Combining Independent and Unbiased Classifiers Using Weighted Average

Luís A. Alexandre
Dept. Matemática/Informática
Univ. Beira Interior and
Instituto de Eng. Biomédica, Portugal
lfbaa@noe.ubi.pt

Aurélio C. Campilho
Instituto de Eng. Biomédica
and Fac. Engineering,
Univ. Porto, Portugal
campilho@fe.up.pt

Mohamed Kamel
Dept. of Systems Design Eng.
Univ. of Waterloo
Ontario, Canada
mkamel@watfor.uwaterloo.ca

Abstract

In a classification problem, improved accuracy can be obtained in many situations by using the combination of several classifiers instead of a single one. In [10], the error reduction that can be obtained by combining unbiased classifiers with independent errors using a simple average, was derived. We present an extension of this result by finding the improvement obtained when combining classifiers using weighted average. We also prove that for unbiased classifiers with independent errors the best combination of N classifiers corresponds to a weighted average, where the combination coefficient of each classifier is equal to $\frac{1}{N}$. This means that in these cases the simple average should be used. We present experiments illustrating our results.

1. Introduction

It is well known that in many situations combining the output of several classifiers leads to an improved classification accuracy [1, 4, 6]. In this paper we show what is the reduction expected when combining classifiers using a weighted average. In [10] the error reduction for the simple average, considering unbiased independent classifiers was presented. We extended this result by showing what is the error reduction for the weighted average, also considering unbiased and independent classifiers. We also show that when considering classifiers with these properties, the weights used should be $\frac{1}{N}$ for N classifiers, which means that the simple average should be used.

We present experiments confirming the theoretical results. We compare the performance of the combination of averaging and weighted averaging of four identical classifiers. In the weighted averaging the coefficients are randomly selected. This is used to illustrate that the simple average is better than any other weight combination, for this type of classifiers.

In section 2 we define the basic concepts needed for the derivation in section 3, and present the results for the simple average of classifiers. In section 3 we derive the expression of the error for the weighted averaging. In section 4, we look for the best choice of combination weights. In section 5, we present some experiments, illustrating the theoretical assertions of previous sections. In the last section we discuss the results obtained.

2. Basic Concepts

2.1. Single classifier

Consider a single classifier whose outputs are expected to approximate *a posteriori* class probabilities. The decision boundaries produced by this classifier are expected to be close to the Bayesian decision boundaries. We consider the problem of overlapping classes. We will also consider that the input to the classifier is one-dimensional (without loss of generality [9]).

The total classification error, E_{tot}, is the sum of the Bayes error, E_{bay}, (which is the theoretical limitation to the performance a classifier or combination of classifiers [2]) with an added error, E_{add}: $E_{tot} = E_{bay} + E_{add}$.

The boundary between classes i and j is the point x^* such that $p(c_i|x^*) = p(c_j|x^*)$. We call x^* the Bayes boundary. The approximation to the *a posteriori* probability $p(c_i|x)$ provided by the classifier is $f_i(x) = p(c_i|x) + \epsilon_i(x)$ where $\epsilon_i(x)$ represents the error introduced by the classifier. The boundary produced by the classifier is the point x_b such that $f_i(x_b) = f_j(x_b)$. The difference between the real and the obtained boundaries is $b = x_b - x^*$.

The added error region due to the error of the classifier is

$$A(b) = \int_{x^*}^{x^*+b} (p(c_j|x) - p(c_i|x))dx.$$

The expected added error region $EA(b)$, which we will call the added error E_{add}, is

$$E_{add} = EA(b) = \int_{-\infty}^{\infty} A(b) f_b(b) db.$$

$f_b(x)$ is the density for the variable b. This is unknown.

2.2. Average of N classifiers

Tumer in [10] showed that the added error using a simple average of N classifiers, E_{add}^{ave}, is given by

$$E_{add}^{ave} = \frac{E_{add}}{N}. \quad (1)$$

This expression is valid for unbiased and independent classifiers. In the following section we show the result for the weighted average, again for unbiased and independent classifiers.

3. Weighted average of N classifiers

In this derivation we will follow closely [10]. Consider that w_j is a positive value used to weight the contribution of classifier j to the combined *a posteriori* probability. The approximation to the *a posteriori* probabilities is now given by

$$f_i^{wa}(x) = \sum_{j=1}^{N} w_j f_i^j(x)$$

where $f_i^j(x)$ represents the approximation to the *a posteriori* probability $p(c_i|x)$ given by classifier j.

