Neurocomputing 000 (2017) 1–11

Contents lists available at ScienceDirect

Neurocomputing

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

Binary classifiers ensemble based on Bregman divergence for multi-class classification

Takashi Takenouchi [a,b,*], Shin Ishii [c]

[a] Future University Hakodate, 116-2, Kamedanakano, Hakodate, Hokkaido 040–8655, Japan
[b] RIKEN Center for Advanced Intelligence Project, Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan
[c] Graduate School of Informatics, Kyoto University, Yoshidahonmachi 36-1, Sakyo-ku, Kyoto 606–8501, Japan

ARTICLE INFO

Article history:
Received 11 April 2016
Revised 29 June 2017
Accepted 1 August 2017
Available online xxx

Communicated by Deng Cheng

Keywords:
Ensemble learning
Multi-class classification
Bregman divergence
Information geometry

ABSTRACT

We propose a novel integration method of binary classifiers for multi-class classification. The proposed method is characterized as a minimization problem of a weighted mixture of Bregman divergence, and employs binary classifiers as class-dependent feature vector. We discuss the statistical properties of the proposed method and the relationship between the proposed method and existing multi-class classification methods, and reveal that many of the existing methods can be formulated as special cases of the proposed method. Small experiments show that the proposed method can effectively incorporate information of multiple binary classifiers into the multi-class classifier.

1. Introduction

Supervised classification constitutes one of the most major topics in machine learning area, and especially the multi-class classification problem with $G(\geq 3)$ classes is an important issue. While methods for the binary classification problem are well established, multi-class classification remains challenging, and methods to achieve it are still being developed. One major approach to multi-class classification directly constructs discriminant functions for multiple classes, for example, by estimating the conditional probability density for each class. So far, several multi-class extensions of SVM [5] and AdaBoost (AdaBoost-M2 and so on) [7] have been proposed. A recent remarkable development within this approach is the Deep learning-based method [15]. Another approach is decoding from an ensemble of classifiers in which the original multi-class classification is decomposed into multiple (typically binary) classification problems so that the multi-class discriminant function is constructed such to integrate the results of the constituentbinary classifiers. This approach is computationally feasible and a lot of studies like Hamming decoding [6], error correcting output coding (ECOC) decoder [26], loss-based decoding [1], multi-class SVM [31] and Bradley–Terry (BT) model-based decoders [12,27,33] have been proposed as instances of this approach. An important issue of that approach is how to decompose the original multi-class classification problem into an easy-to-solve form effectively, which is often terms as an encoding issue. For example, [29] considered a hierarchical structure associated with the decomposition and [17,18,32] incorporated information that is shared by binary classifiers into the resultant multi-class classifier, by means of simultaneously optimizing meta cost functions for integrating and training the binary classifiers. While theoretical analysis of the first approach can be straightforward [34], the cost function becomes usually complicated and is hard to optimize in the sense of computational cost especially when the number of classes, G, is very large. On the other hand, the second approach is easy to implement because fast and sophisticated package programs can be available as constituent binary classifiers, and its computational cost is just proportional to the number of binary classifiers. In this study, we develop a multi-class classification method belonging to the second category, that is, a novel integration method of constituent binary classifiers, which is characterized as the minimization of weighted sum of Bregman divergence.[1] The proposed method can encompass a lot of existing methods based on the en-

* Corresponding author at: Future University Hakodate, 116-2, Kamedanakano, Hakodate, Hokkaido 040–8655, Japan.
E-mail addresses: ttakashi@fun.ac.jp (T. Takenouchi), ishii@i.kyoto-u.ac.jp (S. Ishii).

[1] A short version of this article was presented as a conference paper [28]. While the integration method presented in the previous version used the Kullback–Leibler (KL) divergence, this paper extends the framework to include the general Bregman divergence and presents theoretical discussions and new experiments.

http://dx.doi.org/10.1016/j.neucom.2017.08.004
0925-2312/

semble approach as special cases, and can be interpreted as using class-dependent feature vector constructed from binary classifiers.

