Combining Pairwise Coupling Classifiers Using Individual Logistic Regressions

Nobuhiko Yamaguchi

Faculty of Science and Engineering, Saga University, Saga-shi, 840–8502 Japan

Abstract. Pairwise coupling is a popular multi-class classification approach that prepares binary classifiers separating each pair of classes, and then combines the binary classifiers together. This paper proposes a pairwise coupling combination strategy using individual logistic regressions (ILR-PWC). We show analytically and experimentally that the ILR-PWC approach is more accurate than the individual logistic regressions.

1 Introduction

The object of this paper is to construct K-class classifiers. It is often easier to construct a multi-class classifier by combining multiple binary classifiers than directly construct a multi-class classifier. For example, AdaBoost [1] and support vector machines (SVM) algorithm [2] [3] are basically binary classifiers, and it is difficult to directly expand into multi-class classifiers. Typically, in such case, multi-class classifiers are constructed by decomposing the multi-class problem into multiple binary classification problems that can be handled by the AdaBoost and SVM algorithm. In addition, neural networks [4] are also binary classifiers since each output neuron separates a class from all other classes.

There are many ways to decompose a multi-class problem into multiple binary classification problems: one-per-class, individual logistic regressions [5] and pairwise coupling [6] [7]. One-per-class is one of the simplest approaches for decomposing the multi-class problem. The one-per-class approach prepares K binary classifiers, each of which separates a class from all other classes, and then constructs a multi-class classifier by combining the K binary classifiers. Next, the individual logistic regressions prepare $K-1$ binary classifiers, each of which separates a class i from an arbitrary selected baseline class j. Finally, the pairwise coupling approach prepares $K(K-1)/2$ binary classifiers, each of which separates a class i from a class j. In this paper, we focus on the pairwise coupling approach, and propose a pairwise coupling combination strategy using the individual logistic regressions. In particularly, we investigate the accuracy of our combination strategy in comparison with the individual logistic regressions.

Hastie and Tibshirani [7] show experimentally that the pairwise coupling approach is more accurate than the one-par-class approach. However the accuracy of the pairwise coupling approach has not been almost investigated theoretically. This is because that the combination strategy of the pairwise coupling approach is nonlinear and iterative. On the other hand, individual logistic regressions had

I. King et al. (Eds.): ICONIP 2006, Part II, LNCS 4233, pp. 11–20, 2006.

the same combination problem, but Begg and Gray [5] proposed a simple linear and non-iterative combination strategy with consistent property. For these reasons, we propose a pairwise coupling combination strategy using individual logistic regressions (ILR-PWC), and investigate the accuracy of our combination strategy. As a result, we show that our strategy constructs more accurate multi-class classifiers in comparison with the individual logistic regressions.

This paper is organized as follows. Section 2 explains the pairwise coupling approach proposed by Hastie and Tibshirani [7]. Section 3 explains individual logistic regressions. In section 4, we propose an extension of the pairwise coupling approach, called ILR-PWC, and compare the accuracy of the individual logistic regressions and our approach. Section 5 describes the experimental results.

2 Pairwise Coupling

2.1 Pattern Classification

In K-class classification problems, the task is to assign an input $\boldsymbol{x}_0$ to one of K classes. To solve the problems, we first estimate the posterior probability $p_i^* = P(Y_0 = i|\boldsymbol{x}_0)$ that a given input $\boldsymbol{x}_0$ belongs to a particular class i, with a training set $d = \{(\boldsymbol{x}_n, y_n) \mid 1 \leq n \leq N\}$. We then select the class with the highest posterior probability:

$$y_0 = \arg\max_{1 \leq i \leq K} p_i^* . \tag{1}$$

In the rest of this section, we consider to estimate the posterior probability p_i^* with the training set d.

