory of classifier combination apart from various results and ideas scattered in the literature. A popular analysis of combination schemes is based on the well-known bias–variance dilemma [1]. Tumer and Ghosh [2] showed that combining classifiers using a linear combiner or order-statistics combiner reduces the variance of the actual decision boundaries around the optimal boundary. Kittler et al. [3] developed a common theoretical framework for a class of fixed combination schemes and gave a possible reason why the sum rule often outperforms the product rule. Notwithstanding these theoretical studies, the present paper describes some pessimistic aspects of classifier combination. We prove that there is no 'perfect' combination algorithm suitable for all situations. Such a property, called the "no panacea" principle by Kuncheva [4], appears widely acknowledged, but no strict mathematical proof exists for it.

The 'No Panacea Theorem' (NPT) for classifier combination can be regarded as a generalisation of the 'No Free Lunch' (NFL) theorems [5,6]. Wolper and Macready [6] proved that any two optimisation algorithms are equivalent when their performance is averaged across all possible probability density functions (pdfs). Wolpert [7] further extended the NFL idea to supervised learning and concluded that the performance of all learning algorithms is the same when averaging over all possible prior probability distributions, i.e., establishing the same average performance of all optimisation and supervised learning algorithms across all possible problems. There has been much subsequent work extending and generalising the NFL theorem. The reader is referred to www.no-free-lunch.org for details.

However, the NFL theorem only discusses the *average* performance of algorithms. It does not consider the problem of how good or bad the performance of a specific algorithm would be for a given probability distribution. If there exists a probability distribution that would dictate bad performance for a specified algorithm, what does it look like? This paper will address these two problems. We prove that if the combination functions are continuous and diverse, then we can construct pdfs based on Gaussian mixtures in which the combination algorithm will

* Corresponding author. Tel.: +44 23 80 594577.
E-mail addresses: hh03r@ecs.soton.ac.uk (R. Hu), rid@ecs.soton.ac.uk (R.I. Damper).

0031-3203/$30.00 © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.01.022

yield very bad performance. Although our theorem was originally constructed for multiple classifier combination, it can also be generalised to the area of supervised pattern recognition.

There has also been some research work on constructing objective functions and probability distributions for theorem proving. Oltean [8] has explicitly constructed objective functions where random search outperforms evolutionary algorithms. Antos et al. [9] construct probability distributions to prove that there does not exist a universally superior Bayes error estimation method, no matter how many simulations are performed and how large the sample sizes are. However, the objective functions and probability distributions constructed in these previous works are a little bit 'strange', i.e., they are not likely to be encountered in real-world problems. In this paper, we will prove the NPT based on constructing probability distributions of Gaussian mixtures. By virtue of the central limit theorem [10], Gaussian mixtures are good models for many real-world problems. Note that we have previously proved this theorem in the case of two classifiers and two classes [11]. In this paper, we extend the NPT to cases of multiple classifiers and multiple classes.

We begin in Section 2 by introducing necessary definitions and assumptions. Then in Section 3, we prove the NPT. Section 4 provides examples of the constructed pdfs in the simplest case of two classifiers and two classes. Section 5 provides further examples in which pdfs other than Gaussian mixtures can also make the NPT valid. The relation to the NFL theorems is discussed in Section 6. Finally, Section 7 concludes the paper and outlines consequential problems that need to be solved in the future.

2. Background

Suppose there are M classifiers, the task of each being to assign an input X to one of K classes, $\omega_1, \omega_2, \ldots, \omega_K$. Each classifier generates a set of discriminant functions (or scores), $f^1(X), f^2(X), \ldots, f^K(X)$, respectively. The decision rule in terms of discriminant functions is

$$\text{decide } X \in \omega_S \quad \text{if } S = \arg\max_{k=1}^{K} f^k(X).$$

For these M classifiers, we use $f_1^1(X), f_1^2(X), \ldots, f_1^K(X)$ to represent the scores generated by the first classifier, $f_2^1(X), f_2^2(X), \ldots, f_2^K(X)$ to represent the scores generated by the second classifier, ..., and $f_M^1(X), f_M^2(X), \ldots, f_M^K(X)$ to represent scores generated by the Mth classifier. Thus, we obtain $M \times K$ score functions. For simplicity, we will use $x_1, x_2, \ldots, x_N$ to represent these score functions ($N = M \times K$). If the input has a subscript, such as X_j, we will use $x_{1j}, x_{2j}, \ldots, x_{Nj}$ to represent its scores.

