How to choose the number of classifiers to combine in an ensemble using the maximum combining rule

Helena Ferreira
Dept. Matemática, Univ. Beira Interior, Portugal.
Email: ferreira@ubista.ubi.pt
Luís A. Alexandre
Dept. Informática, Univ. Beira Interior, Portugal.
Email: lfbaa@noe.ubi.pt

Abstract

Consider the classification obtained by using the combination of n classifiers using the maximum function. Since the use of a large number n of classifiers is discouraging, we propose the use of a number $L_n(< n)$ of classifiers which is the smallest number for which the maximum of the *a posteriori* probabilities exceeds p_n, with $p_n \in (0,1)$ previously fixed. We describe the probability law of the random variable L_n and approximate the probability of making a different decision with L_n classifiers than the one that would be made using the total number of classifiers, n. We present some experiments in order to compare the classification using n and L_n classifiers.

Keywords: extreme values, order statistics, combining classifiers, ensembles, maximum combining rule

1 Introduction

It is well known that better results can be obtained in a classification problem if instead of using a single classifier, the combined output of several classifiers is used [6, 10, 9].

There are many ways to do such a combination. One of the possibilities is to use an ensemble of classifiers which consists of a set of identical classifiers. In our case we use neural networks only differing in their initialization [5].

After choosing an ensemble there are still several ways to combine the outputs of the classifiers: sum, product, median, maximum, minimum, using a neural network, etc [7, 1, 2]. In this paper we study the combined output of ensembles using the maximum.

Consider that the number of classifiers that is adequate to use in a problem is n large. We propose to study how many of the n classifiers should be combined, in order to obtain the same decision that would be obtained with the n classifiers.

This paper is organized as follows: the next section presents the notation and basic ideas used in the rest of the paper. Section 3 introduces the model for the problem. In section 4 the main results are derived. Section 5 presents some experiments and the final section lists our conclusions.

2 Notation

Consider a classification problem with r classes. Let $C_i, i = 1, \ldots, n$, represent identical and independent classifiers that produce approximations to the r *a posteriori* probabilities.

We represent the maximum of the *a posteriori* probabilities of classifier i on the pattern X by $Y_i = Y_i(X)$, where X is a continuous random vector and $i = 1, \ldots, n$. The r.v. $Y_1, \ldots, Y_n$ are i.i.d. as a r.v. Y with distribution function F with support (α_F, ω_F).

Consider the classification obtained by using the combination of the n classifiers using the maximum function, i.e., for $j = 1, \ldots, r$,

$$\begin{aligned} &X \in \omega_j \ if \ \max_{i=1,\ldots,n} P_i(\omega_j|X) = \\ &\max_{s=1,\ldots,r} \max_{i=1,\ldots,n} P_i(\omega_s|X) = \max_{i=1,\ldots,n} Y_i \end{aligned} \tag{1}$$

where $P_i(\omega_j|X)$ is the *a posteriori* probability for the j-th class ω_j given by the i-th classifier.

Since the use of a large number n of classifiers (say more than 100) is discouraging, we propose the use of a number $L_n (< n)$ of classifiers. This number is obtained from a level p_n for the maximum of the *a posteriori* probabilities of the classifiers. More precisely, the number L_n of classifiers used is the smallest number for which $\max_{i=1,\ldots,L_n} Y_i > p_n$, with $p_n \in (0, 1)$ previously fixed. If Y_1 is greater than the level p_n then we only consider the decision of C_1. Otherwise we use another classifier for the combination (C_2). If $max(Y_1, Y_2)$ is larger than p_n we use as an output the decision obtained by combining C_1 and C_2. Otherwise we add a new classifier, C_3, and so on. The number of classifiers used in the final combination is L_n such that

$$\max_{i=1,\ldots,L_{n-1}} Y_i \leq p_n \ and \ \max_{i=1,\ldots,L_n} Y_i > p_n$$

Several questions arise:

- How should p_n be defined ?
- What is the probability law of the r.v. L_n ?
- What is the probability that the first L_n classifiers give a different decision than the one obtained by combining the total n classifiers ?