In Section 2, technical preliminaries are reviewed and the most important notion, mixture models, is introduced. Section 3 describes the novel integration method based on a mixture model employing the Bregman divergence. In Section 4, we show that the proposed method includes many of the existing multi-class classification methods as special cases. The statistical properties of the proposed method are discussed in Section 5. Section 6 examines the performance of the proposed method through three kinds of experiments. Section 7 concludes this study.

2. Settings and problem

Let $\boldsymbol{x}$ be an input and $y \in \mathcal{Y} = \{1, \ldots, G\}$ be its class label. The main purpose of multi-class classification ($G \geq 3$), $\boldsymbol{x} \mapsto y$, is to construct a discriminant function $\hat{y}(\boldsymbol{x})$ based on a given dataset of N pairs of input and its class label, $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$. Let us decompose the original G-class classification problem, $\boldsymbol{x} \mapsto y$, into J binary classification problems. Let $W \in \{+1, 0, -1\}^{J \times G}$ be a code word matrix, which is assumed to be given *a priori*, and $\boldsymbol{z}(y) = (z_1(y), \ldots, z_J(y))^T \in \{+1, 0, -1\}^J$ be the yth column of W, where T denotes a transpose. we assume that the vector $\boldsymbol{z}(y)$ equivalently represents the multi-class label y and is called the 'code word' of y. One of the simplest code word matrix for $G = 3$ is:

$$W = \begin{pmatrix} +1 & -1 & 0 \\ +1 & 0 & -1 \\ 0 & +1 & -1 \end{pmatrix}, \tag{1}$$

which is sometimes called 'one-versus-one' coding [23]. Each row of W defines a binary classification problem so that classes having code $+1$ are discriminated from classes having code -1. Classes having code 0 are not utilized for training the binary classifier. Let $f_j(\boldsymbol{x}) \in R$ be a discriminant function constructed by a specific binary classification algorithm such as AdaBoost and Support vector machine (SVM) for the jth binary classification problem, $\boldsymbol{f}(\boldsymbol{x})$ be a vector of $f_j(\boldsymbol{x})$, $\tilde{z}_j(\boldsymbol{x}) = sgn(f_j(\boldsymbol{x})) \in \{+1, -1\}$ be the sign of the discriminant function for the jth binary classification problem ($f_j(\boldsymbol{x})$), and $\tilde{\boldsymbol{z}}(\boldsymbol{x})$ be a vector of $\tilde{z}_j(\boldsymbol{x})$. In this study, we assume that the code word matrix W, the vector $\boldsymbol{f}(\boldsymbol{x})$ of the discriminant function or its sign $\tilde{\boldsymbol{z}}(\boldsymbol{x})$ are given, and we will focus on constructing a multi-class classifier $\hat{y}(\boldsymbol{x})$ for a new input $\boldsymbol{x}$ by integrating $f_j(\boldsymbol{x})$ or $\tilde{z}_j(\boldsymbol{x})$ $(j = 1, \ldots, J)$.

3. Related works

In this section, we describe two related works, Bregman divergence and a generalized mixture of positive measures with the Bregman divergence. The Bregman divergence represents the discrepancy between positive measures and the generalized mixture of positive measures is defined by a minimization problem of the Bregman divergence. We utilize the generalized mixture model for integration of binary classifiers.

3.1. Bregman divergence

We first assume the space of all positive finite measures on $\mathcal{Y}$

$$\mathcal{M} = \left\{ m(y) \;\middle|\; \sum_{y \in \mathcal{Y}} m(y) < \infty \right\}, \tag{2}$$

and its subspace consisting of all probability measures on $\mathcal{Y}$

$$\mathcal{P} = \left\{ m(y) \;\middle|\; \sum_{y \in \mathcal{Y}} m(y) = 1 \right\}. \tag{3}$$