Although many different approaches to classifier combination are possible, in this paper we consider the combination of these M classifiers to be described as finding a set of combination functions $F_k(x_1, x_2, \ldots, x_N)$ $(k = 1, 2, \ldots, K)$ with the following decision rule:

$$\text{decide } X \in \omega_S \quad \text{if } S = \arg\max_{k=1}^{K} F_k(x_1, x_2, \ldots, x_N). \tag{1}$$

A combination function divides the domain D of all points $\{x_1, x_2, \ldots, x_N\}$ into K regions, denoted $D_1, D_2, \ldots, D_K$:

$$D_1 = \left\{ \{x_1, x_2, \ldots, x_N\} \,\middle|\, \arg\max_{k=1}^{K} F_k(x_1, x_2, \ldots, x_N) = 1 \right\},$$
$$D_2 = \left\{ \{x_1, x_2, \ldots, x_N\} \,\middle|\, \arg\max_{k=1}^{K} F_k(x_1, x_2, \ldots, x_N) = 2 \right\},$$
$$\vdots$$
$$D_K = \left\{ \{x_1, x_2, \ldots, x_N\} \,\middle|\, \arg\max_{k=1}^{K} F_k(x_1, x_2, \ldots, x_N) = K \right\}.$$

From decision rule (1), we know that D_i $(i = 1, 2, \ldots, K)$ is the region that the combination algorithm regards as encompassing the ith class.

We define the K joint pdfs of $x_1, x_2, \ldots, x_N$ given the input data as

$$p_1(x_1, x_2, \ldots, x_N) = P(x_1, x_2, \ldots, x_N \mid X \in \omega_1),$$
$$p_2(x_1, x_2, \ldots, x_N) = P(x_1, x_2, \ldots, x_N \mid X \in \omega_2),$$
$$\vdots$$
$$p_K(x_1, x_2, \ldots, x_N) = P(x_1, x_2, \ldots, x_N \mid X \in \omega_K).$$

Then according to our previous definitions, we can obtain the classification error rate given that the correct class is ω_i $(i = 1, 2, \ldots, K)$ as a function of p_i

$$\begin{aligned} P(\text{error} \mid \omega_i) &= \int_{\widetilde{D_i}} p_i(x_1, x_2, \ldots, x_N)\, \mathrm{d}x_1\, \mathrm{d}x_2 \cdots \mathrm{d}x_N \\ &= 1 - \int_{D_i} p_i(x_1, x_2, \ldots, x_N)\, \mathrm{d}x_1\, \mathrm{d}x_2 \cdots \mathrm{d}x_N, \end{aligned} \tag{2}$$

where $\widetilde{D_i}$ is the complement of D_i

$$\widetilde{D_i} = D_1 \cup D_2 \cup \cdots \cup D_{i-1} \cup D_{i+1} \cup \cdots \cup D_K.$$

It refers to the region in which the classification is incorrect.

Based on these definitions, the total classification error rate can be calculated as follows:

$$\begin{aligned} P(\text{error}) &= \sum_{i=1}^{K} P(\omega_i) P(\text{error} \mid \omega_i) \\ &= 1 - \sum_{i=1}^{K} P(\omega_i) \int_{D_i} p_i(x_1, \ldots, x_N)\, \mathrm{d}x_1 \cdots \mathrm{d}x_N. \end{aligned} \tag{3}$$

Here, $P(\omega_1), P(\omega_2), \ldots, P(\omega_K)$ are the prior probabilities that input data X belong to classes $\omega_1, \omega_2, \ldots, \omega_K$, respectively.

In order to build the theorem, two assumptions for the combination functions need to be added.

Assumption 1 (*Continuous assumption*). For each $k \in \{1, 2, \ldots, K\}$, the combination function $F_k(x_1, x_2, \ldots, x_N)$ is