We make the following normalization assumption on the weights $\sum_{j=1}^{N} w_j = 1$. The approximation to the *a posteriori* probabilities given by the combination of classifiers is

$$f_i^{wa}(x) = \sum_{j=1}^{N} w_j (p(c_i|x) + \epsilon_i^j(x))$$

where the superscript wa stands for 'weighted average'.

We split the error term in two: the bias β and a random noise η, as follows: $\epsilon_i^j(x) = \beta_i^j + \eta_i^j(x)$.

Considering that there is no bias in the classifiers' estimation we have $\beta_i^j = 0$ and

$$f_i^{wa}(x) = \sum_{j=1}^{N} w_j (p(c_i|x) + \eta_i^j(x)) = p(c_i|x) + \eta_i^{wa}(x)$$

with

$$\eta_i^{wa}(x) = \sum_{j=1}^{N} w_j \eta_i^j(x). \quad (2)$$

The E_{add}^{wa} is the expected value of the added error region $A(b^{wa})$, which is

$$A(b^{wa}) = \int_{x^*}^{x^*+b^{wa}} (p(c_j|x) - p(c_i|x)) dx.$$

This gives E_{add}^{wa} as

$$E_{add}^{wa} = EA(b^{wa}) = \int_{-\infty}^{\infty} A(b^{wa}) f_{b^{wa}}(b^{wa}) db^{wa}.$$

3.1. Finding $A(b^{wa})$

We will find an expression for $A(b^{wa})$ by making an approximation to the *a posteriori* probability around x^*. We obtain

$$p(c_k|x) = p(c_k|x^*) + (x - x^*) p'(c_k|x^*) + \ldots$$

If we ignore the terms represented by the dots, we have a linear approximation. This way we can calculate $A(b^{wa})$,

$$A(b^{wa}) = \int_{x^*}^{x^*+b^{wa}} (p(c_j|x) - p(c_i|x)) dx$$

$$= \int_{x^*}^{x^*+b^{wa}} (x - x^*)(p'(c_j|x^*) - p'(c_i|x^*)) dx = \frac{s(b^{wa})^2}{2},$$

with

$$s = p'(c_j|x) - p'(c_i|x). \quad (3)$$

We used the fact that in the boundary, $p(c_j|x^*) = p(c_i|x^*)$, by definition of boundary.

3.2. Finding the added error

Using the relation for $A(b^{wa})$, the expression for the added error becomes

$$E_{add}^{wa} = \int_{-\infty}^{\infty} \frac{s(b^{wa})^2}{2} f_{b^{wa}}(b^{wa}) db^{wa}. \quad (4)$$

Also, we can write the expression for the second moment of b^{wa}

$$M_2(b^{wa}(x)) = E(b^{wa})^2(x) = \int_{-\infty}^{\infty} (b^{wa})^2 f_{b^{wa}} db^{wa}.$$

We can use this expression to substitute in (4),

$$E_{add}^{wa} = \frac{s M_2(b^{wa})}{2}.$$

The second moment can be written as a function of the variance of b^{wa} as

$$M_2(b^{wa}) = var(b^{wa}) + M_1^2(b^{wa}).$$

Since the classifiers are unbiased, the first moment of b^{wa} is zero. This means that $M_2(b^{wa}) = var(b^{wa}) = \sigma^2_{b^{wa}}$.

Finally, we can write the added error as

$$E^{wa}_{add} = \frac{s}{2}\sigma^2_{b^{wa}}. \quad (5)$$

We want now to compare this added error with the added error in the case of the single classifier, that is given by [10]

$$E_{add} = \frac{s}{2}\sigma^2_b. \quad (6)$$

3.3. Relation between E_{add} and E^{wa}_{add}

We must find a relation between $\sigma^2_{b^{wa}}$ and σ^2_b. Tumer [10] found the following expression for b: $b = \frac{\epsilon_i - \epsilon_j}{s}$.

Since we are considering the case of unbiased classifiers, this can be reduced to $b = \frac{\eta_i - \eta_j}{s}$. This gives

$$\sigma^2_b = \frac{\sigma^2_{\eta_i} + \sigma^2_{\eta_j}}{s^2} = \frac{2\sigma^2_{\eta_i}}{s^2} \quad (7)$$

since the noise is iid (independent and identically distributed).