Since the combination of $L_n < n$ classifiers is interesting only when the combination of a large number n of classifiers is not possible (or feasible) we will answer the above questions by analyzing the asymptotic behavior of p_n and L_n when $n \to \infty$. The obtained probabilities are thus limits of sequences of probabilities that are interpreted as approximations to the answers for the above questions.

3 The model

Our proposal is to compare the decision criteria in expression 1 with the criteria of the same type but using L_n classifiers, that is, for each $j = 1, \ldots, r$,

$$\begin{array}{l} X \in \omega_j \; if \;\; \max_{i=1,\ldots,L_n} P_i(\omega_j|X) = \\ \max_{s=1,\ldots,r} \max_{i=1,\ldots,L_n} P_i(\omega_s|X) = \max_{i=1,\ldots,L_n} Y_i \end{array} \tag{2}$$

where L_n represents the r.v. 'index of the first occurrence of a value of $Y_1, \ldots, Y_n$ being larger than p_n'.

The reference level $p_n \in (0,1)$ for the maximum of the *a posteriori* probabilities Y_i, determines the behavior of the L_n classifiers to combine. The p_n will be defined such that the number of expected values between $Y_1, \ldots, Y_n$ that exceed p_n is asymptotically constant, when $n \to \infty$. That is,

$$\lim_{n \to +\infty} E \sum_{i=1}^{n} \mathbb{I}_{\{Y_i > p_n\}} = \tau. \tag{3}$$

where $\mathbb{I}_{\{.\}}$ denotes the indicator function.

The previous condition is not strong in the current context, as we shall discuss below, and it is necessary and sufficient for the convergence of the following probability sequence $\{P(\max_{i=1,\ldots,n} Y_i > p_n)\}_{n \geq 1}$. It has been proved [4] that expression 3 is equivalent to

$$\lim_{n \to +\infty} P(\max_{i=1,\ldots,n} Y_i \leq p_n) = exp(-\tau).$$

Since the Y_i, $i = 1, \ldots, n$, all have the same distribution, we can rewrite expression 3 as to make p_n evident. We have

$$E \sum_{i=1}^{n} \mathbb{I}_{\{Y_i > p_n\}} = nE\mathbb{I}_{\{Y > p_n\}} = n(1 - F(p_n))$$

and therefore

$$p_n = F^{-1}(1 - \frac{\tau}{n} + o(\frac{1}{n})), \tag{4}$$

when it exists. For distributions F without discontinuities in ω_F, that is, such that

$$\lim_{x \to +\omega_F} \frac{1 - F(x)}{1 - F(x^-)} = 1$$

we can always choose p_n as $p_n(x) = a_n x + b_n$, with $x \in \mathbb{R}$, $\{a_n > 0\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$ sequences of real numbers [8]. This happens for several distributions that can model each Y_i: uniform, polynomial, truncated exponential, amongst other continuous distributions.

After having defined L_n and $p_n(x) = a_n x + b_n, x \in \mathbb{R}$, we can add:

- since generally the distribution of the a posteriori maxima Y_i is unknown, for the practical use of expression 2 we will estimate p_n from the order statistic $Y_{n - [\frac{\tau}{n}]:n}$ on a sample of dimension n for Y and a fixed positive real τ;

- the criterium in expression 2 will be preferable over the criterium in expression 1 if $L_n < n$ and this only happens if $\max_{i=1,\ldots,n} Y_i > p_n$;

4 Theoretical results

The comparison between 1 and 2 is based on the comparison of the asymptotic behavior of L_n and the r.v. L_n^*, which is equal to the index where the maximum of Y_i is reached for the first time. This is, the comparison between the index of the first time that the value of Y_i is larger than p_n and the index of the first time that the maximum of the of the probabilities Y_i is obtained. In fact, we can consider that the maximum value of the *a posteriori* probabilities Y_i is reached only once, since the probability that this does not happen converges to zero as $n \to \infty$ [4].

