# A 'No Panacea Theorem' for Multiple Classifier Combination

**Roland Hu and R. I. Damper**
School of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK
{hh03r|rid}@ecs.soton.ac.uk

## Abstract

*We introduce the 'No Panacea Theorem' for classifier combination in the two-classifier, two-class case. It states that if the combination function is continuous and diverse, there exists a situation in which the combination algorithm will always give very bad performance. Thus, there is no optimal algorithm, suitable in all situations. From this theorem, we see that the probability density functions (pdf's) play an important role in the performance of combination algorithms, so studying the pdf's becomes the first step in finding a good algorithm.*

## 1. Introduction

For almost any pattern recognition problem, there exist many classifiers which provide potential solutions to it. Combination of these classifiers may provide more accurate recognition 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-know bias-variance dilemma [1]. Tumer and Ghosh [4] showed that combining classifiers using a linear combiner or order statistics combiner reduces the variance of the actual decision boundaries around the optimum boundary. Kittler et al. [2] developed a common theoretical framework for a class of combination schemes and gave a possible reason why the sum rule often outperforms the product rule. Notwithstanding these theoretical studies, this paper describes some 'pessimistic' aspects of classifier combination. We prove that there is no 'perfect' combination algorithm suitable for all situations. Such a property, which is called the 'no panacea' principle by Kuncheva [3], appears widely acknowledged, but no strict mathematical proof exists for it.

The 'No Panacea Theorem' for classifier combination can be regarded as a special case of the 'No Free Lunch' theorem. Wolpert and Macready [8] proved that no optimisation algorithm exists which is always better than any other. In [7], Wolpert further extended the 'No Free Lunch' idea to supervised learning and concluded that the performance of any learning algorithm is the same when averaging over all prior probability distributions, which is very similar to the conclusion of this paper.

Another origin of our proof comes from the Chentsov theorem [5] in statistics, which states that for any estimator $\epsilon_l(A)$ of an unknown probability measure defined on the Borel subsets $A \subset (0,1)$, there exists a measure $P$ for which $\epsilon_l(A)$ does not provide uniform convergence. Our method to construct the probability density functions in Section 3 is very similar to the proof of this theorem.

## 2 Background

Suppose there are two classifiers, each assigning an input $X$ to one of two classes, $\omega_1$ and $\omega_2$, as described by two score functions $f_1(X)$ and $f_2(X)$. The decision rule of the $k$th classifier ($k = 1, 2$) is:

$$\text{Decide} \begin{cases} X \in \omega_1 & : \text{ if } f_k(X) > 0 \\ X \in \omega_2 & : \text{ if } f_k(X) < 0 \end{cases}$$

We will use $x_1$ and $x_2$ to represent $f_1(X)$ and $f_2(X)$. If the input data has a subscript, such as $X_i$, we will use $x_{1i}$ and $x_{2i}$ to represent $f_1(X_i)$ and $f_2(X_i)$. Based on these definitions, every combination algorithm defines a combination function $F(x_1, x_2)$, with the decision rule:

$$\text{Decide} \begin{cases} X \in \omega_1 & : \text{ if } F(x_1, x_2) > 0 \\ X \in \omega_2 & : \text{ if } F(x_1, x_2) < 0 \end{cases} \quad (1)$$

A combination function divides the domain of all points $\{x_1, x_2\}$ into two regions, denoted by $D_{\omega_1}$ and $D_{\omega_2}$.

$$D_{\omega_1} = \{\{x_1, x_2\} | F(x_1, x_2) > 0\}$$
$$D_{\omega_2} = \{\{x_1, x_2\} | F(x_1, x_2) < 0\}$$

Finally, we define the joint probability density functions of $x_1, x_2$ given that the input data satisfy:

$$p_1(x_1, x_2) = P(x_1, x_2|X \in \omega_1)$$
$$p_2(x_1, x_2) = P(x_1, x_2|X \in \omega_2)$$

According to our previous definitions, we can obtain the classification error rate as a function of $p_1$ and $p_2$:

$$P(\text{error}) = P(\omega_1)P(\text{error}|\omega_1) + P(\omega_2)P(\text{error}|\omega_2) \quad (2)$$

$$\text{where } P(\text{error}|\omega_1) = \iint_{D_2} p_1(x_1, x_2)dx_1dx_2$$
$$P(\text{error}|\omega_2) = \iint_{D_1} p_2(x_1, x_2)dx_1dx_2$$

Here $P(\omega_1)$ and $P(\omega_2)$ are the prior probability that an input data $X$ belongs to $\omega_1$ and $\omega_2$ respectively.

