Journal of Machine Learning Research 7 (2006) 85–115 Submitted 5/05; Revised 10/05; Published 1/06

Generalized Bradley-Terry Models and Multi-class Probability Estimates

Tzu-Kuo Huang R93002@CSIE.NTU.EDU.TW
Department of Computer Science, National Taiwan University
Taipei 106, Taiwan

Ruby C. Weng CHWENG@NCCU.EDU.TW
Department of Statistics, National Chengchi University
Taipei 116, Taiwan

Chih-Jen Lin CJLIN@CSIE.NTU.EDU.TW
Department of Computer Science, National Taiwan University
Taipei 106, Taiwan

Editor: Greg Ridgeway

Abstract

The Bradley-Terry model for obtaining individual skill from paired comparisons has been popular in many areas. In machine learning, this model is related to multi-class probability estimates by coupling all pairwise classification results. Error correcting output codes (ECOC) are a general framework to decompose a multi-class problem to several binary problems. To obtain probability estimates under this framework, this paper introduces a generalized Bradley-Terry model in which paired individual comparisons are extended to paired team comparisons. We propose a simple algorithm with convergence proofs to solve the model and obtain individual skill. Experiments on synthetic and real data demonstrate that the algorithm is useful for obtaining multi-class probability estimates. Moreover, we discuss four extensions of the proposed model: 1) weighted individual skill, 2) home-field advantage, 3) ties, and 4) comparisons with more than two teams.

Keywords: Bradley-Terry model, Probability estimates, Error correcting output codes, Support Vector Machines

1. Introduction

The Bradley-Terry model (Bradley and Terry, 1952) for paired comparisons has been broadly applied in many areas such as statistics, sports, and machine learning. It considers a set of k individuals for which

$$P(\text{individual } i \text{ beats individual } j) = \frac{p_i}{p_i + p_j}, \tag{1}$$

and $p_i > 0$ is the overall skill of individual i. Suppose that the outcomes of all comparisons are independent and denote r_{ij} as the number of times that i beats j. Then the negative log-likelihood takes the form

$$l(\mathbf{p}) = -\sum_{i<j} \left(r_{ij} \log \frac{p_i}{p_i + p_j} + r_{ji} \log \frac{p_j}{p_i + p_j} \right). \tag{2}$$

Since $l(\mathbf{p}) = l(\alpha\mathbf{p})$ for any $\alpha > 0$, $l(\mathbf{p})$ is scale invariant. Therefore, it is convenient to assume that $\sum_{i=1}^{k} p_i = 1$ for the sake of identifiability. One can then estimate p_i by

$$\begin{aligned} \min_{\mathbf{p}} \quad & l(\mathbf{p}) \\ \text{subject to} \quad & 0 \le p_j, j = 1, \ldots, k, \sum_{j=1}^{k} p_j = 1. \end{aligned} \tag{3}$$

This approach dates back to (Zermelo, 1929) and has been extended to more general settings. For instance, in sports scenario, extensions to account for the home-field advantage and ties have been proposed. Some reviews are, for example, (David, 1988; Davidson and Farquhar, 1976; Hunter, 2004; Simons and Yao, 1999). The solution of (3) can be solved by a simple iterative procedure:

Algorithm 1

1. Start with any initial $p_j^0 > 0$, $j = 1, \ldots, k$.
2. Repeat $(t = 0, 1, \ldots)$
   (a) Let $s = (t \mod k) + 1$. Define

$$\mathbf{p}^{t+1} \equiv \left[p_1^t, \ldots, p_{s-1}^t, \frac{\sum_{i:i\neq s} r_{si}}{\sum_{i:i\neq s} \frac{r_{si}+r_{is}}{p_s^t+p_i^t}}, p_{s+1}^t, \ldots, p_k^t \right]^T. \tag{4}$$

   (b) Normalize $\mathbf{p}^{t+1}$.

   until $\partial l(\mathbf{p}^t)/\partial p_j = 0, j = 1, \ldots, k$ are satisfied.

This algorithm is so simple that there is no need to use sophisticated optimization techniques. If r_{ij} $\forall i, j$ satisfy some mild conditions, Algorithm 1 globally converges to the unique minimum of (3). A systematic study on the convergence of Algorithm 1 is in (Hunter, 2004).

An earlier work (Hastie and Tibshirani, 1998) in statistics and machine learning considered the problem of obtaining multi-class probability estimates by coupling results from pairwise comparisons. Assume

$$\bar{r}_{ij} \equiv P(\mathbf{x} \text{ in class } i \mid \mathbf{x} \text{ in class } i \text{ or } j)$$

is known. This work estimates $p_i = P(\mathbf{x}$ in class $i)$ by minimizing the (weighted) Kullback-Leibler (KL) distance between $\bar{r}_{ij}$ and $\mu_{ij} \equiv p_i/(p_i + p_j)$:

$$\begin{aligned} \min_{\mathbf{p}} \quad & \sum_{i<j} n_{ij} \left(\bar{r}_{ij} \log \frac{\bar{r}_{ij}}{\mu_{ij}} + \bar{r}_{ji} \log \frac{\bar{r}_{ji}}{\mu_{ji}} \right) \\ \text{subject to} \quad & 0 \le p_j, j = 1, \ldots, k, \sum_{j=1}^{k} p_j = 1, \end{aligned} \tag{5}$$

where n_{ij} is the number of training data in class i or j. By defining $r_{ij} \equiv n_{ij}\bar{r}_{ij}$ and removing constant terms, (5) reduces to the same form as (2), and hence Algorithm 1 can be used to find $\mathbf{p}$. Although one might interpret this as a Bradley-Terry model by treating classes as individuals and r_{ij} as the number that the ith class beats the jth class, it is not indeed. First, r_{ij} (now defined as $n_{ij}\bar{r}_{ij}$) may not be an integer any more. Secondly, r_{ij} are dependent as they share the same training set. However, the closeness between the two motivates us to propose more general models in this paper.

The above approach involving comparisons for each pair of classes is referred to as the "one-against-one" setting in multi-class classification. It is a special case of the framework *error correcting output codes* (ECOC) to decompose a multi-class problem into a number of binary problems (Dietterich and Bakiri, 1995; Allwein et al., 2001). Some classification techniques are two-class based, so this framework extends them to multi-class scenarios. Zadrozny (2002) generalizes the results in (Hastie and Tibshirani, 1998) to obtain probability estimates under ECOC settings. The author proposed an algorithm analogous to Algorithm 1 and demonstrated some experimental results. However, the convergence issue was not discussed. Though the author intended to minimize the KL distance as Hastie and Tibshirani (1998) did, in Section 4.2 we show that their algorithm may not converge to a point with the smallest KL distance.

Motivated from multi-class classification with ECOC settings, this paper presents a generalized Bradley-Terry model where each competition is between two teams (two disjoint subsets of subjects) and team size/members can vary from competition to competition. Then from the outcomes of all comparisons, we fit this general model to estimate the individual skill. Here we propose a simple iterative method to solve the generalized model. The convergence is proved under mild conditions.

The proposed model has some potential applications. For example, in tennis or badminton, if a player participates in many singles and doubles, this general model can combine all outcomes to yield the estimated skill of all individuals. More importantly, for multi-class problems by combining binary classification results, we can also minimize the KL distance and obtain the same optimization problem. Hence the proposed iterative method can be directly applied to obtain the probability estimate under ECOC settings.

This paper is organized as follows. Section 2 introduces a generalized Bradley-Terry model and a simple algorithm to maximize the log-likelihood. The convergence of the proposed algorithm is in Section 3. Section 4 discusses multi-class probability estimates and experiments are in Sections 5 and 6. In Section 7 we discuss four extensions of the proposed model: 1) weighted individual skill, 2) home-field advantage, 3) ties, and 4) comparisons with more than two teams. Discussion and conclusions are in Section 8. A short and preliminary version of this paper appeared in an earlier conference NIPS 2004 (Huang et al., 2005) [1].

2. Generalized Bradley-Terry Model

In this section we study a generalized Bradley-Terry model for approximating individual skill. Consider a group of k individuals: $\{1, \ldots, k\}$. Each time two disjoint subsets I_i^+ and I_i^- form teams for a series of games and $r_i \geq 0$ ($r_i' \geq 0$) is the number of times that I_i^+

1. Programs used are at `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/libsvm-errorcode`.

beats I_i^- (I_i^- beats I_i^+). Thus, we have $I_i \subset \{1,\ldots,k\}, i = 1,\ldots,m$ so that

$$I_i = I_i^+ \cup I_i^-, \qquad I_i^+ \neq \emptyset, I_i^- \neq \emptyset, \text{ and } I_i^+ \cap I_i^- = \emptyset.$$

If the game is designed so that each member is equally important, we can assume that a team's skill is the sum of all its members'. This leads to the following model:

$$P(I_i^+ \text{ beats } I_i^-) = \frac{\sum_{j\in I_i^+} p_j}{\sum_{j\in I_i} p_j}.$$

If the outcomes of all comparisons are independent, then estimated individual skill can be obtained by defining

$$q_i \equiv \sum_{j\in I_i} p_j, \qquad q_i^+ \equiv \sum_{j\in I_i^+} p_j, \qquad q_i^- \equiv \sum_{j\in I_i^-} p_j$$

and minimizing the negative log-likelihood

$$\begin{aligned} \min_{\mathbf{p}} \quad & l(\mathbf{p}) = -\sum_{i=1}^{m} \left(r_i \log(q_i^+/q_i) + r_i' \log(q_i^-/q_i)\right) \\ \text{subject to} \quad & \sum_{j=1}^{k} p_j = 1, 0 \le p_j, j = 1,\ldots,k. \end{aligned} \tag{6}$$

Note that (6) reduces to (3) in the pairwise approach, where $m = k(k-1)/2$ and $I_i, i = 1,\ldots,m$ are as the following:

I_i^+	I_i^-	r_i	r_i'
$\{1\}$	$\{2\}$	r_{12}	r_{21}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\{1\}$	$\{k\}$	r_{1k}	r_{k1}
$\{2\}$	$\{3\}$	r_{23}	r_{32}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\{k-1\}$	$\{k\}$	$r_{k-1,k}$	$r_{k,k-1}$

In the rest of this section we discuss how to solve the optimization problem (6).

2.1 A Simple Procedure to Maximize the Likelihood

The difficulty of solving (6) over (3) is that now $l(\mathbf{p})$ is expressed in terms of q_i^+, q_i^-, q_i but the real variable is $\mathbf{p}$. We propose the following algorithm to solve (6).

Algorithm 2

1. Start with initial $p_j^0 > 0, j = 1, \ldots, k$ and obtain corresponding $q_i^{0,+}, q_i^{0,-}, q_i^0, i = 1, \ldots, m$.
2. Repeat $(t = 0, 1, \ldots)$
   (a) Let $s = (t \mod k) + 1$. Define $\mathbf{p}^{t+1}$ by $p_j^{t+1} = p_j^t,\ \forall j \neq s$, and

$$p_s^{t+1} = \frac{\sum_{i:s\in I_i^+} \frac{r_i}{q_i^{t,+}} + \sum_{i:s\in I_i^-} \frac{r_i'}{q_i^{t,-}}}{\sum_{i:s\in I_i} \frac{r_i+r_i'}{q_i^t}} p_s^t. \tag{7}$$

   (b) Normalize $\mathbf{p}^{t+1}$.
   (c) Update $q_i^{t,+}, q_i^{t,-}, q_i^t$ to $q_i^{t+1,+}, q_i^{t+1,-}, q_i^{t+1}, i = 1, \ldots, m$.

   until $\partial l(\mathbf{p}^t)/\partial p_j = 0, j = 1, \ldots, k$ are satisfied.