The motivation for such comparison is the fact that

$$\max_{i=1,\ldots,n} Y_i = \max_{i=1,\ldots,L_n^*} Y_i$$

and therefore we want to evaluate the possibility of L_n being close to L_n^*.

We now define the variables that are used in the probabilistic results

$$L_n = \begin{cases} inf\{i \in \{1,\ldots,n\} : Y_i > p_n\} & if \ M_n > p_n \\ n & if \ M_n \leq p_n \end{cases}$$

where $M_n = \max_{i=1,\ldots,n} Y_i$,

$$L_n^* = \sum_{j=1}^{n} j \mathbb{I}_{\{Y_j = \max_{i=1,\ldots,n} Y_i\}}$$

which is a.s. defined.

If the event $\{L_n = L_n^* < n\}$ occurs, then both 1 and 2 lead to the same decision and we have combined less than n classifiers.

We start by analyzing the the probability that 2 does not represent any economy when compared with 1. We have, for p_n obeying to 4

$$P(L_n = n) = P(M_n \leq p_n) + P(M_{n-1} \leq p_n, Y_n > p_n) = \\ F^n(p_n) + F^{n-1}(p_n)(1 - F(p_n)),$$

and it follows that

$$\lim_{n \to +\infty} P(L_n = n) = exp(-\tau).$$

This way, an approximation to $P(L_n = n)$ can be regulated by the choice of τ: for classifiers with an expected total number of values larger than p_n approximately equal to a small τ, the probability of $L_n < n$ is high. Let us calculated the probability of criterium 2 being used with $j < n$ classifiers and the expected number of classifiers to combine. We have

$$P(L_n = j | M_n > p_n) = \frac{P(Y_1 \leq p_n, \ldots, Y_{j-1} \leq p_n, Y_j > p_n)}{1 - F^n(p_n)} = \\ \frac{F^{j-1}(p_n)(1 - F(p_n))}{1 - F^n(p_n)} \tag{5}$$

and

$$\lim_{n \to +\infty} E(L_n | M_n > p_n) = \\ \lim_{n \to +\infty} 1 - \frac{F^n(p_n)}{1 - F^n(p_n)} n(1 - F(p_n)) = 1 - \frac{exp(-\tau)}{1 - exp(-\tau)} \tau \tag{6}$$

if p_n satisfies 4.

Proposition 4.1 *For a sequence $\{p_n\}_{n \geq 1}$ satisfying 4 and $0 < k < s < 1$ it holds*
(i) $\lim_{n \to +\infty} P(L_n \leq nk < L_n^* \leq ns) = \frac{s-k}{1-k}(1 - k + kG(x) - G^k(x))$,
(ii) for each $\epsilon \in (0, 1-k)$,
$\lim_{n \to +\infty} P(L_n \leq nk < n(k+\epsilon) < L_n^*) = \frac{1-k+\epsilon}{1-k}(1 - k + kG(x) - G^k(x))$,
where $G(x) = exp(-\tau(x))$.

Proof: (i) If $M_{i,j} = \max_{i \leq s \leq j} X_s$, we can write

$$\begin{array}{l} P(L_n \leq nk < L_n^* \leq ns) = \\ P(M_{[nk]+1,[ns]} \geq M_{1,[nk]}, M_{[nk]+1,[ns]} \geq M_{[ns]+1,n}, M_{1,[nk]} > p_n) = \\ P(a_n^{-1}(M_{[nk]+1,[ns]} - b_n) \geq a_n^{-1}(M_{1,[nk]} - b_n), \\ a_n^{-1}(M_{[nk]+1,[ns]} - b_n) \geq a_n^{-1}(M_{[ns]+1,n} - b_n), a_n^{-1}(M_{1,[nk]} - b_n) > x). \end{array} \quad (7)$$