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

**Assumption 1** *[Continuous assumption]. The combination function $F(x_1, x_2)$ is continuous with respect to $x_1$ and $x_2$. More specifically, for any point $\{x_{10}, x_{20}\}$, and for any $\epsilon > 0$, there is a $\delta = \delta(\epsilon) > 0$ such that*

$$\text{If } \sqrt{(x_1 - x_{10})^2 + (x_2 - x_{20})^2} < \delta, \text{ then } |F(x_1, x_2) - F(x_{10}, x_{20})| < \epsilon$$

A useful corollary can be deduced from the continuous assumption which will be used in our proof of the 'No Panacea Theorem'.

**Corollary 1** *If $F(x_1, x_2)$ is continuous, then for any $F(x_{10}, x_{20}) > 0$ (or $< 0$), there exists an open ball $B(X_0, \delta)$ so that for every $\{x_1, x_2\} \in B(X_0, \delta)$, $F(x_1, x_2) > 0$ (or $< 0$).*

Here $B(X_0, \delta)$ refers to the set of points $\{x_1, x_2\}$ which satisfies the following relationship.

$$\sqrt{(x_1 - x_{10})^2 + (x_2 - x_{20})^2} < \delta$$

**Assumption 2** *[Diverse assumption]. The combination function takes both positive and negative values. That is,*

$$\exists\{x_{1\omega_1}, x_{2\omega_1}\}, \text{ such that } F(x_{1\omega_1}, x_{2\omega_1}) > 0$$
$$\exists\{x_{1\omega_2}, x_{2\omega_2}\}, \text{ such that } F(x_{1\omega_2}, x_{2\omega_2}) < 0$$

This assumption is called diverse because it guarantees the combination function makes diverse decisions.

## 3 Proof of the 'No Panacea Theorem'

We first define the characteristics of the training data. Given $M + N$ training data pints, we assume the first $M$ points, $X_1, X_2, \ldots, X_M$, belong to $\omega_1$ and the following $N$, $X_{M+1}, X_{M+2}, \ldots, X_{M+N}$, to $\omega_2$. Their scores given by the two classifiers are represented as $\{x_{11}, x_{21}\}$, $\{x_{12}, x_{22}\}$, ..., $\{x_{1(M+N)}, x_{2(M+N)}\}$. Now we have the following theorem.

**Theorem 1** *Given the $M + N$ training points as described above, if a combination function $F(x_1, x_2)$ satisfies the continuous and diverse assumptions, then there exist two continuous probability density functions $p_1(x_1, x_2)$ and $p_2(x_1, x_2)$ such that for any given $P > 0$ and any $\epsilon \in (0, 1)$, the following two properties holds:*

1. $p_1(x_{1i}, x_{2i}) > P, \quad i = 1, 2, \ldots, M$
   $p_2(x_{1i}, x_{2i}) > P, \quad i = M + 1, M + 2, \ldots, M + N$

2. *$P(error)$, which is calculated by equation (2), is greater than $1 - \epsilon$.*

For this two-classifier, two-class problem, every combination algorithm needs to generate a combination function $F(x_1, x_2)$ based on the training data $X_1, X_2, \ldots, X_{M+N}$. But, as can be seen from equation (2), the performance of the combination algorithm is not only associated with the function $F(x_1, x_2)$, but also associated with the probability density functions $p_1(x_1, x_2)$ and $p_2(x_1, x_2)$. However, the pdf's $p_1(x_1, x_2)$ and $p_2(x_1, x_2)$ can not be completely revealed by finite training data, so for any combination algorithm, there may exist some pdf's which make the performance very bad. Thus, properties (1) and (2) give criteria for how bad the performance of the combination may be. Property (1) states that there exist pdf's which make the density on the training data very high. Property (2) states that such pdf's also make the error rate very high. Generally, these two properties indicates that for any combination algorithm which satisfies the continuous and diverse assumptions, there exist pdf's which can very possibly generate the training data, but the combination function trained by these data may give very poor performance. The main idea of our proof is to generate Gaussian mixture distributions which have high density in the 'wrong' areas (where the combination function gives incorrect classification).