The gradient of $l(\mathbf{p})$, used in the stopping criterion, is:

$$\begin{aligned}\frac{\partial l(\mathbf{p})}{\partial p_s} &= -\sum_{i=1}^m \left(r_i \frac{\partial \log q_i^+}{\partial p_s} + r_i' \frac{\partial \log q_i^-}{\partial p_s} - (r_i + r_i') \frac{\partial \log q_i}{\partial p_s} \right) \\ &= -\sum_{i:s\in I_i^+} \frac{r_i}{q_i^+} - \sum_{i:s\in I_i^-} \frac{r_i'}{q_i^-} + \sum_{i:s\in I_i} \frac{r_i + r_i'}{q_i}, \qquad s = 1, \ldots, k. \end{aligned} \tag{8}$$

In Algorithm 2, for the multiplicative factor in (7) to be well defined (i.e., non-zero denominator), we need Assumption 1, which will be discussed in Section 3. Eq. (7) is a simple fixed-point type update; in each iteration, only one component (i.e., p_s^t) is modified while the others remain the same. If we apply the updating rule (7) to the pairwise model,

$$p_s^{t+1} = \frac{\sum_{i:s<i} \frac{r_{si}}{p_s^t} + \sum_{i:i<s} \frac{r_{si}}{p_s^t}}{\sum_{i:s<i} \frac{r_{si}+r_{is}}{p_s^t+p_i^t} + \sum_{i:i<s} \frac{r_{is}+r_{si}}{p_s^t+p_i^t}} p_s^t = \frac{\sum_{i:i\neq s} r_{si}}{\sum_{i:i\neq s} \frac{r_{si}+r_{is}}{p_s^t+p_i^t}}$$

reduces to (4).

The updating rule (7) is motivated from using a descent direction to strictly decrease $l(\mathbf{p})$: If $\partial l(\mathbf{p}^t)/\partial p_s \neq 0$ and $p_s^t > 0$, then under suitable assumptions on r_i, r_i',

$$\begin{aligned}\frac{\partial l(\mathbf{p}^t)}{\partial p_s}(p_s^{t+1} - p_s^t) &= \frac{\partial l(\mathbf{p}^t)}{\partial p_s} \left(\frac{\sum_{i:s\in I_i^+} \frac{r_i}{q_i^+} + \sum_{i:s\in I_i^-} \frac{r_i'}{q_i^-} - \sum_{i:s\in I_i} \frac{r_i+r_i'}{q_i}}{\sum_{i:s\in I_i} \frac{r_i+r_i'}{q_i^t}} \right) p_s^t \\ &= \left(-\left(\frac{\partial l(\mathbf{p}^t)}{\partial p_s} \right)^2 p_s^t \right) \Big/ \left(\sum_{i:s\in I_i} \frac{r_i + r_i'}{q_i^t} \right) < 0. \end{aligned} \tag{9}$$

Thus, $p_s^{t+1} - p_s^t$ is a descent direction in optimization terminology since a sufficiently small step along this direction guarantees the strict decrease of the function value. As now we

take the whole direction without searching for the step size, more efforts are needed to prove the strict decrease in the following Theorem 1. However, (9) does hint that (7) is a reasonable update.

Theorem 1 *Let s be the index to be updated at $\mathbf{p}^t$. If*

1. $p_s^t > 0$,
2. $\partial l(\mathbf{p}^t)/\partial p_s \neq 0$, *and*
3. $\sum_{i:s\in I_i}(r_i + r_i') > 0$,

then

$$l(\mathbf{p}^{t+1}) < l(\mathbf{p}^t).$$

The proof is in Appendix A. Note that $\sum_{i:s\in I_i}(r_i + r_i') > 0$ is a reasonable assumption. It means that individual s participates in at least one game.

2.2 Other Methods to Maximize the Likelihood

We briefly discuss other methods to solve (6). For the original Bradley-Terry model, Hunter (2004) discussed how to transform (3) to a logistic regression form: Under certain assumptions[2], the optimal $p_i > 0, \forall i$. Using this property and the constraints $p_j \geq 0, \sum_{j=1}^k p_j = 1$ of (3), we can reparameterize the function (2) by

$$p_s = \frac{e^{\beta_s}}{\sum_{j=1}^k e^{\beta_j}}, \tag{10}$$

and obtain

$$-\sum_{i<j}\left(r_{ij}\log\frac{1}{1+e^{\beta_j-\beta_i}} + r_{ji}\log\frac{e^{\beta_j-\beta_i}}{1+e^{\beta_j-\beta_i}}\right). \tag{11}$$

This is the negative log-likelihood of a logistic regression model. Hence, methods such as iterative weighted least squares (IWLS) (McCullagh and Nelder, 1990) can be used to fit the model. In addition, $\boldsymbol{\beta}$ is now unrestricted, so (3) is transformed to an unconstrained optimization problem. Then conventional optimization techniques such as Newton or Quasi Newton can also be applied.

Now for the generalized model, (6) can still be re-parameterized as an unconstrained problem with the variable $\boldsymbol{\beta}$. However, the negative log-likelihood

$$-\sum_{i=1}^m\left(r_i\log\frac{\sum_{j\in I_i^+}e^{\beta_j}}{\sum_{j\in I_i}e^{\beta_j}} + r_i'\log\frac{\sum_{j\in I_i^-}e^{\beta_j}}{\sum_{j\in I_i}e^{\beta_j}}\right) \tag{12}$$

is not in a form similar to (11), so methods for logistic regression may not be used. Of course Newton or Quasi Newton is still applicable but their implementations are not simpler than Algorithm 2.

2. They will be described in the next section.

3. Convergence of Algorithm 2

Though Theorem 1 has shown the strict decrease of $l(\mathbf{p})$, we must further prove that Algorithm 2 converges to a stationary point of (6). Thus if $l(\mathbf{p})$ is convex, a global optimum is obtained. A vector $\mathbf{p}$ is a stationary (Karash-Kuhn-Tucker) point of (6) if and only if there is a scalar δ and two nonnegative vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\xi}$ such that

$$\nabla f(\mathbf{p})_j = \delta + \lambda_j - \xi_j,$$
$$\lambda_j p_j = 0, \xi_j(1-p_j) = 0, j = 1, \ldots, k.$$

In the following we will prove that under certain conditions Algorithm 2 converges to a point satisfying

$$0 < p_j < 1, \nabla f(\mathbf{p})_j = 0, j = 1, \ldots, k. \tag{13}$$

That is, $\delta = \lambda_j = \xi_j = 0, \forall j$. Problem (6) is quite special as through the convergence proof of Algorithm 2 we show that its optimality condition reduces to (13), the condition without considering constraints. Furthermore, an interesting side-result is that from $\sum_{j=1}^{k} p_j = 1$ and (13), we obtain a point in R^k satisfying $(k+1)$ equations.

If Algorithm 2 stops in a finite number of iterations, then $\partial l(\mathbf{p})/\partial p_j = 0, j = 1, \ldots, k$, which means a stationary point of (6) is already obtained. Thus, we only need to handle the case where $\{\mathbf{p}^t\}$ is an infinite sequence. As $\{\mathbf{p}^t\}_{t=0}^{\infty}$ is in a compact set

$$\{\mathbf{p} \mid 0 \le p_s \le 1, \sum_{j=1}^{k} p_j = 1\},$$

there is at least one convergent subsequence. Assume that $\{\mathbf{p}^t\}, t \in K$ is any such sequence and it converges to $\mathbf{p}^*$. In the following we will show that $\partial l(\mathbf{p}^*)/\partial p_j = 0, j = 1, \ldots, k$.

To prove the convergence of a fixed-point type algorithm (i.e., Lyapunov's theorem), we require $p_s^* > 0, \forall s$. Then if $\partial l(\mathbf{p}^*)/\partial p_s \neq 0$ (i.e., $\mathbf{p}^*$ is not optimal), we can use (7) to find $\mathbf{p}^{*+1} \neq \mathbf{p}^*$, and, as a result of Theorem 1, $l(\mathbf{p}^{*+1}) < l(\mathbf{p}^*)$. This property further leads to a contradiction. To have $p_s^* > 0, \forall s$, for the original Bradley-Terry model, Ford (1957) and Hunter (2004) assume that for any pair of individuals s and j, there is a "path" from s to j; that is, $r_{s,s_1} > 0, r_{s_1,s_2} > 0, \ldots, r_{s_t,j} > 0$. The idea behind this assumption is simple: Since $\sum_{r=1}^{k} p_r^* = 1$, there is at least one $p_j^* > 0$. If in certain games s beats s_1, s_1 beats $s_2, \ldots$, and s_t beats j, then p_s^*, the skill of individual s, should not be as bad as zero. For the generalized model, we make a similar assumption:

Assumption 1 *For any two different individuals s and j, there are $I_{s_0}, I_{s_1}, \ldots, I_{s_t}$, such that either*

1. $r_{s_0} > 0, r_{s_1} > 0, \ldots, r_{s_t} > 0$,
2. $I_{s_0}^+ = \{s\}; I_{s_r}^+ \subset I_{s_{r-1}}, r = 1, \ldots, t; j \in I_{s_t}^-$,

or

1. $r'_{s_0} > 0, r'_{s_1} > 0, \ldots, r'_{s_t} > 0$,

2. $I^-_{s_0} = \{s\}; I^-_{s_r} \subset I_{s_{r-1}}, r = 1, \ldots, t; j \in I^+_{s_t}$.

The idea is that if $p^*_j > 0$, and s beats $I^-_{s_0}$, a subset of I_{s_0} beats $I^-_{s_1}$, a subset of I_{s_1} beats $I^-_{s_2}, \ldots$, and a subset of $I_{s_{t-1}}$ beats $I^-_{s_t}$, which includes j, then p^*_s should not be zero. How this assumption is exactly used is in Appendix B for proving Lemma 2.

Assumption 1 is weaker than that made earlier in (Huang et al., 2005). However, even with the above explanation, this assumption seems to be very strong. Whether the generalized model satisfies Assumption 1 or not, an easy way to fulfill it is to add an additional term

$$-\mu \sum_{s=1}^{k} \log \left(\frac{p_s}{\sum_{j=1}^{k} p_j} \right) \tag{14}$$

to $l(\mathbf{p})$, where μ is a small positive number. That is, for each s, we make an $I_i = \{1, \ldots, k\}$ with $I^+_i = \{s\}, r_i = \mu$, and $r'_i = 0$. As $\sum_{j=1}^{k} p_j = 1$ is one of the constraints, (14) reduces to $-\mu \sum_{s=1}^{k} \log p_s$, which is usually used as a barrier term in optimization to ensure that p_s does not go to zero.

An issue left in Section 2 is whether the multiplicative factor in (7) is well defined. With Assumption 1 and initial $p^0_j > 0, j = 1, \ldots, k$, one can show by induction that $p^t_j > 0, \forall t$ and hence the denominator of (7) is never zero: If $p^t_j > 0$, Assumption 1 implies that there is some i such that $I^+_i = \{j\}$ or $I^-_i = \{j\}$. Then either $\sum_{i:j \in I^+_i} r_i / q^{t,+}_i$ or $\sum_{i:j \in I^-_i} r'_i / q^{t,-}_i$ is positive. Thus, both numerator and denominator in the multiplicative factor are positive, and so is p^{t+1}_j.

The result $p^*_s > 0$ is proved in the following lemma.

Lemma 2 *If Assumption 1 holds, $p^*_s > 0, s = 1, \ldots, k$.*

The proof is in Appendix B.

As the convergence proof will use the strictly decreasing result, we note that Assumption 1 implies the condition $\sum_{i:s \in I_i} (r_i + r'_i) > 0, \forall s$, required by Theorem 1. Finally, the convergence is established:

Theorem 3 *Under Assumption 1, any convergent point of Algorithm 2 is a stationary point of* (6).