2.2 Constructing Binary Classifiers

The structure of pairwise coupling is illustrated in Fig. 1. Pairwise coupling is a multi-class classification approach that prepares $K(K-1)/2$ binary classifiers r_{ij}, $1 \leq i \leq K,\ 1 \leq j < i$, and then estimates the posterior probabilities p_i^* by combining the binary classifiers together. The binary classifiers r_{ij} are trained so as to estimate pairwise class probabilities $\mu_{ij}^* = P(Y_0 = i \mid Y_0 = i \text{ or } Y_0 = j,\ \boldsymbol{x}_0)$. The estimates r_{ij} of μ_{ij}^* are available by training with the ith and jth classes of the training set:

$$d_{ij} = \{(\boldsymbol{x}_n, y_n) \mid y_n = i \text{ or } y_n = j,\ 1 \leq n \leq N\} . \tag{2}$$

Then, using all r_{ij}, the goal is to estimate $p_i^* = P(Y_0 = i|\boldsymbol{x}_0)$, $i = 1, \cdots, K$.

2.3 Estimating Posterior Probabilities

Here, we describe a method for estimating the posterior probabilities p_i^*, proposed by Hastie and Tibshirani [7]. First note that the probabilities μ_{ij}^* can be rewritten as

$$\mu_{ij}^* = P(Y_0 = i \mid Y_0 = i \text{ or } Y_0 = j,\ \boldsymbol{x}_0) = p_i^*/(p_i^* + p_j^*). \tag{3}$$

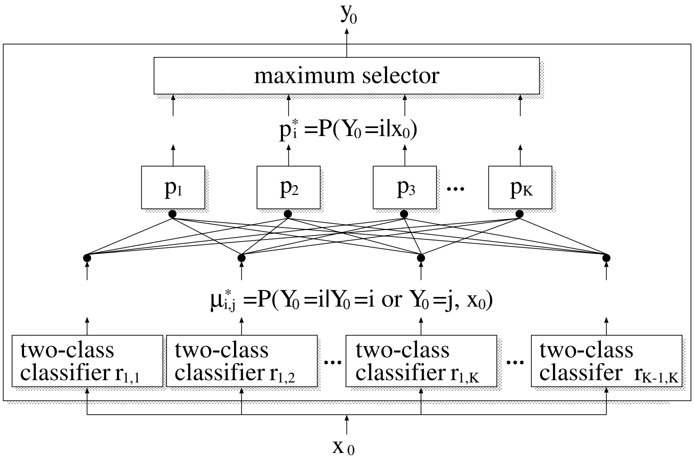
y0
maximum selector
p*i =P(Y0=i|x0)
p1
p2
p3
pK
μ*i,j =P(Y0=i|Y0=i or Y0=j, x0)
two-class classifier r1,1
two-class classifier r1,2
two-class classifier r1,K
two-class classifer rK-1,K
x0

Fig. 1. Structure of pairwise coupling

```
Step 1. Initialize $p_i$ and compute coressponding $\mu_{ij}$.
Step 2. Repeat until conversence:
(a) For each $i = 1, \cdots, K$
        $$p_i \leftarrow p_i \cdot \frac{\sum_{j \neq i}^{K} n_{ij} r_{ij}}{\sum_{j \neq i}^{K} n_{ij} \mu_{ij}}.$$
(b) Renormalize the $p_i$.
(c) Recompute the $\mu_{ij}$.
```

Fig. 2. Algorithm for estimating posterior probabilities

From (3), they consider the model as follows:

$$\mu_{ij} = p_i/(p_i + p_j), \tag{4}$$

and propose to find the estimates p_i of p_i^* so that μ_{ij} are close to the observed r_{ij}. The closeness measure is the Kullback-Leibler (KL) divergence between r_{ij} and μ_{ij}:

$$l(p_1, \cdots, p_K) = \sum_{i=1}^{K} \sum_{j=i+1}^{K} n_{ij} \left[r_{ij} \log \frac{r_{ij}}{\mu_{ij}} + (1 - r_{ij}) \log \frac{1 - r_{ij}}{1 - \mu_{ij}} \right] \tag{5}$$

where n_{ij} is the number of elements in the training set d_{ij}. Hastie and Tibshirani [7] propose to find the estimates p_i that minimize the function l, and also propose to use an iterative algorithm to compute the p_i's as illustrated in Fig. 2.