For $i = i(n), j = j(n)$ such that $\lim_{n \to +\infty} \frac{j(n)-i(n)}{n} = \alpha$, it holds

$$\lim_{n \to +\infty} P(a_n^{-1}(M_{i(n),j(n)} - b_n) \leq x) = P(U_\alpha \leq x) = G^\alpha(x), x \in \mathbb{R}, \quad (8)$$

and $M_{i,j}$ is asymptotically independent of $M_{i',j'}$ when $[i,j] \bigcap [i',j'] = \emptyset$ (see [8]). Therefore, by using 8 in 7 we get

$$\begin{array}{l} \lim_{n \to +\infty} P(L_n \leq nk < L_n^* \leq ns) = \\ P(U_{s-k} \geq U_k, U_{s-k} \geq U_{1-s}, U_k > x) = \\ \int_x^{+\infty} P(x < U_k \leq y) P(U_{1-s} \leq y) dP_{U_{s-k}}(y) = \\ \int_x^{+\infty} (G^k(y) - G^k(x)) G^{1-s}(y) dG^{s-k}(y) = \\ \frac{s-k}{1-k}(1 - k + kG(x) - G^k(x)). \end{array}$$

(ii) The proof of (ii) is omitted since it follows the same lines of argument as above with the obvious modifications.$\triangle$

For a fixed $k \in (0,1)$, the limit of $P(L_n \leq nk < n(k+\epsilon) < L_n^*)$ increases with ϵ, as expected. However, for each ϵ, we can choose small values of $\tau = -logG(x)$, which correspond to high values of p_n, in order to get small values for the approximation $\frac{1-k+\epsilon}{1-k}(1 - k + kG(x) - G^k(x))$, that is, small values for the probability of the event 'the proportion of classifiers L_n/n differs from L_n^*/n more than ϵ' . In fact, $f(k,\epsilon,\tau) = \frac{1-k+\epsilon}{1-k}(1-k+ke^{-\tau}-e^{-k\tau})$ decreases to zero when $\tau \to 0$. Therefore, for practical proposes, we must choose high values for the 'stop *a posteriori* probabilities' p_n in order to have small probability that criteria 1 and 2 will not lead to the same decision.

We present now an overview of the answers to the questions posted at the end of section 2. If we choose for p_n the $(1-\frac{\tau}{n})-th$ quantile of the distribution of Y then the smallest number L_n for which $\max_{i=1,\ldots,L_n} Y_i > p_n$, conditional on $\max_{i=1,\ldots,n} Y_i > p_n$, has a distribution described by expressions 5 and 6. Moreover, the probability that the combination of $L_n \leq nk, k \in (0,1)$, classifiers give a different decision than the one obtained by combining the n classifiers, can be upper bounded by $1 - k + ke^{-\tau} - e^{-k\tau}$.

5 Experiments

We intend to illustrate the results with experiments showing some of the factors involved in the theory derived in the previous sections.

We use an artificial data set with 2 features and 17 classes. These classes were generated using gaussian distributions. The classification problem is complex in the sense that there is a lot of overlaping between the classes. The data set is plotted in figure 1.

Note that we are not interested in the classification error. Our goal is to find out how a combination of L_n classifiers will behave differently from all n (or L_n^*) classifiers.

For each experiment we define the value of n (the maximum number of classifiers) and the value of τ (which should be small, as mentioned above).

The classifiers used are all multi-layer perceptrons (MLPs) with the same topology: 25 neurons in the hidden layer and 17 in the output layer. They are trained for 100 epochs using resilient back-propagation [3]. The training set had 500 patterns and the test set had 100 patterns. These sets were randomly generated for each experiment and were mutually exclusive.

With this experiment we find several parameters which are listed in table 1. These are:

- the value of p_n;
- the number of patterns for which one of the n classifiers exceeded p_n (Valid);
- the number of patterns for which $L_n = L_n^*$;
- the proportion of the valid patterns which have $L_n = L_n^*$ (Prop.);
- the average and the standard deviation of L_n.