The proof is in Appendix C. Though r_i in the Bradley-Terry model is an integer indicating the number of times that team I^+_i beats I^-_i, in the convergence proof we do not use such a property. Hence later for multi-class probability estimates, where r_i is a real number, the convergence result still holds.

Note that a stationary point may be only a saddle point. If (6) is a convex programming problem, then a stationary point is a global minimum. Unfortunately, $l(\mathbf{p})$ may not be convex, so it is not clear whether Algorithm 2 converges to a global minimum or not. The following theorem states that in some cases including the original Bradley-Terry model, any convergent point is a global minimum, and hence a maximum likelihood estimator:

Theorem 4 *Under Assumption 1, if*

1. $|I^+_i| = |I^-_i| = 1, i = 1, \ldots, m$ *or*

2. $|I_i| = k, i = 1, \ldots, m,$

then (6) *has a unique global minimum and Algorithm 2 globally converges to it.*

The proof is in Appendix D. The first case corresponds to the original Bradley-Terry model. Later we will show that under Assumption 1, the second case is related to "one-against-the rest" for multi-class probability estimates. Thus though the theorem seems to be rather restricted, it corresponds to useful situations.

4. Multi-class Probability Estimates

A classification problem is to train a model from data with known class labels and then predict labels of new data. Many classification methods are two-class based approaches and there are different ways to extend them for multi-class cases. Most existing studies focus on predicting class labels but not probability estimates. In this section, we discuss how the generalized Bradley-Terry model can be applied to multi-class probability estimates.

As mentioned in Section 1, there are various ways to decompose a multi-class problem into a number of binary classification problems. Among them, the most commonly used are "one-against-one" and "one-against-the rest." Recently Allwein et al. (2001) proposed a more general framework for the decomposition. Their idea, extended from that of Dietterich and Bakiri (1995), is to associate each class with a row of a $k \times m$ "coding matrix" with all entries from $\{-1, 0, +1\}$. Here m is the number of binary classification problems to be constructed. Each column of the matrix represents a comparison between classes with "-1" and "$+1$," ignoring classes with "0." Note that the classes with "-1" and "$+1$" correspond to our I_i^- and I_i^+, respectively. Then the binary learning method is run for each column of the matrix to obtain m binary decision rules. For a given example, one predicts the class label to be j if the results of the m binary decision rules are "closest" to labels of row j in the coding matrix. Since this coding method can correct errors made by some individual decision rules, it is referred to as *error correcting output codes* (ECOC). Clearly the commonly used "one-against-one" and "one-against-the rest" settings are special cases of this framework.

Given n_i, the number of training data with classes in $I_i = I_i^+ \cup I_i^-$, we assume here that for any given data $\mathbf{x}$,

$$\bar{r}_i = P(\mathbf{x} \text{ in classes of } I_i^+ \mid \mathbf{x} \text{ in classes of } I_i) \tag{15}$$

is available, and the task is to estimate $P(\mathbf{x} \text{ in class } s), s = 1, \ldots, k$. We minimize the (weighted) KL distance between $\bar{r}_i$ and q_i^+/q_i^- similar to (Hastie and Tibshirani, 1998):

$$\min_{\mathbf{p}} \sum_{i=1}^{m} n_i \left(\bar{r}_i \log \frac{\bar{r}_i}{(q_i^+/q_i)} + (1 - \bar{r}_i) \log \frac{1 - \bar{r}_i}{(q_i^-/q_i)} \right). \tag{16}$$

By defining

$$r_i \equiv n_i \bar{r}_i \text{ and } r_i' \equiv n_i(1 - \bar{r}_i), \tag{17}$$

and removing constant terms, (16) reduces to (6), the negative log-likelihood of the generalized Bradley-Terry model. It is explained in Section 1 that one cannot directly interpret

this setting as a generalized Bradley-Terry model. Instead, we minimize the KL distance and obtain the same optimization problem.

We show in Section 5 that many practical "error correcting codes" have the same $|I_i|$, i.e., each binary problem involves the same number of classes. Thus, if data is balanced (all classes have about the same number of instances), then $n_1 \approx \cdots \approx n_m$ and we can remove n_i in (16) without affecting the minimization of $l(\mathbf{p})$. As a result, $r_i = \bar{r}_i$ and $r'_i = 1 - \bar{r}_i$.

In the rest of this section we discuss the case of "one-against-the rest" in detail and the earlier result in (Zadrozny, 2002).

4.1 Properties of the "One-against-the rest" Approach

For this approach, $m = k$ and $I_i, i = 1, \ldots, m$ are

I_i^+	I_i^-	r_i	r'_i
$\{1\}$	$\{2, \ldots, k\}$	r_1	$1 - r_1$
$\{2\}$	$\{1, 3, \ldots, k\}$	r_2	$1 - r_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\{k\}$	$\{1, \ldots, k-1\}$	r_k	$1 - r_k$

Clearly, $|I_i| = k\ \forall i$, so every game involves all classes. Then, $n_1 = \cdots = n_m =$ the total number of training data and the solution of (16) is not affected by n_i. This and (17) suggest that we can solve the problem by simply taking $n_i = 1$ and $r_i + r'_i = 1$, $\forall i$. Thus, (8) can be simplified as

$$\frac{\partial l(\mathbf{p})}{\partial p_s} = -\frac{r_s}{p_s} - \sum_{j:j\neq s} \frac{r'_j}{1 - p_j} + k.$$

Setting $\partial l(\mathbf{p})/\partial p_s = 0\ \forall s$, we have

$$\frac{r_s}{p_s} - \frac{1 - r_s}{1 - p_s} = k - \sum_{j=1}^{k} \frac{r'_j}{1 - p_j}. \tag{18}$$

Since the right-hand side of (18) is the same for all s, we can denote it by δ. If $\delta = 0$, then $p_i = r_i$. This happens only if $\sum_{i=1}^k r_i = 1$. If $\delta \neq 0$, (18) implies

$$p_s = \frac{(1 + \delta) - \sqrt{(1 + \delta)^2 - 4 r_s \delta}}{2\delta}. \tag{19}$$

In Appendix E we show that p_s defined in (19) satisfies $0 \leq p_s \leq 1$. Note that $((1 + \delta) + \sqrt{(1 + \delta)^2 - 4 r_s \delta})/2\delta$ also satisfies (18), but when $\delta < 0$, it is negative and when $\delta > 0$, it is greater than 1. Then the solution procedure is as the following:

If $\sum_{i=1}^k r_i = 1$,
    optimal $\mathbf{p} = [r_1, \ldots, r_k]^T$.
else
    find the root of $\sum_{s=1}^k \frac{(1+\delta) - \sqrt{(1+\delta)^2 - 4 r_s \delta}}{2\delta} - 1 = 0$.
    optimal $p_s =$ (19).

If $\sum_{i=1}^k r_i = 1$, $\mathbf{p} = [r_1, \ldots, r_k]^T$ satisfies $\partial l(\mathbf{p})/\partial p_s = 0\ \forall s$, and thus is the unique optimal solution in light of Theorem 4. For the else part, Appendix E proves that the above equation

of δ has a unique root. Therefore, instead of using Algorithm 2, one can easily solve a one-variable nonlinear equation and obtain the optimal $\mathbf{p}$. This "one-against-the rest" setting is special as we can directly prove the existence of a solution satisfying $k+1$ equations: $\sum_{s=1}^{k} p_s = 1$ and $\partial l(\mathbf{p})/\partial p_s = 0, s = 1, \ldots, k$. Earlier for general models we rely on the convergence proof of Algorithm 2 to show the existence (see the discussion in the beginning of Section 3).

From (19), if $\delta > 0$, larger p_s implies smaller $(1+\delta)^2 - 4r_s\delta$ and hence larger r_s. The situation for $\delta < 0$ is similar. Therefore, the order of $p_1, \ldots, p_k$ is the same as that of $r_1, \ldots, r_k$:

Theorem 5 *If $r_s \geq r_t$, then $p_s \geq p_t$.*

This theorem indicates that results from the generalized Bradley-Terry model are reasonable estimates.

4.2 An Earlier Approach

Zadrozny (2002) was the first to address the probability estimates using error-correcting codes. By considering the same optimization problem (16), she proposes a heuristic updating rule

$$p_s^{t+1} \equiv \frac{\sum_{i:s\in I_i^+} r_i + \sum_{i:s\in I_i^-} r_i'}{\sum_{i:s\in I_i^+} \frac{n_i q_i^{t,+}}{q_i^t} + \sum_{i:s\in I_i^-} \frac{n_i q_i^{t,-}}{q_i^t}} p_s^t, \tag{20}$$

but does not provide a convergence proof. For the "one-against-one" setting, (20) reduces to (4) in Algorithm 1. However, we will show that under other ECOC settings, the algorithm using (20) may not converge to a point with the smallest KL distance. Taking the "one-against-the rest" approach, if $k = 3$ and $r_1 = r_2 = 3/4, r_3 = 1/2$, for our approach Theorem 5 implies $p_1 = p_2$. Then (18) and $p_1 + p_2 + p_3 = 1$ give

$$\frac{3}{4p_1} - \frac{1}{4(1-p_1)} = \frac{1}{2p_3} - \frac{1}{2(1-p_3)} = \frac{1}{2(1-2p_1)} - \frac{1}{4p_1}.$$

This leads to a solution

$$\mathbf{p} = [15 - \sqrt{33}, 15 - \sqrt{33}, 2\sqrt{33} - 6]^T/24, \tag{21}$$

which is also unique according to Theorem 4. If this is a convergent point by using (20), then a further update from it should lead to the same point (after normalization). Thus, the three multiplicative factors must be the same. Since we keep $\sum_{i=1}^{k} p_i^t = 1$ in the algorithm, with the property $r_i + r_i' = 1$, for this example the factor in the updating rule (20) is

$$\frac{r_s + \sum_{i:i\neq s} r_i'}{p_s^t + \sum_{i:i\neq s}(1-p_i^t)} = \frac{k - 1 + 2r_s - \sum_{i=1}^{k} r_i}{k - 2 + 2p_s^t} = \frac{2r_s}{1 + 2p_s^t}. \tag{22}$$

Clearly the $\mathbf{p}$ obtained earlier in (21) by our approach of minimizing the KL distance does not result in the same value for (22). Thus, in this case Zadrozny (2002)'s approach fails to converge to the unique solution of (16) and hence lacks a clear interpretation.

5. Experiments: Simulated Examples

In the following two sections, we present experiments on multi-class probability estimates using synthetic and real-world data. In implementing Algorithm 2, we use the following stopping condition:

$$\max_{s:s\in\{1,\ldots,k\}} \left| \frac{\sum_{i:s\in I_i^+} \frac{r_i}{q_i^{t,+}} + \sum_{i:s\in I_i^-} \frac{r_i'}{q_i^{t,-}}}{\sum_{i:s\in I_i} \frac{r_i+r_i'}{q_i^t}} - 1 \right| < 0.001,$$

which implies that $\partial l(\mathbf{p}^t)/\partial p_s, s = 1, \ldots, k$ are all close to zero.

5.1 Data Generation

We consider the same setting in (Hastie and Tibshirani, 1998; Wu et al., 2004) by defining three possible class probabilities:

(a) $p_1 = 1.5/k$, $p_j = (1 - p_1)/(k - 1)$, $j = 2, \ldots, k$.

(b) $k_1 = k/2$ if k is even, and $(k + 1)/2$ if k is odd; then $p_1 = 0.95 \times 1.5/k_1$, $p_i = (0.95 - p_1)/(k_1 - 1)$ for $i = 2, \ldots, k_1$, and $p_i = 0.05/(k - k_1)$ for $i = k_1 + 1, \ldots, k$.

(c) $p_1 = 0.95 \times 1.5/2$, $p_2 = 0.95 - p_1$, and $p_i = 0.05/(k - 2), i = 3, \ldots, k$.