The statistical version of the Bregman divergence between two conditional measures $\mu(y|\boldsymbol{x})$ and $\nu(y|\boldsymbol{x})$ on $\mathcal{M}$ [21] is defined as

$$D_U(\mu, \nu) = \int p(\boldsymbol{x}) \sum_{y \in \mathcal{Y}} \Big\{ U(\xi(\nu(y|\boldsymbol{x}))) - U(\xi(\mu(y|\boldsymbol{x}))) - \mu(y|\boldsymbol{x})(\xi(\nu(y|\boldsymbol{x})) - \xi(\mu(y|\boldsymbol{x}))) \Big\} d\boldsymbol{x}, \tag{4}$$

where $p(\boldsymbol{x})$ is the marginal distribution of $\boldsymbol{x}$ and $U(\zeta)$ is a strictly convex and monotonically increasing function on R, and $u(\zeta) = U'(\zeta)$ is its derivative. The function $\xi(\zeta) = u^{-1}(\zeta)$ is the inverse function of $u(\zeta)$. The Bregman divergence is a pseudo-distance for measuring the discrepancy between two conditional measures, and is 0 if and only if the two conditional measures are equivalent, $\mu = \nu$. Note that the Bregman divergence is not in general symmetric with respect to μ and ν, therefore, it is not a proper distance. One possible advantage of formulation (4) is that we can directly plug-in the empirical distribution into μ. The following are examples of the function U.

1. Exponential type: $U(z) = \exp(z)$, $u(z) = \exp(z)$, $\xi(z) = \log z$.
   The Bregman divergence with the exponential function is equivalent to the Kullback–Leibler (KL) divergence,
   $$KL(\mu, \nu) = \int p(\boldsymbol{x}) \sum_{y \in \mathcal{Y}} \mu(y|\boldsymbol{x}) \log \frac{\mu(y|\boldsymbol{x})}{\nu(y|\boldsymbol{x})} - \mu(y|\boldsymbol{x}) + \nu(y|\boldsymbol{x}) d\boldsymbol{x}. \tag{5}$$
2. β-divergence: $U(z, \beta) = \frac{1}{\beta+1}(1 + \beta z)^{\frac{\beta+1}{\beta}}$, $u(z, \beta) = (1 + \beta z)^{\frac{1}{\beta}}$, $\xi(z, \beta) = \frac{z^\beta - 1}{\beta}$.
   The Bregman divergence with the $U(z, \beta)$ above is called the β-divergence,
   $$D_\beta(\mu, \nu) = \int p(\boldsymbol{x}) \sum_{y \in \mathcal{Y}} \left\{ \frac{\nu(y|\boldsymbol{x})^{\beta+1} - \mu(y|\boldsymbol{x})^{\beta+1}}{\beta + 1} - \mu(y|\boldsymbol{x}) \frac{\nu(y|\boldsymbol{x})^\beta - \mu(y|\boldsymbol{x})^\beta}{\beta} \right\} d\boldsymbol{x}. \tag{6}$$
   Note that the β-divergence reduces to the KL divergence when $\beta \to 0$. The β-divergence is frequently used for robust inference [4,19], and is closely related to the Tsallis entropy [30].
3. η-divergence: $U(z, \eta) = (1 + \eta)\exp(z) - \eta z$, $u(z, \eta) = (1 + \eta)\exp(z) - \eta$, $\xi(z, \eta) = \log \frac{z + \eta}{1 + \eta}$.
   The Bregman divergence with the $U(z, \eta)$ defined above is called the η-divergence,
   $$D_\eta(\mu, \nu) = \int p(\boldsymbol{x}) \sum_{y \in \mathcal{Y}} \left\{ (\mu(y|\boldsymbol{x}) + \eta) \log \frac{\mu(y|\boldsymbol{x}) + \eta}{\nu(y|\boldsymbol{x}) + \eta} - \mu(y|\boldsymbol{x}) + \nu(y|\boldsymbol{x}) \right\} d\boldsymbol{x}. \tag{7}$$
   The η-divergence has also been used to robustify classification algorithms and is useful for probabilistic modeling of mislabeling [24,25].

3.2. Mixture of positive measures by the Bregman divergence

In the following two subsections, we review the mixture models of positive measures which are defined in terms of the Bregman divergence. Suppose that we have positive measure functions $q_j(y) \in \mathcal{M}$ $(j = 1, \ldots, J)$ and their weights w_j $(\sum_{j=1}^J w_j = 1, w_j \geq 0)$. Based on the Bregman divergence, we can define 'mixture models' for the given positive measure functions [10,20], that are useful in our multi-class classification context.

Please cite this article as: T. Takenouchi, S. Ishii, Binary classifiers ensemble based on Bregman divergence for multi-class classification, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.08.004