Neural Networks 75 (2016) 126–140

Contents lists available at ScienceDirect

Neural Networks

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

Cross-validation of matching correlation analysis by resampling matching weights

Hidetoshi Shimodaira

Division of Mathematical Science, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama-cho, Toyonaka, Osaka, Japan

ARTICLE INFO

Article history:
Received 29 March 2015
Received in revised form 23 September 2015
Accepted 11 December 2015
Available online 22 December 2015

Keywords:
Cross-validation
Resampling
Multiple domains
Canonical correlation analysis
Spectral graph embedding
Associative memory

ABSTRACT

The strength of association between a pair of data vectors is represented by a nonnegative real number, called matching weight. For dimensionality reduction, we consider a linear transformation of data vectors, and define a *matching error* as the weighted sum of squared distances between transformed vectors with respect to the matching weights. Given data vectors and matching weights, the optimal linear transformation minimizing the matching error is solved by the spectral graph embedding of Yan et al. (2007). This method is a generalization of the canonical correlation analysis, and will be called as matching correlation analysis (MCA). In this paper, we consider a novel sampling scheme where the observed matching weights are randomly sampled from underlying true matching weights with small probability, whereas the data vectors are treated as constants. We then investigate a cross-validation by resampling the matching weights. Our asymptotic theory shows that the cross-validation, if rescaled properly, computes an unbiased estimate of the matching error with respect to the true matching weights. Existing ideas of cross-validation for resampling data vectors, instead of resampling matching weights, are not applicable here. MCA can be used for data vectors from multiple domains with different dimensions via an embarrassingly simple idea of coding the data vectors. This method will be called as cross-domain matching correlation analysis (CDMCA), and an interesting connection to the classical associative memory model of neural networks is also discussed.

1. Introduction

We have N data vectors of P dimensions. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^P$ be the data vectors, and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T \in \mathbb{R}^{N\times P}$ be the data matrix. We also have *matching weights* between the data vectors. Let $w_{ij} = w_{ji} \geq 0$, $i, j = 1, \ldots, N$, be the matching weights, and $\boldsymbol{W} = (w_{ij}) \in \mathbb{R}^{N\times N}$ be the matching weight matrix. The matching weight w_{ij} represents the strength of association between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. For dimensionality reduction, we will consider a linear transformation from $\mathbb{R}^P$ to $\mathbb{R}^K$ for some $K \leq P$ as

$$\boldsymbol{y}_i = \boldsymbol{A}^T\boldsymbol{x}_i, \quad i = 1, \ldots, N,$$

or $\boldsymbol{Y} = \boldsymbol{XA}$, where $\boldsymbol{A} \in \mathbb{R}^{P\times K}$ is the linear transformation matrix, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N \in \mathbb{R}^K$ are the transformed vectors, and $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)^T \in \mathbb{R}^{N\times K}$ is the transformed matrix. Observing $\boldsymbol{X}$ and $\boldsymbol{W}$, we would like to find $\boldsymbol{A}$ that minimizes the *matching error*

$$\phi = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2$$

under some constraints. We expect that the distance between $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ will be small when w_{ij} is large, so that the locations of transformed vectors represent both the locations of the data vectors and the associations between data vectors. The optimization problem for finding $\boldsymbol{A}$ is solved by the spectral graph embedding for dimensionality reduction of Yan et al. (2007). Similarly to principal component analysis (PCA), the optimal solution is obtained as the eigenvectors of the largest K eigenvalues of some matrix computed from $\boldsymbol{X}$ and $\boldsymbol{W}$. In Section 3, this method will be formulated by specifying the constraints on the transformed vectors and also regularization terms for numerical stability. We will call the method as *matching correlation analysis* (MCA), since it is a generalization of the classical canonical correlation analysis (CCA) of Hotelling (1936). The matching error will be represented by *matching correlations* of transformed vectors, which correspond to the canonical correlations of CCA.