All classes are competitive in case (a), but only two dominate in (c). For given I_i, $i = 1, \ldots, m$, we generate r_i by adding some noise to q_i^+/q_i and then check if the proposed method obtains good probability estimates. Since q_i^+/q_i of these three cases are different, it is difficult to have a fair way of adding noise. Furthermore, various ECOC settings (described later) will also result in different q_i^+/q_i. Though far from perfect, here we try two ways:

1. An "absolute" amount of noise:

$$r_i = \min(\max(\epsilon, \frac{q_i^+}{q_i} + 0.1N(0, 1)), 1 - \epsilon). \tag{23}$$

   Then $r_i' = 1 - r_i$. Here $\epsilon = 10^{-7}$ is used so that all r_i, r_i' are positive.

   This is the setting considered in (Hastie and Tibshirani, 1998).

2. A "relative" amount of noise:

$$r_i = \min(\max(\epsilon, \frac{q_i^+}{q_i}(1 + 0.1N(0, 1))), 1 - \epsilon). \tag{24}$$

   r_i' and ϵ are set in the same way.

5.2 Results of Various ECOC Settings

We consider the four encodings used in (Allwein et al., 2001) to generate I_i:

1. "1vs1": the pairwise approach (Eq. (5)).

2. "1vsrest": the "One-against-the rest" approach in Section 4.1.

3. "dense": $I_i = \{1, \ldots, k\}$ for all i. I_i is randomly split to two equally-sized sets I_i^+ and I_i^-. $[10 \log_2 k]$ [3] such splits are generated. That is, $m = [10 \log_2 k]$.

   Intuitively, more combinations of subjects as teams give more information and may lead to a better approximation of individual skill. Thus, we would like to select a diversified $I_i^+, I_i^-, i = 1, \ldots, m$. Following Allwein et al. (2001), we repeat the selection 100 times. For each collection of $I_i^+, I_i^-,\ i = 1, \ldots, m$, we calculate the smallest distance between any pair of (I_i^+, I_i^-) and (I_j^+, I_j^-). A larger value indicates better quality of the coding, so we pick the one with the largest value. For the distance between any pair of (I_i^+, I_i^-) and (I_j^+, I_j^-), Allwein et al. (2001) consider a generalized Hamming distance defined as follows:

$$\sum_{s=1}^{k} \begin{cases} 0 & \text{if } s \in I_i^+ \cap I_j^+ \text{ or } s \in I_i^- \cap I_j^-, \\ 1 & \text{if } s \in I_i^+ \cap I_j^- \text{ or } s \in I_i^- \cap I_j^+, \\ 1/2 & \text{if } s \notin I_i \text{ or } s \notin I_j. \end{cases}$$

4. "sparse": I_i^+, I_i^- are randomly drawn from $\{1, \ldots, k\}$ with $E(|I_i^+|) = E(|I_i^-|) = k/4$. Then $[15 \log_2 k]$ such splits are generated. Similar to "dense," we repeat the procedure 100 times to find a good coding.

The way of adding noise may favor some ECOC settings. Since in general

$$\frac{q_i^+}{q_i} \text{ for "1vs1"} \gg \frac{q_i^+}{q_i} \text{ for "1vsrest,"}$$

adding $0.1N(0,1)$ to q_i^+/q_i result in very inaccurate r_i for "1vsrest." On the other hand, if using a relative way, noise added to r_i and r_i' for "1vsrest" is smaller than that for "1vs1." This analysis indicates that using the two different noise makes the experiment more complete.

Figures 1 and 2 show results of adding an "absolute" amount of noise. Two criteria are used to evaluate the obtained probability estimates: Figures 1 presents averaged accuracy rates over 500 replicates for each of the four encodings when $k = 2^2, 2^3, \ldots, 2^6$. Figure 2 gives the (relative) mean squared error (MSE):

$$\text{MSE} = \frac{1}{500} \sum_{j=1}^{500} \left(\sum_{i=1}^{k} (\hat{p}_i^j - p_i)^2 / \sum_{i=1}^{k} p_i^2 \right), \tag{25}$$

where $\hat{\mathbf{p}}^j$ is the probability estimate obtained in the jth of the 500 replicates. Using the same two criteria, Figures 3 and 4 present results of adding a "relative" amount of noise.

3. We use $[x]$ to denote the nearest integer value of x.

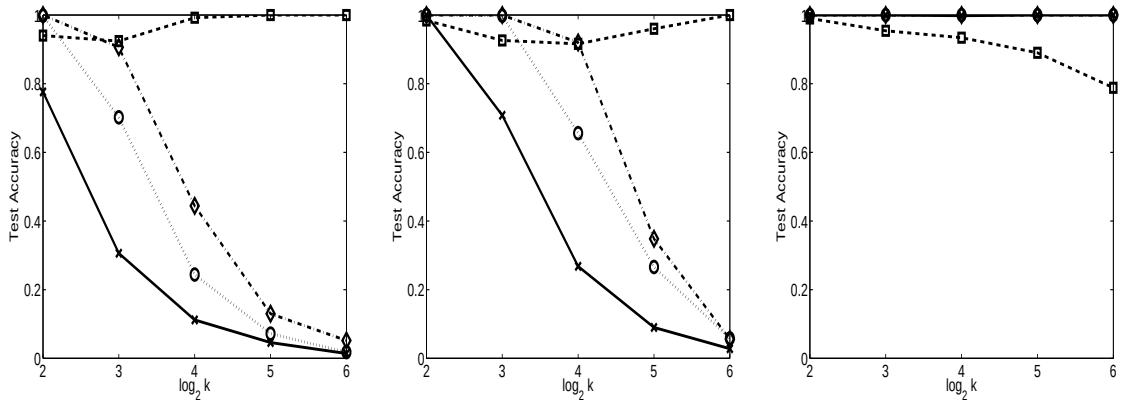


Figure 1: Accuracy of predicting the true class by four encodings and (23) for generating noise: "1vs1" (dashed line, square marked), "1vsrest" (solid line, cross marked), "dense" (dotted line, circle marked), "sparse" (dashdot line, diamond marked). Sub-figures 1(a), 1(b) and 1(c) correspond to the three settings of class probabilities in Section 5.1.

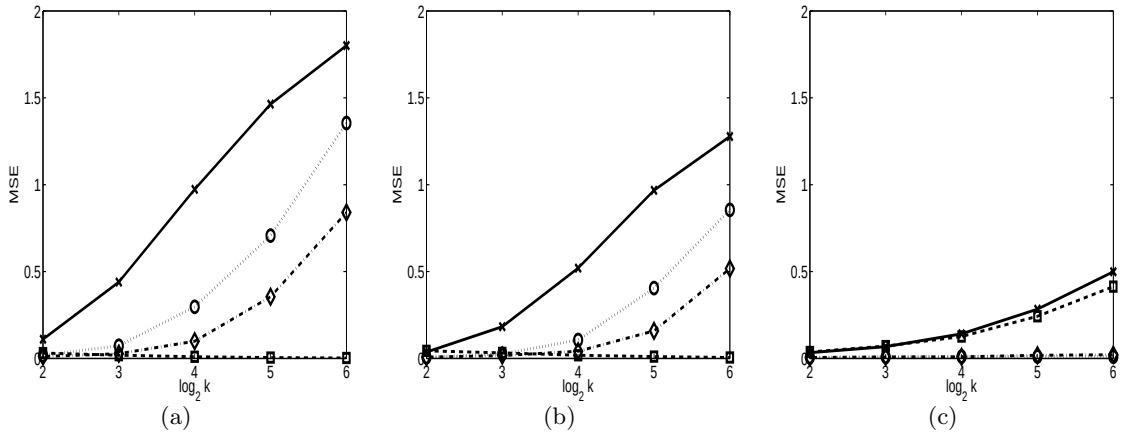


Figure 2: MSE by four encodings and (23) for generating noise. The legend is the same as that of Figure 1.

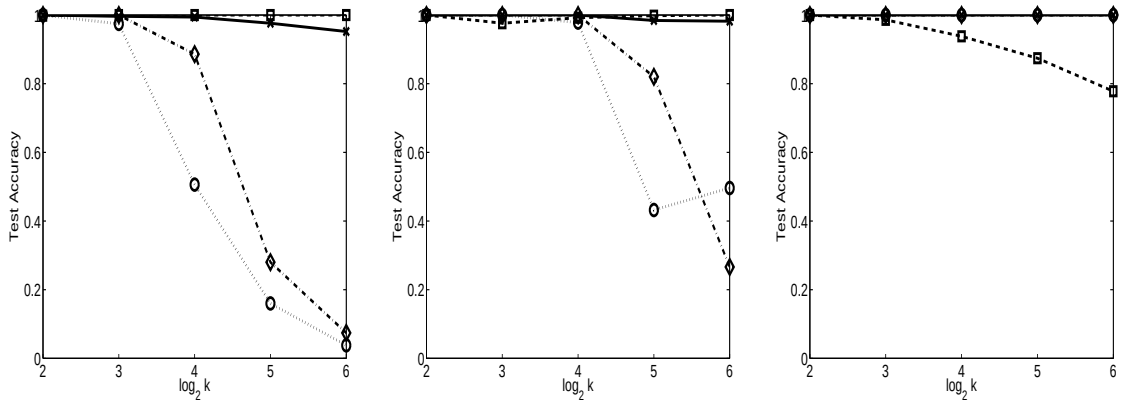


Figure 3: Accuracy of predicting the true class by four encodings and (24) for generating noise. The legend is the same as that of Figure 1.

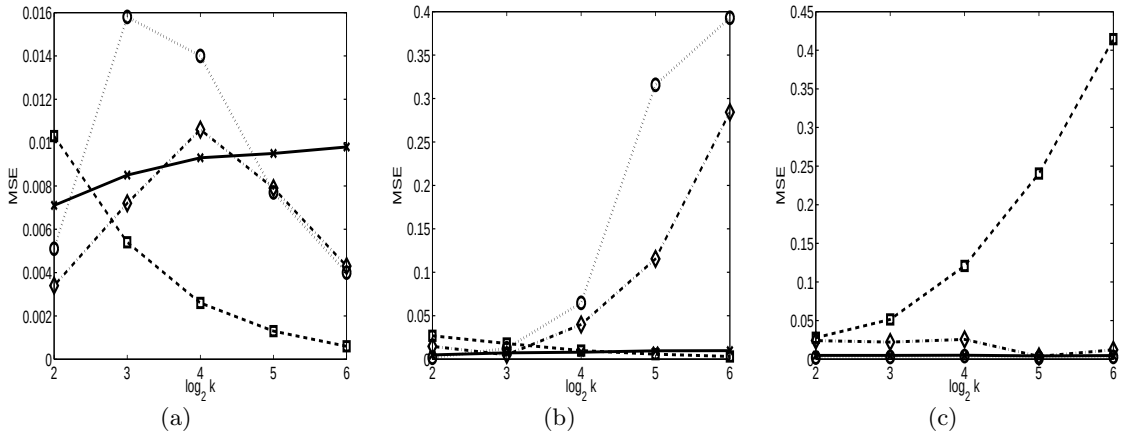
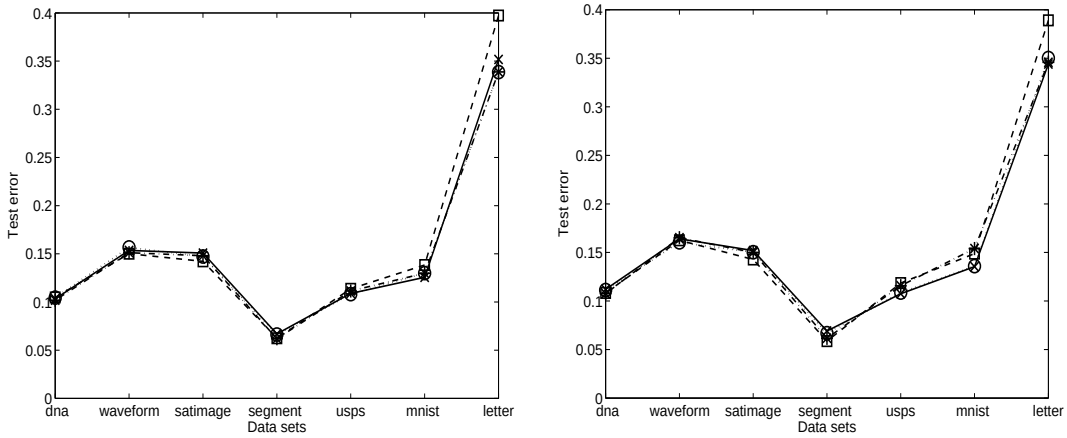


