Multi-Category Classification by Soft-Max Combination of Binary Classifiers

K. Duan, S. S. Keerthi, W. Chu, S. K. Shevade, A. N. Poo
Department of Mechanical Engineering, National University of Singapore
Gatsby Computational Neuroscience Unit, University College London
Department of Computer Science and Automation, Indian Institute of Science

2003/06/12

– The fourth International Workshop on Multiple Classifier Systems –

Multi-category classification Using Binary Classifiers

- **All-Together Methods**: Training speed is usually slow.
- **One-Versus-All Methods**: For M-category classification, construct M one-versus-all binary classifiers, each to distinguish one class from all other classes.

  Implementation strategy: Winner-Takes-All
- **One-Versus-One Methods:** For M-category classification, construct $M(M-1)/2$ binary classifiers, each to distinguish one class from another.

  Implementation strategy: Max-Wins-Voting etc.

Multi-category classification Using Binary Classifiers

- **Pairwise Coupling**: For one-versus-one binary classifiers with probabilistic outputs, such as kernel logistic regression.

  Central idea: Couple $M(M-1)/2$ pairwise class probability estimates to obtain estimates of posterior probabilities for M classes.

- **Our Methods:** Combine one-versus-others or one-versus-one binary classifiers through soft-max functions to obtain posterior class probabilities.

Soft-Max Combination of Binary Classifiers

M classes and l labelled training data $(\mathbf{x}_1, y_1)$, $\cdots$, $(\mathbf{x}_l, y_l)$, where $\mathbf{x}_i \in \mathbf{R}^m$ and $y_i \in \{1, \cdots, M\}$.

- Combination of **One-Versus-All** binary classifiers;
- Combination of **One-Versus-One** binary classifiers;
- Relation to Previous Work.

Soft-Max Combination of One-Versus-All Classifiers

Denote the output of the kth binary classifier (class c_k versus the rest) for $\mathbf{x}_i$ as r_k^i.

Posteriori probabilities obtained through a soft-max function

$$P_k^i = \mathrm{Prob}(c_k|\mathbf{x}_i) = \frac{e^{w_k r_k^i + w_{ko}}}{z^i} \tag{1}$$

where $z^i = \sum_{k=1}^{M} e^{w_k r_k^i + w_{ko}}$ is a normalization term to ensure $\sum_{k=1}^{M} P_k^i = 1$.

Soft-Max Combination of One-Versus-All Classifiers

$\mathbf{w} = \{(w_1, w_{1o}), \ldots, (w_M, w_{Mo})\}$ can be designed to minimize a penalized NLL:

$$\begin{aligned} \min \qquad & E(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 - C\sum_{i=1}^{l} \log P_{y_i}^i \qquad (2) \\ \text{subject to} \qquad & w_k, w_{ko} > 0 \ , \ k = 1, \cdots, M \end{aligned}$$

Auxiliary Variables:

$$s_k = \log(w_k) \ , \ s_{ko} = \log(w_{ko})$$

Soft-Max Combination of One-Versus-All Classifiers

The optimization problem can be solved using gradient methods. Following formulas give gradients wrt auxiliary variables:

$$\frac{\partial E}{\partial s_k} = \frac{\partial E}{\partial w_k}\frac{\partial w_k}{\partial s_k} = \left(w_k + C \sum_{y_i = k} \left(P_k^i - 1\right) r_k^i + C \sum_{y_i \neq k} P_k^i r_k^i \right) w_k$$

$$\frac{\partial E}{\partial s_{ko}} = \frac{\partial E}{\partial w_{ko}}\frac{\partial w_{ko}}{\partial d s_{ko}} = \left(w_{ko} + C \sum_{y_i = k} \left(P_k^i - 1\right) + C \sum_{y_i \neq k} P_k^i \right) w_{ko}$$

Soft-Max Combination of One-Versus-One Classifiers

Denote the output of one-versus-one classifier C_{kt} for $\mathbf{x}_i$ as r^i_{kt}. We have $r^i_{tk} = -r^i_{kt}$.