For the case of b^{wa} we define $x_b = x^* + b^{wa}$ and we have

$$f^{wa}_i(x_b) = f^{wa}_j(x_b) \Leftrightarrow$$

$$p(c_i|x^*) + b^{wa}p'(c_i|x^*) + \eta^{wa}_i(x_b) =$$

$$p(c_j|x^*) + b^{wa}p'(c_j|x^*) + \eta^{wa}_j(x_b) \Leftrightarrow$$

$$b^{wa} = \frac{\eta^{wa}_i(x_b) - \eta^{wa}_j(x_b)}{s},$$

again using (3).

To find $\sigma^2_{b^{wa}}$ we start by noting that since the noise has zero mean, the second moment of b^{wa} is its variance. To find the second moment, we start by calculating

$$(b^{wa})^2 = \frac{1}{s^2}((\eta^{wa}_i)^2 + (\eta^{wa}_j)^2 - 2\eta^{wa}_i\eta^{wa}_j).$$

Since $E\eta^{wa}_i = 0$ (noise with zero mean) and the η_k are independent [1], we get

$$\sigma^2_{b^{wa}} = E(b^{wa})^2 = \frac{1}{s^2}(\sigma^2_{\eta^{wa}_i} + \sigma^2_{\eta^{wa}_j}) \quad (8)$$

Now we must rewrite $\sigma^2_{\eta^{wa}_i}$ as a function of σ^2_η, which is the same as to say, as a function of $\sigma^2_{\eta_i}$.

[1] This independence allows the expected value of the product to be the product of the expected values.

3.4. Relation between $\sigma^2_{\eta^{wa}_i}$ and $\sigma^2_{\eta_i}$

Consider (2). Since η^j_i is a noise, its expected value is zero. This implies that the expected value of η^{wa}_i is also zero. This way the variance of η^{wa}_i is equal to its second moment

$$\sigma^2_{\eta^{wa}_i} = M_2(\eta^{wa}_i).$$

To calculate the second moment, we start by finding

$$(\eta^{wa}_i)^2 = (\sum_{j=1}^{N} w_j\eta^j_i)^2 = (w_1\eta^1_i + \ldots + w_N\eta^N_i)^2$$

$$= (w_1\eta^1_i)^2 + \ldots + (w_N\eta^N_i)^2 + 2w_1\eta^1_i w_2\eta^2_i + \ldots +$$
$$2w_1\eta^1_i w_N\eta^N_i + \ldots + 2w_{N-1}\eta^{N-1}_i w_N\eta^N_i.$$

Since we are considering iid noise, the expected value of the product equals the product of the expected values, and since the expected values of the noise are zero we have

$$M_2(\eta^{wa}_i) = E(\eta^{wa}_i)^2 = w_1^2 E(\eta^1_i)^2 + \ldots + w_N^2 E(\eta^N_i)^2.$$

This leads to

$$\sigma^2_{\eta^{wa}_i} = \sum_{j=1}^{N} w_j^2 \sigma^2_{\eta^j_i}. \quad (9)$$

3.5. Relation between the added errors

Since the classifiers are independent and the noise is iid, $\sigma^2_{\eta^j_i} = \sigma^2_{\eta_i}$. The added error in the case of a single classifier is given by (6), and using (7) we find

$$E_{add} = \frac{\sigma^2_{\eta_i}}{s}. \quad (10)$$

We can now rewrite (5) using (8) and (9).

$$E^{wa}_{add} = \frac{s}{2}\sigma^2_{b^{wa}} = \frac{1}{s}\sigma^2_{\eta^{wa}_i} = \frac{1}{s}\sum_{j=1}^{N} w_j^2\sigma^2_{\eta_i}.$$

Finally, using (10) we get the relation between the added error when combining using a weighted average and the added error for the single classifier

$$E^{wa}_{add} = E_{add}\sum_{j=1}^{N} w_j^2.$$

We can see that in the case of equal weights, $\frac{1}{N}$, we get the result for the simple average (see (1)).

4. What is the best choice for the combination weights ?

Given the previous section results it is interesting to know what should the choice for the weights w_j be in order to produce the smallest added error.