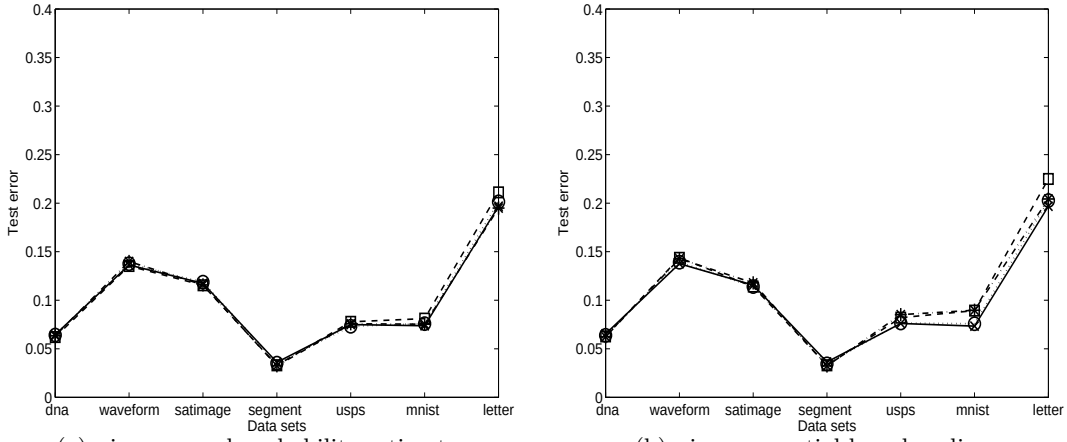
Figure 4: MSE by four encodings and (24) for generating noise. The legend is the same as that of Figure 1.

Clearly, following our earlier analysis on adding noise, results of "1vsrest" in Figures 3 and 4 are much better than those in Figures 1 and 2. In all figures, "dense and "sparse" are less competitive in cases (a) and (b) when k is large. Due to the large $|I_i^+|$ and $|I_i^-|$, the model is unable to single out a clear winner when probabilities are more balanced. For "1vs1," it is good for (a) and (b), but suffers some losses in (c), where the class probabilities are highly unbalanced. Wu et al. (2004) have observed this shortcoming and proposed a quadratic model for the "1vs1" setting.



(a) via proposed probability estimates (b) via exponential-loss decoding

Figure 5: Testing error on smaller (300 training, 500 testing) data sets by four encodings: "1vs1" (dashed line, square marked), "1vsrest" (solid line, cross marked), "dense" (dotted line, circle marked), "sparse" (dashdot line, asterisk marked).



(a) via proposed probability estimates (b) via exponential-loss decoding

Figure 6: Testing error on larger (800 training, 1000 testing) data sets by four encodings. The legend is the same as that of Figure 5.

Results here indicate that the four encodings perform very differently under various conditions. Later in experiments for real data, we will see that in general the situation is closer to case (c), and all four encodings are practically viable[4].

4. Experiments here are done using MATLAB (`http://www.mathworks.com`), and the programs are available at `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/libsvm-errorcode/generalBT.zip`.

6. Experiments: Real Data

In this section we present experimental results on some real-world multi-class problems. There are two goals of experiments here:

1. Check the viability of the proposed multi-class probability estimates. We hope that under reasonable ECOC settings, equally good probabilities are obtained.

2. Compare with the standard ECOC approach without extracting probabilities. This is less important than the first goal as the paper focuses on probability estimates. However, as the classification accuracy is one of the evaluation criteria used here, we can easily conduct a comparison.

6.1 Data and Experimental Settings

We consider data sets used in (Wu et al., 2004): dna, satimage, segment, and letter from the Statlog collection (Michie et al., 1994), waveform from UCI Machine Learning Repository (Blake and Merz, 1998), USPS (Hull, 1994), and MNIST (LeCun et al., 1998). Except dna, which takes two possible values 0 and 1, each attribute of all other data is linearly scaled to $[-1, 1]$. The data set statistics are in Table 1.

Table 1: Data Set Statistics

dataset	dna	waveform	satimage	segment	USPS	MNIST	letter
#classes	3	3	6	7	10	10	26
#attributes	180	21	36	19	256	784	16

After data scaling, we randomly select smaller (300/500) and larger (800/1,000) training/testing sets from thousands of points for experiments. 20 such selections are generated and results are averaged[5].

We use the same four ways in Section 5 to generate I_i. All of them have $|I_1| \approx \cdots \approx |I_m|$. With the property that these multi-class problems are reasonably balanced, we set $n_i = 1$ in (16).

We consider support vector machines (SVM) (Boser et al., 1992; Cortes and Vapnik, 1995) with the RBF (Radial Basis Function) kernel $e^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2}$ as the binary classifier. An improved version (Lin et al., 2003) of (Platt, 2000) obtains r_i using SVM decision values. It is known that SVM may not give good probability estimates (e.g., Zhang (2004)), but Platt (2000) and Wu et al. (2004) empirically show that using decision values from cross validation yields acceptable results in practice. In addition, SVM is sometimes sensitive to parameters, so we conduct a selection procedure before testing. Details can be found in Figure 4 of (Wu et al., 2004). The code is modified from LIBSVM (Chang and Lin, 2001), a library for support vector machines.

5. All training/testing sets used are at `http://www.csie.ntu.edu.tw/~cjlin/papers/svmprob/data`.

6.2 Evaluation Criteria and Results

For these real data sets, there are no true probability values available. We consider the same three evaluation criteria used in (Wu et al., 2004):

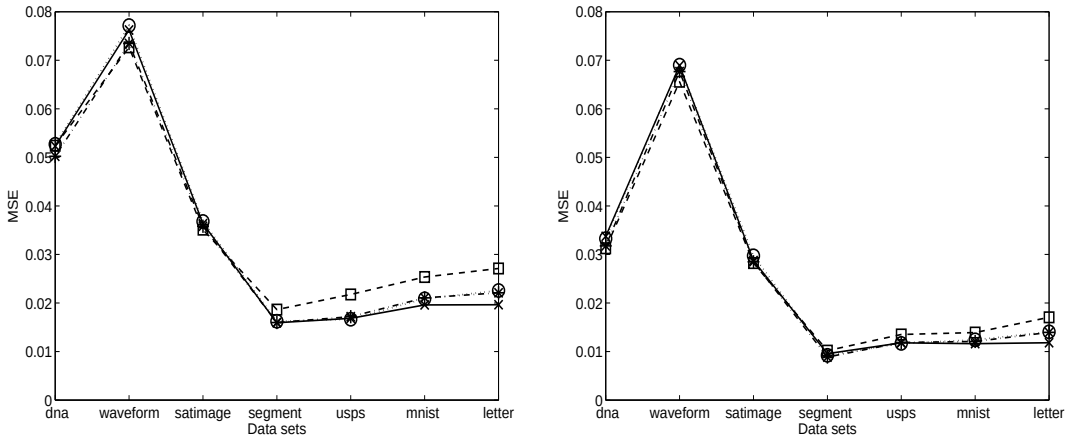
1. Test errors. Averages of 20 errors for smaller and larger sets are in Figures 5(a) and 6(a), respectively.

2. MSE (Brier Score).
$$\frac{1}{l}\sum_{j=1}^{l}\left(\sum_{i=1}^{k}(I_{y_j=i}-\hat{p}_i^j)^2\right),$$
where l is the number of test data, $\hat{\mathbf{p}}^j$ is the probability estimate of the jth data, y_j is the true class label, and $I_{y_j=i}$ is an indicator function (1 if $y_j = i$ and 0 otherwise). This measurement (Brier, 1950), popular in meteorology, satisfies the following property:
$$\arg\min_{\hat{\mathbf{p}}} E_Y[\sum_{i=1}^{k}(I_{Y=i}-\hat{p}_i)^2] \equiv \arg\min_{\hat{\mathbf{p}}}\sum_{i=1}^{k}(\hat{p}_i-p_i)^2,$$
where Y, a random variable for the class label, has the probability distribution $\mathbf{p}$. Brier score is thus useful when the true probabilities are unknown. We present the average of 20 Brier scores in Figure 7.

3. Log loss:
$$\frac{-1}{l}\sum_{j=1}^{l}\log\hat{p}_{y_j}^j,$$
where $\hat{\mathbf{p}}^j$ is the probability estimate of the jth data and y_j is its actual class label. It is another useful criterion when true probabilities are unknown:
$$\min_{\hat{\mathbf{p}}} E_Y[-\sum_{i=1}^{k}\log\hat{p}_i\cdot I_{Y=i}] \equiv \min_{\hat{\mathbf{p}}} -\sum_{i=1}^{k} p_i\log\hat{p}_i$$
has the minimum at $\hat{p}_i = p_i, i = 1,\ldots,k$. Average of 20 splits are presented in Figure 8.

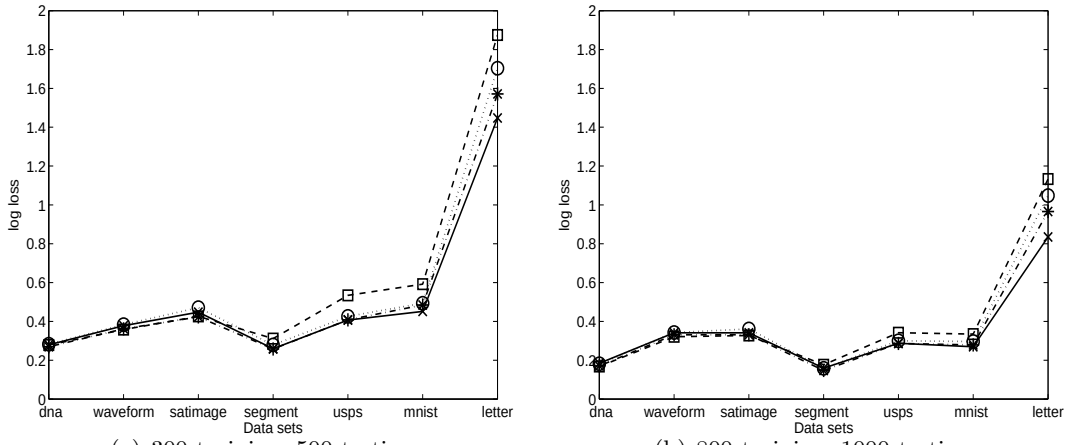
Results of using the three criteria all indicate that the four encodings are quite competitive. Such an observation suggests that in practical problems class probabilities may resemble those specified in case (c) in Section 5; that is, only few classes dominate. Wu et al. (2004) is the first one pointing out this resemblance. In addition, all figures show that "1vs1" is slightly worse than others in the case of larger k (e.g., letter). Earlier Wu et al. (2004) proposed a quadratic model, which gives better probability estimates than the Bradley-Terry model for "1vs1."

In terms of the computational time, because the number of binary problems for "dense" and "sparse" ($[10\log_2 k]$ and $[15\log_2 k]$, respectively) are larger than k, and each binary problem involves many classes of data (all and one half), their training time is longer than that of "1vs1" and "1vsrest." "Dense" is particularly time consuming. Note that though "1vs1" solves $k(k-1)/2$ SVMs, each is small via using only two classes of data.



