Pattern Recognition Letters 84 (2016) 99–106

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Bayes covariant multi-class classification☆

Ondrej Šuch[a,b,*], Santiago Barreda[c]

[a] Fakulta riadenia a informatiky, Žilinská Univerzita v Žiline, Univerzitná 8215/1, Žilina, 010 26, Slovakia
[b] Mathematical Institute of Slovak Academy of Sciences, Ďumbierska 1, Banská Bystrica, 974 11, Slovakia
[c] University of California, Davis, Linguistics Department, 469 Kerr Hall, One Shields Avenue, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 9 December 2015
Available online 26 August 2016

Keywords:
Multi-class classification
Bradley–Terry model
Bayes classifier
Combining binary classifiers
TIMIT
Vowel classification

ABSTRACT

We consider multi-class classification models built from complete sets of pairwise binary classifiers. The Bradley–Terry model is often used to estimate posterior distributions in this setting. We introduce the notion of Bayes covariance, which holds if the multi-class classifier respects multiplicative group action on class priors. As a consequence, a Bayes covariant method yields the same result whether new priors are considered before or after combination of the individual classifiers, which has several practical advantages for systems with feedback. In the paper, we construct a Bayes covariant combining method and compare it with previously published methods in both Monte Carlo simulations as well as on a practical speech frame recognition task.

1. Introduction

Statistical and machine-learning classification methods have found widespread applications in industry, as well as in scientific research. Successful applications include optical character recognition [1], speech recognition systems [2], automated medical diagnoses [3] and credit-risk scoring [4]. Although in some practical applications binary decisions may be sufficient (e.g. cancer/no cancer decision), most applications require correct classification among multiple classes.

Broadly speaking, multi-class classification will pose a more challenging problem than binary classification. One reason for this is that the set of boundaries among multiple classes may be more complex and thus may be harder to learn than the boundary between two classes. Another reason is that several powerful machine learning methods for classification of two classes have no direct analogues for multiple classes, making these methods inapplicable for those faced with a multi-class problem. Important examples of such methods include support vector machines [5,6] and Adaboost [7].

There are many ways to reduce the multi-class classification of K classes to binary classification subproblems. One common approach is one-vs-all classification when one trains K classifiers to distinguish each class from all of the rest [8]. Another common approach is all-vs-all when one trains $\binom{K}{2}$ pairwise classifiers [9]. Other approaches have been proposed based on error correcting coding theory [10–14] and on training statistical meta-classifiers [15].

In our work, we consider the question of combining the output of binary classifiers in an all-vs-all setting. Some reasons to consider this approach rather than the one-vs-all approach [8] include:

- larger number of parameters allow for more powerful models,
- simpler and faster training of individual classifiers compared to one-vs-all ([8, pp. 123–124]),
- when samples are densely packed in Euclidean space, the all-vs-all boundaries should be simpler, and thus easier to learn than one-vs-all boundaries; for example English vowels lie essentially in a 2-dimensional space [16],
- larger number of binary models allows for some tolerance of imprecision of individual classifiers ([17], [8, p. 102], [18]). Imprecise computation is typical for neuromorphic circuits for classification problems, which on the other hand are highly parallel and highly energy efficient [19,20].

Bayes theorem provides a rigorous foundation of classification. The theorem explains the crucial role played by class priors on the outcome of classification (cf. (2)). Usually, class priors are a fixed quantity during classification. However, in multi-tiered systems with feedback, one may desire to reevaluate evidence with different priors based on feedback from other tiers. For instance, a typical automated speech recognition system consists of three parts – an acoustic model, a lexicon and a language model [21].

☆ This paper has been recommended for acceptance by Maria De Marsico.

* Corresponding author at: University of Žilina, Žilina, Slovakia. Fax: +421 415134312.
E-mail addresses: ondrej.such@fri.uniza.sk, ondrej.such@gmail.com (O. Šuch).

http://dx.doi.org/10.1016/j.patrec.2016.08.014
0167-8655/

At first, one may classify a word-initial sound based on the prior for word-initial sounds from the acoustic model. As subsequent phonemes are recognized, the system may narrow down the list of possible words based on lexical and language constraints, yielding a different prior on the word-initial sound. Final classification of the word initial phoneme could then be derived from this latter prior.

The effect of varying class priors on posterior probabilities is predicted by Bayes theorem. In all-vs-all classification, this effect can be applied in two ways: either before, or after combining the results of individual pairwise classifiers. We propose to investigate Bayes covariant methods, which we define as those for which either combining method yields the same result.

2. The Bradley–Terry model

Consider the situation when each of a set of pairwise binary classifiers produces not only a $0-1$ decision, but actually estimates the two class conditional probability. Namely, each pairwise classifier trained to decide between classes C_i and C_j also estimates the posterior probability of belonging to a particular class, $r_{ij} = p(C_i|\mathbf{x}, C_i \vee C_j)$ given a vector of observed features $\mathbf{x}$. This additional information should allow the multi-class model to produce more accurate classification. Moreover, and independently of the potential for greater classification accuracy, in some applications it may be desirable to obtain an estimate of the posterior distribution $p(C_i \mid \mathbf{x})$. For example, there are many uses for posterior information, e.g. for loss minimization decisions, for subsequent processing by a temporal model like HMM, for explanation of perception experiments [22–24].