3 Individual Logistic Regressions

The object of this paper is to propose a pairwise coupling combination strategy using individual logistic regressions [5]. The individual logistic regressions are

K-class classification approaches that combine $K-1$ binary classifiers. In this section, we describe the individual logistic regressions, and in the next section, we propose a pairwise coupling combination strategy using the individual logistic regressions (ILR-PWC).

3.1 Background

Multinomial logistic regressions [8] are popular approaches for solving multi-class classification problems. However, at the time when the individual logistic regressions were proposed, most statistical software packages included only simple binary logistic regressions, but did not include the multinomial logistic regressions. For this reason, Begg and Gray [5] proposed the individual logistic regressions which approximate the multinomial logistic regressions by combining multiple binary logistic regressions. They show that the approximation algorithm is not maximum likelihood but is consistent [5]D In addition, some experiments [5] [9] show that the efficiency loss of the approximation is small. For these reasons, the individual logistic regressions are still used to approximate the multinomial logistic regressions.

The rest of this section is organized as follows. Section 3.2 and 3.3 describe the logistic regressions and multinomial logistic regressions, respectively. In section 3.4, we describe a method for approximating the multinomial logistic regressions by using the individual logistic regressions.

3.2 Logistic Regressions

Logistic regressions are one of the most widely used techniques for solving binary classification problems. In the logistic regressions, the posterior probabilities p_i^*, $i \in \{1, 2\}$, are represented as the following:

$$\pi_1 = \frac{\exp(\eta)}{1 + \exp(\eta)}, \quad \pi_2 = 1 - \pi_1 \tag{6}$$

where η is a function of an input $\boldsymbol{x}_0$. For example, η is a linear function of the input $\boldsymbol{x}_0$, that is,

$$\eta = \boldsymbol{\alpha}^T \boldsymbol{x}_0 + \beta, \tag{7}$$

and the parameters $\boldsymbol{\alpha}$, β are estimated by the maximum likelihood method. In this paper, η is an arbitrary function of $\boldsymbol{x}_0$. Note that if you choose an appropriate η, the model in (6) can represent some kinds of binary classification systems, such as neural networks, logitBoost [10], etc.

3.3 Multinomial Logistic Regressions

Multinomial logistic regressions are one of the techniques for solving multi-class classification problems. In the multinomial logistic regressions, the posterior probabilities p_i^*, $i \in \{1, \cdots, K\}$, are represented as the following:

$$\pi_i^j = \begin{cases} \dfrac{\exp(\eta_i^j)}{1+\sum_{k\neq j}^K \exp(\eta_k^j)} & \text{if} \quad i \neq j \\ \\ \dfrac{1}{1+\sum_{k\neq j}^K \exp(\eta_k^j)} & \text{otherwise} \end{cases} \tag{8}$$

where j is a baseline class and η_i^j is a function of an input $\boldsymbol{x}_0$. For example, η_i^j is a linear function of the input $\boldsymbol{x}_0$, that is,

$$\eta_i^j = {\boldsymbol{\alpha}_i^j}^T \boldsymbol{x}_0 + \beta_i^j, \tag{9}$$

and the parameters $\boldsymbol{\alpha}_i^j$, β_i^j are estimated by the maximum likelihood method. As in the case of the logistic regressions, η_i^j is an arbitrary function of $\boldsymbol{x}_0$, and the baseline class j is an arbitrary class.