(a) 300 training, 500 testing (b) 800 training, 1000 testing

Figure 7: MSE by four encodings. The legend is the same as that of Figure 5.



(a) 300 training, 500 testing (b) 800 training, 1000 testing

Figure 8: Log loss by four encodings. The legend is the same as that of Figure 5.

To check the effectiveness of the proposed model in multi-class classification, we compare it with a standard ECOC-based strategy which does not produce probabilities: exponential loss-based decoding by Allwein et al. (2001). Let $\hat{f}_i$ be the decision function of the ith binary classifier, and $\hat{f}_i(\mathbf{x}) > 0$ (< 0) specifies that data $\mathbf{x}$ to be in classes in I_i^+ (I_i^-). This approach determines the predicted label by the following rule:

$$\text{predicted label } = \arg\min_s \left(\sum_{i:s\in I_i^+} e^{-\hat{f}_i} + \sum_{i:s\in I_i^-} e^{\hat{f}_i} \right).$$

Testing errors for smaller and larger sets are in Figures 5(b) and 6(b), respectively. Comparing them with results by the proposed model in Figures 5(a) and 6(a), we observe that both approaches have very similar errors. Therefore, in terms of predicting class labels only, our new method is competitive.

7. Extensions of the Generalized Bradley-Terry Model

In addition to multiclass probability estimates, the proposed generalized Bradley-Terry model, as mentioned in Section 1, has some potential applications in sports. We consider in this section several extensions based on common sport scenarios and show that, with a slight modification of Algorithm 2, they can be easily solved as well.

7.1 Weighted Individual Skill

In some sports, team performance is highly affected by certain positions. For example, many people think guards are relatively more important than centers and forwards in basketball games. We can extend the generalized Bradley-Terry model to this case: Define

$$\bar{q}_i \equiv \sum_{j:j\in I_i} w_{ij}p_j, \quad \bar{q}_i^+ \equiv \sum_{j:j\in I_i^+} w_{ij}p_j, \quad \bar{q}_i^- \equiv \sum_{j:j\in I_i^-} w_{ij}p_j,$$

where $w_{ij} > 0$ is a given weight parameter reflecting individual j's position in the game between I_i^+ and I_i^-. By minimizing the same negative log-likelihood function (6), estimated individual skill can be obtained. Here Algorithm 2 can still be applied but with the updating rule replaced by

$$p_s^{t+1} = \frac{\sum_{i:s\in I_i^+} \frac{r_i w_{is}}{\bar{q}_i^{t,+}} + \sum_{i:s\in I_i^-} \frac{r_i' w_{is}}{\bar{q}_i^{t,-}}}{\sum_{i:s\in I_i} \frac{(r_i+r_i')w_{is}}{\bar{q}_i^t}} p_s^t, \tag{26}$$

which is derived similarly to (7) so that the multiplicative factor is equal to one when $\partial l(\mathbf{p})/\partial p_s = 0$. The convergence can be proved similarly. However, it may be harder to obtain the global optimality: Case 1 in Theorem 4 still holds, but Case 2 may not since $\bar{q}_i$ needs not be equal to one (the proof of Case 2 requires $q_i = 1$, which is guaranteed by $|I_i| = k$).

7.2 Home-field Advantage

The original home-field advantage model (Agresti, 1990) is based on paired individual comparisons. We can incorporate its idea into our proposed model by taking

$$P(I_i^+ \text{ beats } I_i^-) = \begin{cases} \frac{\theta q_i^+}{\theta q_i^+ + q_i^-} & \text{if } I_i^+ \text{ is home,} \\ \frac{q_i^+}{q_i^+ + \theta q_i^-} & \text{if } I_i^- \text{ is home,} \end{cases}$$

where $\theta > 0$ measures the strength of the home-field advantage or disadvantage. Note that θ is an unknown parameter to be estimated, while the weights w_{ij} in Section 7.1 are given.

Let $\bar{r}_i \geq 0$ and $\bar{r}_i' \geq 0$ be the number of times that I_i^+ wins and loses at home, respectively. For I_i^+'s away games, we let $\tilde{r}_i \geq 0$ and $\tilde{r}_i' \geq 0$ be the number of times that I_i^+ wins and loses. The minimization of the negative log-likelihood function thus becomes:

$$\min_{\mathbf{p},\theta} \; l(\mathbf{p},\theta) = \\ -\sum_{i=1}^{m} \left(\bar{r}_i \log \frac{\theta q_i^+}{\theta q_i^+ + q_i^-} + \tilde{r}_i \log \frac{q_i^+}{q_i^+ + \theta q_i^-} + \bar{r}_i' \log \frac{q_i^-}{\theta q_i^+ + q_i^-} + \tilde{r}_i' \log \frac{\theta q_i^-}{q_i^+ + \theta q_i^-} \right)$$

under the constraints in (6) and the condition $\theta \geq 0$.

To apply Algorithm 2 on the new optimization problem, we must modify the updating rule. For each s, $\partial l(\mathbf{p},\theta)/\partial p_s = 0$ leads to the following rule[6]:

$$p_s^{t+1} = \frac{\sum_{i:s\in I_i^+} \frac{\bar{r}_i+\tilde{r}_i}{q_i^+} + \sum_{i:s\in I_i^-} \frac{\bar{r}'_i+\tilde{r}'_i}{q_i^-}}{\sum_{i:s\in I_i^+} \left(\frac{\theta(\bar{r}_i+\bar{r}'_i)}{\theta q_i^+ + q_i^-} + \frac{\tilde{r}_i+\tilde{r}'_i}{q_i^+ + \theta q_i^-}\right) + \sum_{i:s\in I_i^-} \left(\frac{\bar{r}_i+\bar{r}'_i}{\theta q_i^+ + q_i^-} + \frac{\theta(\tilde{r}_i+\tilde{r}'_i)}{q_i^+ + \theta q_i^-}\right)} p_s^t. \tag{27}$$

For θ, from $\partial l(\mathbf{p},\theta)/\partial\theta = 0$, we have

$$\theta^{t+1} = \frac{\sum_{i=1}^m (\bar{r}_i + \tilde{r}'_i)}{\sum_{i=1}^m \left(\frac{q_i^+(\bar{r}_i+\bar{r}'_i)}{\theta^t q_i^+ + q_i^-} + \frac{q_i^-(\tilde{r}_i+\tilde{r}'_i)}{q_i^+ + \theta^t q_i^-}\right)}. \tag{28}$$

Unlike the case of updating p_s^t, there is no need to normalize θ^{t+1}. The algorithm then cyclically updates $p_1,\ldots,p_k$, and θ. If p_s is updated, we can slightly modify the proof of Theorem 1 and obtain the strict decrease of $l(\mathbf{p},\theta)$. Moreover, Appendix F gives a simple derivation of $l(\mathbf{p}^t,\theta^{t+1}) < l(\mathbf{p}^t,\theta^t)$. Thus, if we can ensure that θ^t is bounded above, then under a modified version of Assumption 1 where $\max(\bar{r}_{s_i},\tilde{r}_{s_i}) > 0$ replaces $r_{s_i} > 0$, the convergence of Algorithm 2 (i.e., Theorem 3) still holds by a similar proof.

7.3 Ties

Suppose ties are possible between teams. Extending the model proposed in (Rao and Kupper, 1967), we consider:

$$\begin{aligned} P(I_i^+ \text{ beats } I_i^-) &= \frac{q_i^+}{q_i^+ + \theta q_i^-}, \\ P(I_i^- \text{ beats } I_i^+) &= \frac{q_i^-}{\theta q_i^+ + q_i^-}, \text{ and} \\ P(I_i^+ \text{ ties } I_i^-) &= \frac{(\theta^2-1) q_i^+ q_i^-}{(q_i^+ + \theta q_i^-)(\theta q_i^+ + q_i^-)}, \end{aligned}$$

where $\theta > 1$ is a threshold parameter to be estimated.

Let t_i be the number of times that I_i^+ ties I_i^- and r_i, r'_i defined as before. We then minimize the following negative log-likelihood function:

$$\begin{aligned} &\min_{\mathbf{p},\theta} \; l(\mathbf{p},\theta) \\ &= -\sum_{i=1}^m \left(r_i \log \frac{q_i^+}{q_i^+ + \theta q_i^-} + r'_i \log \frac{q_i^-}{\theta q_i^+ + q_i^-} + t_i \log \frac{(\theta^2-1) q_i^+ q_i^-}{(q_i^+ + \theta q_i^-)(\theta q_i^+ + q_i^-)} \right) \\ &= -\sum_{i=1}^m \left(r_i \log \frac{q_i^+}{q_i^+ + \theta q_i^-} + r'_i \log \frac{q_i^-}{\theta q_i^+ + q_i^-} + t_i \log \frac{\theta q_i^+}{\theta q_i^+ + q_i^-} + t_i \log \frac{\theta q_i^-}{q_i^+ + \theta q_i^-} \right) \qquad (29) \\ &\quad -\sum_{i=1}^m t_i \log \frac{\theta^2-1}{\theta^2} \qquad (30) \end{aligned}$$

6. For convenience, $q_i^{t,+}(q_i^{t,-})$ is abbreviated as $q_i^+(q_i^-)$. The same abbreviation is used in the updating rule in Sections 7.3 and 7.4.

under the constraints in (6) and the condition $\theta > 1$.

For updating p_s^t, θ is considered as a constant and (29) is in a form of the Home-field model, so the rule is similar to (27). The strict decrease of $l(\mathbf{p}, \theta)$ can be established as well. For updating θ, we have

$$\theta^{t+1} = \frac{1}{2C_t} + \sqrt{1 + \frac{1}{4C_t^2}}, \tag{31}$$

where

$$C_t = \frac{1}{2\sum_{i=1}^m t_i} \left(\sum_{i=1}^m \frac{(r_i + t_i) q_i^-}{q_i^+ + \theta^t q_i^-} + \sum_{i=1}^m \frac{(r_i' + t_i) q_i^+}{\theta^t q_i^+ + q_i^-} \right).$$

The derivation and the strict decrease of $l(\mathbf{p}, \theta)$ are in Appendix F. If we can ensure that $1 < \theta^t < \infty$ and modify Assumption 1 as in Section 7.2, the convergence of Algorithm 2 also holds.

7.4 Multiple Team Comparisons

In this type of comparison, a game may include more than two participants, and the result is a ranking of the participants. For a game of three participants. Pendergrass and Bradley (1960) proposed using

$$\begin{aligned} & P(i \text{ best, } j \text{ in the middle, and } k \text{ worst}) \\ = \; & P(i \text{ beats } j \text{ and } k) \cdot P(j \text{ beats } k) \\ = \; & \frac{p_i}{p_i + (p_j + p_k)} \cdot \frac{p_j}{p_j + p_k}. \end{aligned}$$

A general model introduced in (Placket, 1975) is:

$$P(a(1) \to a(2) \to \cdots \to a(k)) = \prod_{i=1}^k \frac{p_{a(i)}}{p_{a(i)} + p_{a(i+1)} + \cdots + p_{a(k)}}, \tag{32}$$

where $a(i), 1 \le i \le k$ is the ith ranked individual and $\to$ denotes the relation "is ranked higher than." A detailed discussion of this model is in (Hunter, 2004, Section 5).