We will now describe the Bradley–Terry model [25] which is commonly used to combine the output of pairwise classifiers [17,18]. By Bayes theorem we know that the output of the classifier comparing C_i and C_j is:

$$r_{ij} = \frac{p(\mathbf{x} \mid C_i)p(C_i)}{p(\mathbf{x}, C_i \vee C_j)} = \frac{p(\mathbf{x} \mid C_i)p(C_i)}{p(\mathbf{x} \mid C_i)p(C_i) + p(\mathbf{x} \mid C_j)p(C_j)}, \tag{1}$$

and for the desired multi-class posterior one has:

$$p(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_i)p(C_i)}{\sum_j p(\mathbf{x} \mid C_j)p(C_j)}. \tag{2}$$

Setting $p_i = p(\mathbf{x} \mid C_i)p(C_i)$ we arrive at the Bradley–Terry system of equations:

$$r_{ij} = \frac{p_i}{p_i + p_j}, \tag{3}$$

whose solution gives the desired posterior distribution:

$$p(C_i \mid \mathbf{x}) = \frac{p_i}{\sum_j p_j}. \tag{4}$$

We will assume that $r_{ij} + r_{ji} = 1$ and moreover that $r_{ij} > 0$ for $i \neq j$. The equality is a very natural requirement on the binary classifiers. The requirement that r_{ij} is positive may simply be achieved by removing from consideration all categories C_i for which there is $j \neq i$ such that $r_{ij} = 0$. We shall call matrices $\mathbf{R} = (r_{ij})$ that satisfy these requirements *feasible*. Our goal is then to construct estimates $\hat{p}_i$ of p_i based on a feasible matrix $\mathbf{R}$. We will refer to algorithms yielding such estimates as *combining methods*. We are interested in algorithms that do not involve training in contrast to statistical combining methods [15].

3. Bayes covariant combining methods

As outlined in the previous section, the Bradley–Terry model explicitly depends on priors $\pi_i = p(C_i)$. Consider what happens to r_{ij} if the prior changes to $\pi'_i = q_i\pi_i$. On the one hand from (3) we have:

$$\frac{1}{r_{ij}} - 1 = \frac{p_j}{p_i} = \frac{p(\mathbf{x} \mid C_j)p(C_j)}{p(\mathbf{x} \mid C_i)p(C_i)}, \tag{5}$$

so that binary posteriors $\mathbf{R} = (r_{ij})$ should transform to $\mathbf{R}' = (r'_{ij})$ satisfying:

$$\frac{1}{r'_{ij}} - 1 = \frac{q_j}{q_i}\left(\frac{1}{r_{ij}} - 1\right). \tag{6}$$

When the dependence on $\mathbf{q}$ is important we will indicate it by right upper exponent $\mathbf{R}' = \mathbf{R}^{\mathbf{q}}$.

On the other hand, we may examine the effect of altering the prior probabilities on estimated solutions $\hat{p}_i$ produced by a combining method M which takes as input the matrix $\mathbf{R} = (r_{ij})$. Writing:

$$\hat{p}_i = M(\mathbf{R}) \tag{7}$$

$$\hat{p}'_i = M(\mathbf{R}') \tag{8}$$

we may expect from (2) that:

$$(\hat{p}'_1, \hat{p}'_2, \ldots) \propto (q_1\hat{p}_1, q_2\hat{p}_2, \ldots). \tag{9}$$

If this relationship holds for any feasible matrix $\mathbf{R}$ and any positive reweighing vector $\mathbf{q} = (q_1, \ldots, q_K)$ we say that the combining method M is *Bayes covariant*. In the next sections, we will constructively prove the existence of Bayes covariant combining methods.

4. A Bayes covariant combining method for three categories

In this section, we restrict ourselves to three-category problems ($K = 3$). Let us start by introducing the *3-symmetry* condition. Consider the situation when binary classifiers report a feasible matrix satisfying

$$r_{12} = r_{23} = r_{31}. \tag{10}$$

By our assumption also $r_{21} = r_{32} = r_{13} = 1 - r_{12}$. These data are completely symmetrical and thus it is natural to expect that a combining method M gives preference to no category. Formally, a combining method M satisfies 3-symmetry if and only if:

$$M\begin{pmatrix} \cdot & t & 1-t \\ 1-t & \cdot & t \\ t & 1-t & \cdot \end{pmatrix} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \quad \text{for any } t \in (0,1). \tag{11}$$

Theorem 1. *There exists a unique Bayes covariant combining method for $K = 3$ that has 3-symmetry.*

Proof. Let M be a Bayes covariant combining method with 3-symmetry. We will show that for any feasible matrix $\mathbf{R}$ there exists $\mathbf{q} > 0$ such that $\mathbf{R}^{\mathbf{q}} = (r'_{ij})$ has entries satisfying (10). If we set $\mathbf{q} = (1, A, B)$ then from (6) we have for $\mathbf{R}^{\mathbf{q}} = (r'_{ij})$:

$$A\left(\frac{1}{r_{12}} - 1\right) = \frac{1}{r'_{12}} - 1 \tag{12}$$

$$\frac{1}{B}\left(\frac{1}{r_{31}} - 1\right) = \frac{1}{r'_{31}} - 1 \tag{13}$$

$$\frac{B}{A}\left(\frac{1}{r_{23}} - 1\right) = \frac{1}{r'_{23}} - 1 \tag{14}$$