3.4 Individual Logistic Regressions

Individual logistic regressions are techniques for approximating K-class multinomial logistic regressions by combining $K-1$ binary logistic regressions. As in the case of the multinomial logistic regressions, the individual logistic regressions represent the posterior probabilities p_i^* as (8), but the function η_i^j is approximated by using $K-1$ binary logistic regressions. In the following sentence, we describe the method for approximating the function η_i^j.

First, we select a class j and prepare $K-1$ binary logistic regressions π_{ij}, $i = 1, \cdots, j-1,\ j+1, \cdots, K$. The binary logistic regressions π_{ij} are trained so as to estimate the probabilities $\mu_{ij}^* = P(Y_0 = i \mid Y_0 = i \text{ or } Y_0 = j,\ \boldsymbol{x}_0)$. Namely, we prepare $K-1$ logistic regressions

$$\pi_{ij} = \frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})} \tag{10}$$

and train the π_{ij}'s with the training set d_{ij} in (2).

The function η_{ij} in (10) can be considered as an estimate of $\log p_i^*/p_j^*$ by the following expansion:

$$\eta_{ij} = \log \frac{\pi_{ij}}{1-\pi_{ij}} \approx \log \frac{\mu_{ij}^*}{1-\mu_{ij}^*} = \log \frac{p_i^*}{p_j^*}, \tag{11}$$

and the function η_i^j in (8) can be also considerd as an estimate of $\log p_i^*/p_j^*$ by the following expansion:

$$\eta_i^j = \log \frac{\pi_i^j}{\pi_j^j} \approx \log \frac{p_i^*}{p_j^*}. \tag{12}$$

From this equality, replacing the function η_i^j in (8) with the function η_{ij} in (10), we can approximate the multinomial logistic regression of the baseline class j as follows:

$$\pi_i^j = \begin{cases} \dfrac{\exp(\eta_{ij})}{1+\sum_{k\neq j}^{K}\exp(\eta_{kj})} & \text{if} \quad i \neq j \\ \dfrac{1}{1+\sum_{k\neq j}^{K}\exp(\eta_{kj})} & \text{otherwise.} \end{cases} \tag{13}$$

4 ILR-PWC

4.1 Pattern Classification Problem of ILR-PWC

In this paper, we propose a pairwise coupling combination strategy using individual logistic regressions (ILR-PWC). As in the case of the pairwise coupling approach, the ILR-PWC approach prepares $K(K-1)/2$ binary classifiers r_{ij}, and then combines the binary classifiers together. In the ILR-PWC approach, however, logistic regression is used as the binary classifier r_{ij}, that is,

$$r_{ij} = \frac{\exp(g_{ij})}{1+\exp(g_{ij})} \tag{14}$$

where g_{ij} is an arbitrary function of an input $\boldsymbol{x}_0$. The logistic regression r_{ij} is trained so as to estimate probability μ_{ij}^* with the training set d_{ij}. Then, the goal is to estimate the posterior probabilities p_i^* by using all r_{ij}.

To estimate the posterior probabilities, Hastie and Tibshirani [7] proposed a nonlinear and iterative algorithm, but it is difficult to investigate the accuracy. From this reason, we propose a two-stage estimation strategy. In the first stage, we construct K multinomial logistic regressions using the $K(K-1)/2$ logistic regressions r_{ij}. In the second stage, we estimate the posterior probabilities p_i^* using the K multinomial logistic regressions. In this paper, we show that the optimal estimates of p_i^* can be derived as a linear combination of the K multinomial logistic regressions, and we investigate the accuracy of our estimation strategy in comparison with individual logistic regressions.

The rest of this section is organized as follows. In section 4.2, we propose a method for constructing K multinomial logistic regressions by using individual logistic regressions. In section 4.3, we estimate the posterior probabilities p_i^* using the K multinomial logistic regressions. In section 4.4, we investigate the accuracy of the ILR-PWC approach.