*Proof.* Because $F(x_1, x_2)$ satisfies the diverse assumption, there exist two points $\{x_{1\omega_1}, x_{2\omega_1}\} \in \omega_1$ and $\{x_{1\omega_2}, x_{2\omega_2}\} \in \omega_2$, so that $F(x_{1\omega_1}, x_{2\omega_1}) > 0$ and $F(x_{1\omega_2}, x_{2\omega_2}) < 0$. Because $F(x_1, x_2)$ is continuous, by Corollary (1), there exist $\delta_1$ and $\delta_2$ which make $B(X_{\omega_1}, \delta_1) \subseteq D_1$ and $B(X_{\omega_2}, \delta_2) \subseteq D_2$.

Now we will prove that the following forms of $p_1$ and $p_2$ satisfy these properties.

$$p_1(x_1,x_2) = \left(\frac{M}{M+1}\epsilon\right) t_{11}(x_1,x_2) + \left(1-\frac{M}{M+1}\epsilon\right) t_{12}(x_1,x_2)$$

$$p_2(x_1,x_2) = \left(\frac{N}{N+1}\epsilon\right) t_{21}(x_1,x_2) + \left(1-\frac{N}{N+1}\epsilon\right) t_{22}(x_1,x_2)$$

Here $t_{11}$ and $t_{21}$ are mixture Gaussian distributions, and $t_{12}$ and $t_{22}$ are Gaussian distributions.

$$t_{11}(x_1,x_2) = \frac{1}{M}\frac{1}{2\pi\sigma_1^2}\sum_{j=1}^{M} e^{\left\{-\frac{(x_1-x_{1j})^2+(x_2-x_{2j})^2}{2\sigma_1^2}\right\}}$$

$$t_{12}(x_1,x_2) = \frac{1}{2\pi\sigma_{\omega_2}^2} e^{\left\{-\frac{(x_1-x_{1\omega_2})^2+(x_2-x_{2\omega_2})^2}{2\sigma_{\omega_2}^2}\right\}}$$

$$t_{21}(x_1,x_2) = \frac{1}{N}\frac{1}{2\pi\sigma_2^2}\sum_{j=M+1}^{M+N} e^{\left\{-\frac{(x_1-x_{1j})^2+(x_2-x_{2j})^2}{2\sigma_2^2}\right\}}$$

$$t_{22}(x_1,x_2) = \frac{1}{2\pi\sigma_{\omega_1}^2} e^{\left\{-\frac{(x_1-x_{1\omega_1})^2+(x_2-x_{2\omega_1})^2}{2\sigma_{\omega_1}^2}\right\}}$$

$\sigma_1$, $\sigma_2$, $\sigma_{\omega_1}$ and $\sigma_{\omega_2}$ are parameters to be decided. In the following, we prove that when $\sigma_1$, $\sigma_2$, $\sigma_{\omega_1}$ and $\sigma_{\omega_2}$ are small enough, property (1) and (2) will hold.

We firstly prove that when $\sigma_1$ is small enough, $P_1(x_{1i},x_{2i}) > P$ for $i = 1,2,\ldots,M$:

$$\begin{aligned} p_1(x_{1i},x_{2i}) &\geq \frac{\epsilon}{2\pi(M+1)\sigma_1^2} e^{\left\{-\frac{(x_{1i}-x_{1i})^2+(x_{2i}-x_{2i})^2}{2\sigma_1^2}\right\}} \\ &= \frac{\epsilon}{2\pi(M+1)\sigma_1^2} \end{aligned}$$

So if we choose:

$$\sigma_1 < \sqrt{\frac{\epsilon}{2\pi(M+1)P}} \tag{3}$$

we will always have $p_1(x_{1i},x_{2i}) > P$. The same deduction can be used to prove that if:

$$\sigma_2 < \sqrt{\frac{\epsilon}{2\pi(N+1)P}} \tag{4}$$

then $p_2(x_{1i},x_{2i}) > P$ $(i = M+1, M+2, \ldots, M+N)$. Thus, we have proved property (1).