The posteriori probabilities can be obtained through a soft-max function

$$P^i_k = \mathrm{Prob}(c_k|\mathbf{x}_i) = \frac{e^{\sum_{t \neq k} w_{kt} r^i_{kt} + w_{ko}}}{z^i} \tag{3}$$

where $z^i = \sum_{k=1}^{M} e^{\sum_{t \neq k} w_{kt} r^i_{kt} + w_{ko}}$ is a normalization term.

Soft-Max Combination of One-Versus-One Classifiers

The weight parameters $\mathbf{w}$ can be designed to minimize a penalized NLL:

$$\begin{aligned} \min \qquad & E(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 - C\sum_{i=1}^{l} \log P_{y_i}^{i} \qquad (4) \\ \text{subject to} \qquad & w_{kt}, w_{ko} > 0\ ,\ k, t = 1, \cdots, M \text{ and } t \neq k \end{aligned}$$

Auxiliary Variables:

$$s_{kt} = \log(w_{kt})\ ,\ s_{ko} = \log(w_{ko})$$

Soft-Max Combination of One-Versus-One Classifiers

The optimization problem can be solved using gradient methods. Following formulas give gradients wrt auxiliary variables:

$$\frac{\partial E}{\partial s_{kt}} = \frac{\partial E}{\partial w_{kt}} \frac{\partial w_{kt}}{\partial s_{kt}} = \left(w_{kt} + C \sum_{y_i = k} \left(P_k^i - 1 \right) r_{kt}^i + C \sum_{y_i \neq k} P_k^i r_{kt}^i \right) w_{kt}$$

$$\frac{\partial E}{\partial s_{ko}} = \frac{\partial E}{\partial w_{ko}} \frac{\partial w_{ko}}{\partial s_{ko}} = \left(w_{ko} + C \sum_{y_i = k} \left(P_k^i - 1 \right) + C \sum_{y_i \neq k} P_k^i \right) w_{ko}$$

Relation to Previous Work

- The following parametric model is used by Platt (1999) to fit the posteriori probability

$$\mathrm{Prob}(c_1|\mathbf{x}_i) = \frac{1}{1+e^{Af_i+B}}\ , \tag{5}$$

where f_i is the output of SVMs.

- One-Versus-All case with $M = 2$, $r_1^i = f_i$, and $r_2^i = -r_1^i$:

$$\mathrm{Prob}(c_1|\mathbf{x}_i) = \frac{1}{1+e^{-(w_1+w_2)f_i+(w_{2o}-w_{1o})}}\ . \tag{6}$$

Relation to Previous Work

- One-Versus-One case with $M = 2$, $r^i_{12} = f_i$ and $r^i_{21} = -r^i_{12}$:

$$\mathrm{Prob}(c_1|\mathbf{x}_i) = \frac{1}{1 + e^{-(w_{12}+w_{21})f_i + (w_{2o} - w_{1o})}} \,. \tag{7}$$

- Therefore, our soft-max combination methods can be viewed as natural extensions of Platt's sigmoid-fitting idea to multi-category classification.

Practical Issues in Soft-Max Design

- 5-fold cross validation for soft-max design

  The original training data is partitioned into 5 folds with each fold containing equal percentage of examples of one particular class.

- Regularization Parameter C

  We select optimal C by the validation estimates of error rate and negative log-likelihood.

- Simplified soft-max function design

  We may omit the use of regularization.

Numerical Study

- Soft-max combination of SVM one-versus-all classifiers: standard design and simplified soft-max function design
- Soft-max combination of SVM one-versus-one classifiers: standard design and simplified soft-max function design
- Winner-Takes-All of one-versus-all classifiers: SVM, SVM with Platt's posterior probabilities (PSVM) and kernel logistic regression (KLOGR)
- Max-Wins-Voting of one-versus-one classifiers: SVM, PSVM and KLOGR
- Pairwise coupling of one-versus-one classifiers: PSVM and KLOGR.

Results and Conclusions

- Winner-Takes-All of KLOGR seems best among all one-versus-all classifiers.
- Max-Wins-Voting of KLOGR seems best among all one-versus-one classifiers.
- Overall, Pairwise-Coupling of KLOGR seems slightly better.
- The proposed soft-max combination methods with simplified combination function design are competitive and simpler to design.
- They provide new ways of obtaining posteriori probability estimates from binary classifiers whose outputs are not probabilistic values.