By looking at the results we can make several remarks:

- as τ decreases p_n increases;
- as p_n increases the number of valid cases decreases since it becomes less probable that the *a posteriori* probabilities exceed the value of p_n;
- as p_n increases the number of cases where $L_n = L_n^*$ increases. This is reasonable since we are demanding a higher value for the maximum of the *a posteriori* probabilities and thus the necessary value will be harder to obtain requiring more classifiers in such a way that L_n becomes near L_n^*;

Table 1: Results for the several experiments

Experiment	1	2	3
n	200	200	200
Tau	3	6.67	10
p_n	0.891	0.749	0.479
Valid	23	33	100
$L_n = L_n^*$	15	15	6
Prop.	65.2 %	45.5 %	6%
Ave(L_n)	33.70	49.94	3.24
StD(L_n)	38.35	44.76	5.02

- the average of L_n should decrease when we decrease p_n, but that was not the case for the second experiment where this value increased when compared to the value of experiment 1. This disagreement can always occur in any experiment used to illustrate *probabilistic* results.

6 Conclusion

By choosing an order statistic from a sample of the maximum of the *a posteriori* probability distribution we obtain:

- a stopping criterium for the number of classifiers to combine;
- approximation for the probability of different decision than the one obtained by combining the total number of classifiers.

We presented experiments illustrating these results.

We are currently extending these results to the case of combining using simple average instead of the maximum.

References

[1] L.A. Alexandre, A.C. Campilho, and M. Kamel. Combining independent and unbiased classifiers using weighted average. In *Proceedings of the 15th International Conference on Pattern Recognition*, volume 2, pages 495–498, Barcelona, Spain, September 2000. IEEE Press.

[2] L.A. Alexandre, A.C. Campilho, and M. Kamel. On combining classifiers using sum and product rules. *Pattern Recognition Letters*, 22(12):1283–1289, October 2001.

[3] H. Demuth and M. Beale. *Neural Network Toolbox User's Guide.* The MathWorks, Inc., 1998.

[4] J. Galambos. *The assymptotic theory of extreme order statistics.* New York: Wiley, 1978.

[5] L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Trans. PAMI*, 12(10):993–1001, October 1990.

[6] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: a review. *IEEE Trans. PAMI*, 22(1):4–37, January 2000.

[7] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Trans. PAMI*, 20(3):226–239, March 1998.

[8] M.R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes.* Springer-Verlag, 1983.

[9] D. Opitz and R. Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, (11):169–198, 1999.

[10] N. Ueda. Optimal linear combination of neural networks for improving classification performance. *IEEE Trans. PAMI*, 22(2):207–215, February 2000.

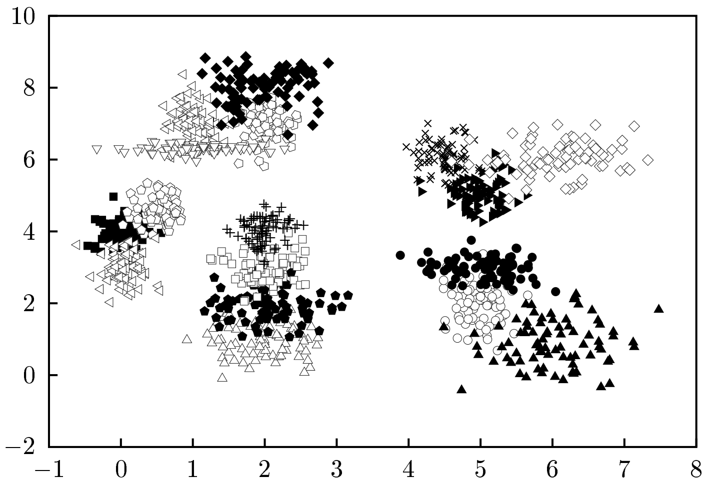

Figure 1: Data set for the 17 class artificial problem.