For property (2), we will prove that when $\sigma_{\omega_1}$ and $\sigma_{\omega_2}$ are small enough, both $P(\text{error}|\omega_1)$ and $P(\text{error}|\omega_2)$ are greater than $1-\epsilon$.

$$\begin{aligned} P(\text{error}|\omega_1) &= \iint_{D_2} p_1(x_1,x_2)dx_1dx_2 \\ &\geq \iint_{D_2} \left(1-\frac{M}{M+1}\epsilon\right) t_{12}(x_1,x_2)dx_1dx_2 \end{aligned}$$

Since $B(X_{\omega_2},\delta_2) \subseteq D_2$, we have:

$$\begin{aligned} P(\text{error}|\omega_1) &\geq \left(1-\frac{M}{M+1}\epsilon\right) \times \iint_{B(X_{\omega_2},\delta_2)} t_{12}(x_1,x_2)dx_1dx_2 \\ &= \left(1-\frac{M}{M+1}\epsilon\right) \frac{1}{2\pi\sigma_{\omega_2}^2} \times \iint_{B(X_{\omega_2},\delta_2)} e^{\left\{-\frac{(x_1-x_{1\omega_2})^2+(x_2-x_{2\omega_2})^2}{2\sigma_{\omega_2}^2}\right\}} dx_1dx_2 \\ &\geq \left(1-\frac{M}{M+1}\epsilon\right) \times \left(\int_{x_{1\omega_2}-\frac{\delta_2}{\sqrt{2}}}^{x_{1\omega_2}+\frac{\delta_2}{\sqrt{2}}} \frac{1}{\sqrt{2\pi}\sigma_{\omega_2}} e^{\left\{-\frac{(x_1-x_{1\omega_2})^2}{2\sigma_{\omega_2}^2}\right\}} dx_1\right) \times \left(\int_{x_{2\omega_2}-\frac{\delta_2}{\sqrt{2}}}^{x_{2\omega_2}+\frac{\delta_2}{\sqrt{2}}} \frac{1}{\sqrt{2\pi}\sigma_{\omega_2}} e^{\left\{-\frac{(x_2-x_{2\omega_2})^2}{2\sigma_{\omega_2}^2}\right\}} dx_2\right) \\ &= \left(1-\frac{M}{M+1}\epsilon\right) \left(\int_{\frac{-\delta_2}{\sqrt{2}}}^{\frac{\delta_2}{\sqrt{2}}} e^{\left\{-\frac{x^2}{2\sigma_{\omega_2}^2}\right\}} dx\right)^2 \end{aligned}$$

For a Gaussian distribution with mean 0 and variance $\sigma$, we have the Chernoff bound [6] for the integral.

$$\begin{aligned} P(-\delta \leq X \leq \delta) &= \int_{-\delta}^{\delta} \frac{1}{\sqrt{2\pi}\sigma} e^{\left\{-\frac{x^2}{2\sigma^2}\right\}} dx \\ &\geq 1 - 2e^{\left\{-\frac{\delta^2}{2\sigma^2}\right\}} \end{aligned}$$

Thus we can finally obtain:

$$P(\text{error}|\omega_1) \geq \left(1-\frac{M}{M+1}\epsilon\right)\left(1-2e^{\left\{-\frac{\delta_2^2}{4\sigma_{\omega_2}^2}\right\}}\right)^2$$

So if we choose:

$$\sigma_{\omega_2} < \frac{\delta_2}{2\sqrt{\ln 2 - \ln\left(1-\sqrt{\frac{M+1-(M+1)\epsilon}{M+1-M\epsilon}}\right)}} \tag{5}$$

then $P(\text{error}|\omega_1)$ will be greater than $1-\epsilon$. Similarly, if:

$$\sigma_{\omega_1} < \frac{\delta_1}{2\sqrt{\ln 2 - \ln\left(1-\sqrt{\frac{N+1-(N+1)\epsilon}{N+1-N\epsilon}}\right)}} \tag{6}$$

then $P(\text{error}|\omega_2)$ will be greater than $1-\epsilon$. If both $P(\text{error}|\omega_1)$ and $P(\text{error}|\omega_1)$ are greater than $1-\epsilon$, from equation (2), the total error rate $P(\text{error})$ is also greater than $1-\epsilon$. Thus, we have proved property (2).

## 4 Example

Suppose we have four training data points, two of which belong to $\omega_1$ and two belong to $\omega_2$ (i.e., $M = 2$ and $N = 2$). The scores of the two data points from $\omega_1$ are given as $\{x_{11}, x_{21}\} = \{1, 2\}$ and $\{x_{12}, x_{22}\} = \{2, 1\}$; and the scores of the two data points from $\omega_2$ are given as $\{x_{13}, x_{23}\} = \{-1, -2\}$ and $\{x_{14}, x_{24}\} = \{-2, -1\}$. We assume that the combination function $F(x_1, x_2)$ follows the simple sum rule:

$$\text{Decide} \begin{cases} X \in \omega_1 & : \text{ if } x_1 + x_2 > 0 \\ X \in \omega_2 & : \text{ if } x_1 + x_2 < 0 \end{cases}$$

It is obvious that this rule is continuous and diverse. We can choose the corresponding $\{x_{1\omega_1}, x_{2\omega_1}\} = \{1, 1\}$, and $\{x_{1\omega_2}, x_{2\omega_2}\} = \{-1, -1\}$. For the sum rule, there is a $\delta_1 = 1$ which makes $B(\{1, 1\}, \delta_1) \in D_1$, and a $\delta_2 = 1$ which makes $B(\{-1, -1\}, \delta_2) \in D_2$. Finally, we choose $\epsilon = 0.1$ and $P = 2$.