4.2 Constructing Multinomial Logistic Regressions

In this section, we propose a method for constructing K multinomial logistic regressions using the $K(K-1)/2$ logistic regressions r_{ij}. First, note that r_{ij} and π_{ij} in (10) are the same estimate because they are trained so as to estimate the same probability μ_{ij}^* with the same training set d_{ij}. We can therefore approximate multinomial logistic regressions with individual logistic regressions in section 3.4. That is, we can approximate a multinomial logistic regression of a baseline class j as the following:

$$p_i^j = \begin{cases} \dfrac{\exp(g_{ij})}{1 + \sum_{k \neq j}^{K} \exp(g_{kj})} & \text{if} \quad i \neq j \\ \\ \dfrac{1}{1 + \sum_{k \neq j}^{K} \exp(g_{kj})} & \text{otherwise.} \end{cases} \tag{15}$$

The ILR-PWC approach prepares K multinomial logistic regressions p_i^j of the baseline class $j = 1, \cdots, K$ using (15).

4.3 Estimating Posterior Probabilities

In this section, we consider to estimate the posterior probabilities p_i^* using the K multinomial logistic regressions p_i^j. In the ILR-PWC approach, we find the estimate p_i of p_i^* so that p_i is close to the estimates $p_i^1, \cdots, p_i^K$ of the K multinomial logistic regressions. The closeness measure is the Kullback-Leibler (KL) divergence between p_i and $p_i^1, \cdots, p_i^K$. Noting further that the sum of probabilities is 1, we can write the problem of estimating p_i^* as follows:

$$\text{minimize} \quad \sum_{i=1}^{K} \sum_{j=1}^{K} p_i^j \log \frac{p_i^j}{p_i} \quad \text{subject to} \quad \sum_{i=1}^{K} p_i = 1. \tag{16}$$

In the rest of this subsection, we solve this constrained optimization problem.

We use the Lagrange multiplier method to derive the optimal estimate p_i. We first define an objective function L as follows:

$$L(p_1, \cdots, p_K, \lambda) = \sum_{i=1}^{K} \sum_{j=1}^{K} p_i^j \log \frac{p_i^j}{p_i} - \lambda \left\{ \sum_{i=1}^{K} p_i - 1 \right\}. \tag{17}$$

Differentiating the function L with respect to the p_i and Lagrange multiplier λ, we can obtain

$$\sum_{i=1}^{K} p_i = 1, \tag{18}$$

$$p_i = -\frac{1}{\lambda} \sum_{j=1}^{K} p_i^j. \tag{19}$$

Substituting (19) into (18), we obtain $\lambda = -K$. Further substituting $\lambda = -K$ into (19), we can derive the optimal estimate p_i as follows:

$$p_i = \frac{1}{K} \sum_{j=1}^{K} p_i^j. \tag{20}$$

Thus, we construct a multi-class classifier by using (1), (14), (15) and (20), and we call this strategy ILR-PWC (pairwise coupling combination strategy using individual logistic regressions).

4.4 Investigation of Accuracy of ILR-PWC

In this section, we compare the accuracy of the ILR-PWC approach with individual logistic regressions. Here, we use the estimation error of posterior probabilities to evaluate the accuracy of a multi-class classifier. First, we define the accuracy of the ILR-PWC approach as (21). In the same way, we define the accuracy of the individual logistic regressions as (22), but R_i^{ilr} is defined using the mean of all baseline classes since we can select an arbitrary class as the baseline class.

$$R_i^{ilr-pwc} = \mathrm{E}\left\{(p_i^* - p_i)^2\right\} \tag{21}$$

$$R_i^{ilr} = \frac{1}{K}\sum_{j=1}^{K}\mathrm{E}\left\{(p_i^* - p_i^j)^2\right\} \tag{22}$$

We can obtain (23) by transforming (21) into (24), and we can therefore show that the ILR-PWC approach is more accurate than the individual logistic regressions.