With similar ideas, we consider a more general setting: Each game may include more than two participating teams. Assume that there are k individuals and N games resulting in N rankings; the mth game involves g_m disjoint teams. Let $I_m^i \subset \{1, \ldots, k\}$ be the ith ranked team in the mth game, $1 \le i \le g_m$, $1 \le m \le N$. We consider the model:

$$P(I_m^1 \to I_m^2 \to \cdots \to I_m^{g_m}) = \prod_{i=1}^{g_m} \frac{\sum_{s:s \in I_m^i} p_s}{\sum_{j=i}^{g_m} \sum_{s:s \in I_m^j} p_s}. \tag{33}$$

Defining

$$q_m^i = \sum_{s:s \in I_m^i} p_s,$$

we minimize the negative log-likelihood function:

$$\min_{\mathbf{p}} l(\mathbf{p}) = -\sum_{m=1}^{N}\sum_{i=1}^{g_m} \log \frac{q_m^i}{\sum_{j=i}^{g_m} q_m^j} \tag{34}$$

under the constraints in (6).

In fact, (34) is a special case of (6). Each ranking can be viewed as the result of a series of paired team comparisons: the first ranked team beats the others, the second ranked team beats the others except the first, and so on; for each paired comparison, $r_i = 1$ and $r'_i = 0$. Therefore, Algorithm 2 can be applied and the updating rule is:

$$p_s^{t+1} = \frac{\sum_{j:s\in I_j} (q_j^{\phi_j(s)})^{-1}}{\sum_{j:s\in I_j}\sum_{i=1}^{\phi_j(s)} (\sum_{v=i}^{g_j} q_j^v)^{-1}} p_s^t, \tag{35}$$

where $\phi_j(s)$ is the rank of the team that individual s belongs to in the jth game and $I_j = \cup_{i=1}^{g_j} I_j^i$.

We explain in detail how (35) is derived. Since teams are disjoint in one game and (33) implies that ties are not allowed, $\phi_j(i)$ is unique under a given i. In the jth game, individual s appears in $\phi_j(s)$ paired comparisons:

$$\begin{array}{lll} I_j^1 & \text{vs.} & I_j^2 \cup \cdots \cup I_j^{g_j}, \\ I_j^2 & \text{vs.} & I_j^3 \cup \cdots \cup I_j^{g_j}, \\ & \vdots & \\ I_j^{\phi_j(s)} & \text{vs.} & I_j^{\phi_j(s)+1} \cup \cdots \cup I_j^{g_j}. \end{array}$$

From (7), the numerator of the multiplicative factor involves winning teams that individual s is in, so there is only one (i.e., $\phi_j(s)$) in each game that s joins; the denominator involves teams of both sides, so it is in the form of $\sum_{i=1}^{\phi_j(s)} (\sum_{v=i}^{g_j} q_j^v)^{-1}$.

8. Discussion and Conclusions

We propose a generalized Bradley-Terry model which gives individual skill from group competition results. We develop a simple iterative method to maximize the log-likelihood and prove the convergence. The new model has many potential applications. In particular, minimizing the negative log likelihood of the proposed model coincides with minimizing the KL distance for multi-class probability estimates under error correcting output codes. Hence the iterative scheme is useful for finding class probabilities. Similar to the original Bradley-Terry model, we can extend the proposed generalized model to other settings such as home-field advantages, ties, and multiple team comparisons.

Investigating more practical applications using the proposed model is certainly an important future direction. The lack of convexity of $l(\mathbf{p})$ also requires more studies. In Section 5, the "sparse" coding has $E(|I_i^+|) = E(|I_i^-|) = k/4$, and hence is not covered by Theorem 4 which proves the global optimality. However, this coding is competitive with others in Section 6. If possible, we hope to show in the future that in general the global optimality holds.

Acknowledgments

This work was supported in part by the National Science Council of Taiwan via the grants NSC 92-2213-E-002-062 and NSC 92-2118-M-004-003.

References

A. Agresti. *Categorical Data Analysis*. Wiley, New York, 1990.

Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1: 113–141, 2001. ISSN 1533-7928.

C. L. Blake and C. J. Merz. UCI repository of machine learning databases. Technical report, University of California, Department of Information and Computer Science, Irvine, CA, 1998. Available at `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.

R. A. Bradley and M. Terry. The rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324–345, 1952.

G. W. Brier. Verification of forecasts expressed in probabilities. *Monthly Weather Review*, 78:1–3, 1950.

Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

C. Cortes and V. Vapnik. Support-vector network. *Machine Learning*, 20:273–297, 1995.

H. A. David. *The method of paired comparisons*. Oxford University Press, New York, second edition, 1988.

R. R. Davidson and P. H. Farquhar. A bibliography on the method of paired comparisons. *Biometrics*, 32:241–252, 1976.

Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995. URL `citeseer.ist.psu.edu/dietterich95solving.html`.

L. R. Jr. Ford. Solution of a ranking problem from binary comparisons. *American Mathematical Monthly*, 64(8):28–33, 1957.

T. Hastie and R. Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26(1):451–471, 1998.

Tzu-Kuo Huang, Ruby C. Weng, and Chih-Jen Lin. A generalized Bradley-Terry model: From group competition to individual skill. In *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, 2005.

J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, May 1994.

David R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32:386–408, 2004.

Yann LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. MNIST database available at `http://yann.lecun.com/exdb/mnist/`.

Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C. Weng. A note on Platt's probabilistic outputs for support vector machines. Technical report, Department of Computer Science, National Taiwan University, 2003. URL `http://www.csie.ntu.edu.tw/~cjlin/papers/plattprob.ps`.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. CRC Press, 2nd edition, 1990.

D. Michie, D. J. Spiegelhalter, and C. C. Taylor. *Machine Learning, Neural and Statistical Classification*. Prentice Hall, Englewood Cliffs, N.J., 1994. Data available at `http://www.ncc.up.pt/liacc/ML/statlog/datasets.html`.

R. N. Pendergrass and R. A. Bradley. Ranking in triple comparisons. In Ingram Olkin, editor, *Contributions to Probability and Statistics*. Stanford University Press, Stanford, CA, 1960.

R. L. Placket. The analysis of permutations. *Applied Statistics*, 24:193–202, 1975.

J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, Cambridge, MA, 2000. MIT Press. URL `citeseer.nj.nec.com/platt99probabilistic.html`.

P. V. Rao and L. L. Kupper. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62:194–204, 1967. [Corrigendum J. Amer. Statist. Assoc. 63 1550-1551].

G. Simons and Y.-C. Yao. Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *The Annals of Statistics*, 27(3): 1041–1060, 1999.

Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004. URL `http://www.csie.ntu.edu.tw/~cjlin/papers/svmprob/svmprob.pdf`.

B. Zadrozny. Reducing multiclass to binary by coupling probability estimates. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 1041–1048. MIT Press, Cambridge, MA, 2002.

E. Zermelo. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29:436–460, 1929.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004.

Appendix A. Proof of Theorem 1

Define

$$q_{i\backslash s}^{t,+} \equiv \sum_{j\in I_i^+, j\neq s} p_j^t, \text{ and } q_{i\backslash s}^{t} \equiv \sum_{j\in I_i^-, j\neq s} p_j^t.$$

Using

$$-\log x \geq 1 - \log y - x/y \text{ with equality if and only if } x = y,$$

we have

$$Q^1(p_s) \geq l([p_1^t, \ldots, p_{s-1}^t, p_s, p_{s+1}^t, \ldots, p_k^t]^T) \text{ with equality if } p_s = p^t,$$

where

$$\begin{aligned} Q^1(p_s) &\equiv -\sum_{i:s\in I_i^+} r_i \left(\log(q_{i\backslash s}^{t,+} + p_s) - \frac{q_{i\backslash s}^t + p_s}{q_i^t} - \log q_i^t + 1 \right) - \\ &\quad \sum_{i:s\in I_i^-} r_i' \left(\log(q_{i\backslash s}^{t,-} + p_s) - \frac{q_{i\backslash s}^t + p_s}{q_i^t} - \log q_i^t + 1 \right) \\ &= -\sum_{i:s\in I_i^+} r_i \log(q_{i\backslash s}^{t,+} + p_s) - \sum_{i:s\in I_i^-} r_i' \log(q_{i\backslash s}^{t,-} + p_s) + \\ &\quad \sum_{i:s\in I_i} (r_i + r_i') \left(\frac{q_{i\backslash s}^t + p_s}{q_i^t} + \log q_i^t - 1 \right). \end{aligned}$$

For $0 < \lambda < 1$, we have

$$\log(\lambda x + (1-\lambda)y) \geq \lambda \log x + (1-\lambda)\log y \text{ with equality if and only if } x = y.$$

With $p_s^t > 0$,

$$\begin{aligned} & \log(q_{i\backslash s}^{t,+} + p_s) \\ = \; & \log\left(\frac{q_{i\backslash s}^{t,+}}{q_i^{t,+}} \cdot 1 + \frac{p_s^t}{q_i^{t,+}} \cdot \frac{p_s}{p_s^t} \right) + \log(q_i^{t,+}) \\ \geq \; & \frac{p_s^t}{q_i^{t,+}} (\log p_s - \log p_s^t) + \log q_i^{t,+} \text{ with equality if } p_s = p_s^t. \end{aligned}$$

Then

$$Q^2(p_s) \geq l([p_1^t, \ldots, p_{s-1}^t, p_s, p_{s+1}^t, \ldots, p_k^t]^T) \text{ with equality if } p_s = p_s^t,$$

where

$$\begin{aligned} & Q^2(p_s) \\ \equiv & -\sum_{i:s\in I_i^+} r_i \left(\frac{p_s^t}{q_i^{t,+}}(\log p_s - \log p_s^t) + \log q_i^{t,+} \right) - \\ & \sum_{i:s\in I_i^-} r_i' \left(\frac{p_s^t}{q_i^{t,-}}(\log p_s - \log p_s^t) + \log q_i^{t,-} \right) + \sum_{i:s\in I_i} (r_i + r_i') \left(\frac{q_{i\setminus s}^t + p_s}{q_i^t} + \log q_i^t - 1 \right). \end{aligned}$$

As we assume $p_s^t > 0$ and $\sum_{i:s\in I_i}(r_i + r_i') > 0$, $Q^2(p_s)$ is a strictly convex function of p_s. By $dQ^2(p_s)/dp_s = 0$,

$$\left(\sum_{i:s\in I_i^+} \frac{r_i}{q_i^{t,+}} \sum_{i:s\in I_i^-} \frac{r_i'}{q_i^{t,-}} \right) \frac{p_s^t}{p_s} = \sum_{i:s\in I_i} \frac{r_i + r_i'}{q_i^t}$$

leads to the updating rule. Thus, if $p_s^{t+1} \neq p_s^t$, then

$$l(\mathbf{p}^{t+1}) \leq Q^2(p_s^{t+1}) < Q^2(p_s^t) = l(\mathbf{p}^t).$$

Appendix B. Proof of Lemma 2

If the result does not hold, there is an index $\bar{s}$ and an infinite index set T such that

$$\lim_{t\in T, t\to\infty} p_{\bar{s}}^t = p_{\bar{s}}^* = 0.$$

Since $\sum_{s=1}^k p_s^t = 1, \forall t$ and k is finite,

$$\lim_{t\in T, t\to\infty} \sum_{s=1}^k p_s^t = \sum_{s=1}^k p_s^* = 1.$$

Thus, there is an index j such that

$$\lim_{t\in T, t\to\infty} p_j^t = p_j^* > 0. \tag{36}$$

Under Assumption 1, one of the two conditions linking individual $\bar{s}$ and j must hold. As both cases are similar, we consider only the first here. With $p_{\bar{s}}^* = 0$ and $I_{s_0}^+ = \{\bar{s}\}$, we claim that $p_u^* = 0, \forall u \in I_{s_0}^-$. If this claim is wrong, then

$$\begin{aligned} l(\mathbf{p}^t) &= -\sum_{i=1}^m \left(r_i \log \frac{q_i^{t,+}}{q_i^t} + r_i' \log \frac{q_i^{t,-}}{q_i^t} \right) \\ &\geq -r_{s_0} \log \frac{q_{s_0}^{t,+}}{q_{s_0}^t} \\ &= -r_{s_0} \log \frac{p_{\bar{s}}^t}{p_{\bar{s}}^t + \sum_{u\in I_{s_0}^-} p_u^t} \\ &\to \infty \text{ when } t \in T, t \to \infty. \end{aligned}$$

This result contradicts Theorem 1, which implies $l(\mathbf{p}^t)$ is bounded above by $l(\mathbf{p}^0)$. Thus, $p_u^* = 0, \forall u \in I_{s_0}$. With $I_{s_1}^+ \subset I_{s_0}$, we can use the same way to prove $p_u^* = 0, \forall u \in I_{s_1}$. Continuing the same derivation, in the end $p_u^* = 0, \forall u \in I_{s_t}$. Since $j \in I_{s_t}^-$, $p_j^* = 0$ contradicts (36) and the proof is complete.