This is a constraint problem: the E_{add}^{wa} is proportional to the sum of the squared weights. We should find the minimum of the sum of squared weights given that they should sum to 1. Let us define $F(w_1, ..., w_N) = \sum_{j=1}^{N} w_j^2$. The constraint is $G(w_1, \ldots, w_N) = \sum_{j=1}^{N} w_j = 1$. We build the function $u = F + \lambda G$. The λ is the Lagrange multiplier. We solve the system of $N+1$ equations $\frac{\partial u}{\partial w_j} = 0, j = 1, \ldots, N$ and $G = 1$. The minimum of F is found for $w_j = \frac{1}{N}, j = 1, ..., N$. (The method finds an extremum. We know it is the minimum by testing F on other points).

This means that the simple average is the best (in the sense that it minimizes the added error) method for doing an weighted average of several unbiased independent classifiers. For correlated classifiers, consult [5]. The derivation therein is not applicable to our case since the expressions require the inversion of singular matrices.

5. Experiments

In this section we present some experiments that confirm our findings. Since the results obtained on the previous sections considered classifiers with iid noise, we used 4 identical multi-layer perceptrons (MLPs) only differing in the weight initialization (which is done randomly). The other requisite was that the classifiers should approximate *a posteriori* probabilities, and these neural networks do so [8]. The MLPs are trained for 300 epochs using resilient backpropagation. We used three data sets from the UCI repository [3]: pima-diabetes (768 points, 8 features and 2 classes), segmentation (2310 points, 19 features and 7 classes) and sonar (208 points, 60 features and 2 classes). The topologies for the MLPs for each dataset are [8,6,2], [25,20,7] and [5,60,2], respectively. Training and test sets were generated randomly each time the experiment was repeated, each containing half of the data points. Table 1 presents the average errors for the experience.

Table 1. Average classification errors

Name	*Diabetes*	*Segmentation*	*Sonar*
C1	27.83(3.69)	22.56(5.12)	31.37(6.60)
C2	28.79(3.85)	23.17(5.39)	32.50(6.22)
C3	28.31(3.85)	22.09(5.77)	32.69(6.07)
C4	29.61(4.38)	21.48(5.32)	33.52(5.93)
S.ave.	26.33(2.72)	14.44(2.55)	28.90(4.11)
W.ave.	28.01(3.41)	21.77(4.28)	32.25(5.59)

The columns show the average and the standard deviation for the four isolated classifiers and the two combinations: simple average and weighted average for the four data sets. The experience was repeated 50 times. Each time the weights for the weighted average combination, were generated randomly. We can see that the best result is given by the simple average as derived in section 4. Note that we did not spend much time in tuning the individual classifiers, since the important issue was comparative performance. Even so, the results are compatible with recent ones [7].

6. Conclusions

We showed how the added error varies when combining unbiased and independent classifiers using a weighted average. We also showed that, for these classifiers, the weights used for a weighted average should be $\frac{1}{N}$ for a combination of N classifiers. This means that the simple average should be used. We also present experiments that confirm our results. We are currently investigating other possibilities for combining classifiers including the cases with no assumptions on their biases or on their error distributions.

References

[1] L. Alexandre, A. Campilho, and M. Kamel. Combining unbiased and independent classifiers using geometric mean. In *Proceedings of the 11th Portuguese Conference on Pattern Recognition*, pages 399–401, Porto, Portugal, May 2000.

[2] R. Battiti and A. Colla. Democracy in neural networks: Voting schemes for classification. *Neural Networks*, 7(4):691–707, 1994.

[3] E. K. C. Blake and C. Merz. UCI repository of machine learning databases, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html.

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

[5] S. Hashem. Optimal linear combinations of neural networks. *Neural Networks*, 10(4):599–614, 1997.

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

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

[8] M. Richard and R. Lippmann. Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation*, (3):461–483, 1991.

[9] K. Tumer and J. Ghosh. Boundary variance reduction for improved classification through hybrid networks. In *AeroSense 95: Applications and Science of Artificial Neural Netwoks, Proceedings of the SPIE*, volume 2492, pages 573–584, Orlando, 1995.

[10] K. Tumer and J. Gosh. Linear and order statistics combiners for pattern classification. In A. Sharkey, editor, *Combining Artificial Neural Nets*, pages 127–161. Springer-Verlag, 1999.