Eqns. (3), (4), (5) and (6) yield $\sigma_1 = \sigma_2 = 0.0515$ and $\sigma_{\omega_1} = \sigma_{\omega_2} = 0.2304$. Figure 1 shows $p_1(x_1, x_2)$ and $p_2(x_1, x_2)$ obtained by setting $\sigma_1$, $\sigma_2$, $\sigma_{\omega_1}$ and $\sigma_{\omega_2}$ as above. Figure 1(a) shows that more than 90% ($1 - \epsilon = 0.9$) of the probability that the input data belong to $\omega_1$ is accumulated near the point $\{-1, -1\}$. At this point, the sum rule gives an incorrect classification. It can also be seen that high probability also exists near the training data $\{1, 2\}$ and $\{2, 1\}$, which indicates that in such a distribution, it is very possible to have these training data, but impossible to obtain correct classification by the sum rule.

It may be argued that such a 'strange' probability distribution, which is so biased in the 'wrong' areas and near the training data, is not a distribution that nature 'favours'. However, in situations which are not so extreme, we can show that a given combination rule also can not guarantee good performance. However, space precludes further examples.

## 5 Conclusions

We have proved the 'No Panacea Theorem' for classifier combination, which states that if the combination function is continuous and diverse, there exists a situation in which the combination algorithm will make very bad performance. Thus, there is no optimal combination algorithm which is suitable in any situation. Although the proof is based on the two-classifier and two-class problem, it can be generated to the case of multiple classifiers and multiple classes.

Our aim in presenting this theorem is not to criticise any particular algorithms for combining classifiers, but rather to point out the difficulties we might encounter in this area. From this theorem, we see that a good combination algorithm is not only dependent on the combination function, but also on the probability density functions, so studying the pdf's becomes the first step in finding a good combination algorithm.

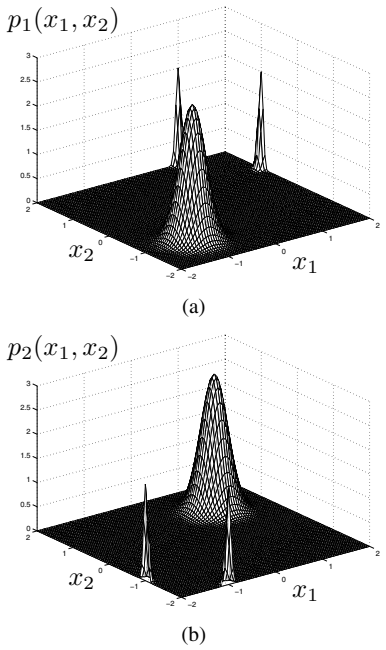


**Figure 1. An example of the probability density functions which give bad performance for combination by the sum rule. (a)** $p_1(x_1, x_2)$**; (b)** $p_2(x_1, x_2)$

## References


[1] S. Geman, E. Bienenstock, and R. Doursat. Neural network and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[2] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[3] L. I. Kuncheva. *Combining Pattern Classifiers–Methods and Algorithms*, chapter 4, pages 111–149. John Wiley & Sons, Inc, 2004.

[4] K. Tumer and J. Ghosh. Analysis of decision boundaries in linearly combined neural classifiers. *Pattern Recognition*, 29(2):341–348, 1996.

[5] V. N. Vapnik. *Statistical Learning Theory*, chapter 2, pages 59–78. John Wiley & Sons, INC., 1998.

[6] S. G. Wilson. *Digital Modulation and Coding*, chapter 2, pages 18–140. Prentice Hall, 1996.

[7] D. H. Wolpert. The supervised learning no-free-lunch theorems. In *Proceedings of the 6th Online World Conference on Soft Computing in Industrial Applications*, 2001. Available at `http://www.no-free-lunch.org/Wolp01a.pdf`.

[8] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.