Appendix C. Proof of Theorem 3

Recall that we assume $\lim_{t\in K, t\to\infty} \mathbf{p}^t = \mathbf{p}^*$. For each $\mathbf{p}^t, t \in K$, there is a corresponding index s for the updating rule (7). Thus, one of $\{1, \ldots, k\}$ must be considered infinitely many times. Without loss of generality, we assume that all $\mathbf{p}^t, t \in K$ have the same corresponding s. If $\mathbf{p}^*$ does not satisfy

$$\frac{\partial l(\mathbf{p}^*)}{\partial p_j} = 0, \ j = 1, \ldots, k,$$

starting from $s, s+1, \ldots, k, 1, \ldots, s-1$, there is a first component $\bar{s}$ such that $\partial l(\mathbf{p}^*)/\partial p_{\bar{s}} \neq 0$. As $p_{\bar{s}}^* > 0$, by applying one iteration of Algorithm 2 on $p_{\bar{s}}^*$, and using Theorem 1, we obtain $\mathbf{p}^{*+1} \neq \mathbf{p}^*$ and

$$l(\mathbf{p}^{*+1}) < l(\mathbf{p}^*). \tag{37}$$

Since $\bar{s}$ is the first index so that the partial derivative is not zero,

$$\frac{\partial l(\mathbf{p}^*)}{\partial p_s} = 0 = \cdots = \frac{\partial l(\mathbf{p}^*)}{\partial p_{\bar{s}-1}}.$$

Thus, at the tth iteration,

$$\lim_{t\in K, t\to\infty} p_s^{t+1} = \lim_{t\in K, t\to\infty} \frac{\sum_{i:s\in I_i^+} \frac{r_i}{q_i^{t,+}} + \sum_{i:s\in I_i^-} \frac{r_i'}{q_i^{t,-}}}{\sum_{i:s\in I_i} \frac{r_i + r_i'}{q_i^t}} p_s^t = \frac{\sum_{i:s\in I_i^+} \frac{r_i}{q_i^{*,+}} + \sum_{i:s\in I_i^-} \frac{r_i'}{q_i^{*,-}}}{\sum_{i:s\in I_i} \frac{r_i + r_i'}{q_i^*}} p_s^* = p_s^*$$

and hence

$$\lim_{t\in K, t\to\infty} \mathbf{p}^{t+1} = \lim_{t\in K, t\to\infty} \mathbf{p}^t = \mathbf{p}^*.$$

Assume $\bar{s}$ corresponds to the $\bar{t}$th iteration, by a similar derivation,

$$\lim_{t\in K, t\to\infty} \mathbf{p}^{t+1} = \cdots = \lim_{t\in K, t\to\infty} \mathbf{p}^{\bar{t}} = \mathbf{p}^*$$

and

$$\lim_{t\in K, t\to\infty} \mathbf{p}^{\bar{t}+1} = \mathbf{p}^{*+1}.$$

Thus, with (37),

$$\lim_{t\in K, t\to\infty} l(\mathbf{p}^{\bar{t}+1}) = l(\mathbf{p}^{*+1}) < l(\mathbf{p}^*)$$

contradicts the fact that

$$l(\mathbf{p}^*) \leq \cdots \leq l(\mathbf{p}^t), \forall t.$$

Appendix D. Proof of Theorem 4

The first case reduces to the original Bradley-Terry model, so we can directly use existing results. As explained in Section 3, Assumption 1 goes back to (Hunter, 2004, Assumption 1). Then the results in Section 4 of (Hunter, 2004) imply that (6) has a unique global minimum and Algorithm 2 globally converges to it.

For the second case, as $q_i = \sum_{j=1}^{k} p_j = 1$, $l(\mathbf{p})$ can be reformulated as

$$\bar{l}(\mathbf{p}) = -\sum_{i=1}^{m} (r_i \log q_i^+ + r_i' \log q_i^-), \tag{38}$$

which is a convex function of $\mathbf{p}$. Then solving (6) is equivalent to minimizing (38). One can also easily show that they have the same set of stationary points.

From Assumption 1, for each p_j, there is i such that either $I_i^+ = \{s\}$ with $r_i > 0$, or $I_i^- = \{s\}$ with $r_i' > 0$. Therefore, either $-r_i \log p_s$ or $-r_i' \log p_s$ appears in (38). Since they are strictly convex functions of p_s, the summation on all $s = 1, \ldots, k$ makes (38) a strictly convex function of $\mathbf{p}$. Hence (6) has a unique global minimum, which is also (6)'s unique stationary point. From Theorem 3, Algorithm 2 globally converges to this unique minimum.

Appendix E. Solving Nonlinear Equations for the "One-against-the-rest" Approach

We show that if $\delta \neq 0$, p_s defined in (19) satisfies $0 \leq p_s \leq 1$. If $\delta > 0$, then

$$(1+\delta) \geq \sqrt{(1+\delta)^2 - 4r_s\delta},$$

so $p_s \geq 0$. The situation for $\delta < 0$ is similar. To prove $p_s \leq 1$, we consider three cases:

1. $\delta \geq 1$.
   Clearly,
   $$p_s = \frac{(1+\delta) - \sqrt{(1+\delta)^2 - 4r_s\delta}}{2\delta}. \leq \frac{1+\delta}{2\delta} \leq 1.$$

2. $0 < \delta < 1$.
   With $0 \leq r_s \leq 1$, we have $4\delta - 4r_s\delta \geq 0$ and
   $$(1+\delta)^2 - 4r_s\delta \geq 1 - 2\delta + \delta^2.$$
   Using $0 < \delta < 1$,
   $$p_s = \frac{(1+\delta) - \sqrt{(1+\delta)^2 - 4r_s\delta}}{2\delta}. \leq \frac{1+\delta-(1-\delta)}{2\delta} = 1.$$

3. $\delta < 0$.
   Now $4\delta - 4r_s\delta \leq 0$, so
   $$0 \leq (1+\delta)^2 - 4r_s\delta \leq 1 - 2\delta + \delta^2.$$
   Then
   $$-\sqrt{(1+\delta)^2 - 4r_s\delta} \geq \delta - 1.$$
   Adding $1+\delta$ on both sides and dividing them by 2δ leads to $p_s \leq 1$.

To find δ by solving $\sum_{s=1}^{k} p_s - 1 = 0$, the discontinuity at $\delta = 0$ is a concern. A simple calculation shows

$$\lim_{\delta \to 0} \sum_{s=1}^{k} \frac{(1+\delta) - \sqrt{(1+\delta)^2 - 4r_s\delta}}{2\delta} - 1 = \sum_{s=1}^{k} r_s - 1.$$

One can thus define the following continuous function:

$$f(\delta) = \begin{cases} \sum_{s=1}^{k} r_s - 1 & \text{if } \delta = 0, \\ \sum_{s=1}^{k} \frac{(1+\delta) - \sqrt{(1+\delta)^2 - 4r_s\delta}}{2\delta} - 1 & \text{otherwise.} \end{cases}$$

Since

$$\lim_{\delta \to -\infty} f(\delta) = k - 1 > 0 \text{ and } \lim_{\delta \to \infty} f(\delta) = -1 < 0,$$

$f(\delta) = 0$ has at least one root. Next we show that $f(\delta)$ is strictly decreasing: Consider $\delta \neq 0$, then

$$f'(\delta) = \sum_{s=1}^{k} \frac{-1 + \frac{1+\delta-2r_s\delta}{\sqrt{(1+\delta)^2 - 4r_s\delta}}}{2\delta^2}.$$

If $1+\delta-2r_s\delta \leq 0$, then of course $f'(\delta) < 0$. For the other case, first we use $-4r_s\delta^2 + 4r_s^2\delta^2 < 0$ to obtain

$$(1+\delta)^2 - 4r_s\delta(1+\delta) + 4r_s^2\delta^2 < (1+\delta)^2 - 4r_s\delta.$$

Since $1 + \delta - 2r_s\delta > 0$, taking the square root on both sides leads to $f'(\delta) < 0$.

Therefore,

$$f(\delta) = 0 \text{ has a unique solution at } \delta > 0 \ (< 0) \text{ if } \sum_{s=1}^{k} r_s - 1 > 0 \ (< 0).$$

Appendix F. Update θ for Models with Home-field Advantages or Ties

For the Home-field model, we use the "minorizing"-function approach in (Hunter, 2004):

$$\begin{aligned} & \text{Terms of } l(\mathbf{p}^t, \theta) \text{ related to } \theta \\ = \; & -\sum_{i=1}^{m} \big((\bar{r}_i + \tilde{r}'_i) \log \theta - (\bar{r}_i + \bar{r}'_i) \log(\theta q_i^+ + q_i^-) - (\tilde{r}_i + \tilde{r}'_i) \log(q_i^+ + \theta q_i^-) \big) \\ \leq \; & -\sum_{i=1}^{m} \Bigg((\bar{r}_i + \tilde{r}'_i) \log \theta - (\bar{r}_i + \bar{r}'_i) \big(-1 + \log(\theta^t q_i^+ + q_i^-) + \frac{\theta q_i^+ + q_i^-}{\theta^t q_i^+ + q_i^-} \big) \\ & \qquad -(\tilde{r}_i + \tilde{r}'_i) \big(-1 + \log(q_i^+ + \theta^t q_i^-) + \frac{q_i^+ + \theta q_i^-}{q_i^+ + \theta^t q_i^-} \big) \Bigg) \\ \equiv \; & Q(\theta). \end{aligned}$$

The inequality becomes equality if $\theta = \theta^t$. Thus, $Q'(\theta) = 0$ leads to (28) and $l(\mathbf{p}^t, \theta^{t+1}) < l(\mathbf{p}^t, \theta^t)$.

For the model which allows ties, we again define a minorizing function of θ.

$$\begin{aligned}
& \text{Terms of } l(\mathbf{p}^t, \theta) \text{ related to } \theta \\
= & -\sum_{i=1}^{m} \left(t_i \log(\theta^2 - 1) - (r_i + t_i) \log(q_i^+ + \theta q_i^-) - (r_i' + t_i) \log(\theta q_i^+ + q_i^-) \right) \\
\leq & -\sum_{i=1}^{m} \Bigg(t_i \log(\theta^2 - 1) - (r_i + t_i) \left(1 + \log(q_i^+ + \theta^t q_i^-) + \frac{q_i^+ + \theta q_i^-}{q_i^+ + \theta^t q_i^-} \right) \\
& \qquad -(r_i' + t_i) \left(1 + \log(\theta^t q_i^+ + q_i^-) + \frac{\theta q_i^+ + q_i^-}{\theta^t q_i^+ + q_i^-} \right) \Bigg) \\
\equiv & \; Q(\theta).
\end{aligned}$$

Then $Q'(\theta) = 0$ implies

$$\sum_{i=1}^{m} \left(\frac{2\theta t_i}{\theta^2 - 1} - \frac{(r_i + t_i) q_i^-}{q_i^+ + \theta^t q_i^-} - \frac{(r_i' + t_i) q_i^+}{\theta^t q_i^+ + q_i^-} \right) = 0$$

and hence θ^{t+1} is defined as in (31).