$$R_i^{ilr-pwc} \leq R_i^{ilr} \tag{23}$$

$$\begin{aligned}\mathrm{E}\left\{(p_i^* - p_i)^2\right\} &= \mathrm{E}\left\{(p_i^* - \frac{1}{K}\sum_{j=1}^{K}p_i^j)^2\right\}\\ &= \mathrm{E}\left\{\frac{1}{K^2}(\sum_{j=1}^{K}(p_i^* - p_i^j))^2\right\}\\ &\leq \mathrm{E}\left\{\frac{1}{K}\sum_{j=1}^{K}(p_i^* - p_i^j)^2\right\}\end{aligned} \tag{24}$$

where the last inequality is obtained by the Cauchy-Schwarz inequality.

5 Computer Simulation

We present an experimental evaluation on 7 data sets from the UCI machine learning repository [11], including glass, hayes-roth, iris, led, letter, segment and vehicle. A summary of data sets is given in Table 1. For comparison, we tested three different approaches; one-per-class (OPC), pairwise coupling (PWC) and individual logistic regressions (ILR). In our experiment, as individual binary classifiers r_{ij}, we employ feedforward neural networks with one output unit and 10 hidden units.

To evaluate our approach, we used the evaluation technique 10-fold cross-validation method, which consists of randomly dividing the data into 10 equal-sized groups and performing ten different experiments. In each run, nine of the ten groups are used to train the classifiers and the remaining group is held out for

Table 1. Experimental data set

Data Set	Entries	Attributes	Classes
glass	214	9	6
hayes-roth	132	5	3
iris	150	4	3
led	700	7	10
letter	20000	16	26
segment	2310	19	7
vehicle	846	18	4

Table 2. Average misclassification rates

dataset	OPC	PWC	ILR	ILR-PWC
glass	39.7	34.1	37.6	35.0
hayes-roth	38.0	30.4	35.6	30.4
iris	4.7	4.0	6.5	4.0
led	27.9	27.9	31.4	27.3
letter	39.8	18.6	33.3	17.5
segment	8.6	6.6	12.7	6.9
vehicle	23.2	21.2	25.3	21.3
average	26.0	20.4	26.1	20.3

the evaluation. Table 2 shows the average misclassification rates of 10 runs of 10-fold cross-validations. From Table 2, we can see that the misclassification rate of the ILR-PWC approach is better than that of the ILR approach. From Table 2, we can see that the maximal difference of misclassification rates between the PWC and ILR-PWC approach is 1.1% in letter data and the performance of the PWC and ILR-PWC approach are almost the same.

6 Conclusion

In this paper, we have focused on combining binary classifiers of pairwise coupling and have proposed a pairwise coupling combination strategy using individual logistic regressions (ILR-PWC). In particular, we have investigated the accuracy of the ILR-PWC approach, and as a result, we have shown that our combination strategy is more accurate than individual logistic regressions.

References

1. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences **55** (1997) 119–139
2. Cortes, C., Vapnik, V.: Support vector networks. Machine Learning **20** (1995) 273–297

3. Vapnik, V.: The nature of statistical learning theory, Springer (1995)
4. Rumelhart, D., Hinton, G., Williams, R.: Learning internal representations by error propagation. In: Rumelhart, D., McClelland, J. et al. (eds.): Parallel Distributed Processing: Volume 1: Foundations, MIT Press, Cambridge (1987) 318–362
5. Begg, C., Gray, R.: Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika **71** (1984) 11–18
6. Friedman, J.: Another approach to polychotomous classification. Technical Report, Statistics Department, Stanford University (1996)
7. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. The Annals of Statistics **26** (1998) 451–471
8. Agresti, A.: Categorical Data Analysis. John Wiley & Sons (1990).
9. Hosmer, D., Lemeshow, S.: Applied logistic regression, 2nd ed. Wiley-Interscience (2000)
10. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting. Annals of statistics **28** (2000) 337–374
11. Blake, C., Merz, C.: UCI repository of machine learning databases (1998)