MCA will be called as *cross-domain matching correlation analysis* (CDMCA) when we have data vectors from multiple domains with different sample sizes and different dimensions. Let D be the number of domains, and $d = 1, \ldots, D$ denote each domain. For example, domain $d = 1$ may be for image feature vectors, and domain $d = 2$ may be for word vectors computed by word2vec

E-mail address: shimo@sigmath.es.osaka-u.ac.jp.

http://dx.doi.org/10.1016/j.neunet.2015.12.007
0893-6080/

(Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) from texts, where the matching weights between the two domains may represent tags of images in a large dataset, such as Flickr. From domain d, we get data vectors $\boldsymbol{x}_i^{(d)} \in \mathbb{R}^{p_d}$, $i = 1, \ldots, n_d$, where n_d is the number of data vectors, and p_d is the dimension of the data vector. Typically, p_d is hundreds, and n_d is thousands to millions. We would like to retrieve relevant words from an image query, and alternatively retrieve images from a word query. Given matching weights across/within domains, we attempt to find linear transformations of data vectors from multiple domains to a "common space" of lower dimensionality so that the distances between transformed vectors well represent the matching weights. This problem is solved by an embarrassingly simple idea of coding the data vectors, which is similar to that of Daumé III (2007). Each data vector from domain d is represented by an augmented data vector $\boldsymbol{x}_i$ of dimension $P = \sum_{d=1}^{D} p_d$, where only p_d dimensions are for the original data vector and the rest of $P - p_d$ dimensions are padded by zeros. In the case of $D = 2$ with $p_1 = 2$, $p_2 = 3$, say, a data vector $(1, 2)^T$ of domain 1 is represented by $(1, 2, 0, 0, 0)^T$, and $(3, 4, 5)^T$ of domain 2 is represented by $(0, 0, 3, 4, 5)^T$. The number of total augmented data vectors is $N = \sum_{d=1}^{D} n_d$. Note that the above mentioned "embarrassingly simple coding" is not actually implemented by padding zeros in computer software; only the nonzero elements are stored in memory, and CDMCA is in fact implemented very efficiently for sparse $\boldsymbol{W}$. CDMCA is illustrated in a numerical example of Section 2. CDMCA is further explained in Appendix A.1, and an interesting connection to the classical associative memory model of neural networks (Kohonen, 1972; Nakano, 1972) is also discussed in Appendix A.2.

CDMCA is solved by applying the single-domain version of MCA described in Section 3 to the augmented data vectors, and thus we only discuss the single-domain version in this paper. This formulation of CDMCA includes a wide class of problems of multivariate analysis, and similar approaches are very popular recently in pattern recognition and vision (Correa, Eichele, Adalı, Li, & Calhoun, 2010; Gong, Ke, Isard, & Lazebnik, 2014; Kan, Shan, Zhang, Lao, & Chen, 2012; Shi, Liu, Fan, & Yu, 2013; Wang, He, Wang, Wang, & Tan, 2013; Yuan & Sun, 2014; Yuan, Sun, Zhou, & Xia, 2011). CDMCA is equivalent to the method of Nori, Bollegala, and Kashima (2012) for multinomial relation prediction if the matching weights are defined by cross-products of the binary matrices representing relations between objects and instances. CDMCA is also found in Huang, Shan, Zhang, Lao, and Chen (2013) for the case of $D = 2$. CDMCA reduces to the multi-set canonical correlation analysis (MCCA) (Kettenring, 1971; Takane, Hwang, & Abdi, 2008; Tenenhaus & Tenenhaus, 2011) when $n_1 = \cdots = n_D$ with cross-domain matching weight matrices being proportional to the identity matrix. It becomes the classical CCA by further letting $D = 2$, or it becomes PCA by letting $p_1 = p_2 = \cdots = p_D = 1$.

In this paper, we discuss a cross-validation method for computing the matching error of MCA. In Section 4, we will define two types of matching errors, i.e., fitting error and true error, and introduce cross-validation (cv) error for estimating the true error. In order to argue distributional properties of MCA, we consider the following sampling scheme. First, the data vectors are treated as constants. Similarly to the explanatory variables in regression analysis, we perform conditional inference given data matrix $\boldsymbol{X}$, although we do not avoid assuming that $\boldsymbol{x}_i$'s are sampled from some probability distribution. Second, the matching weights w_{ij} are randomly sampled from underlying true matching weights $\bar{w}_{ij}$ with small probability $\epsilon > 0$. The value of ϵ is unknown and it should not be used in our inference. Let $z_{ij} = z_{ji} \in \{0, 1\}$, $i, j = 1, \ldots, N$, be samples from Bernoulli trial with success probability ϵ, where the number of independent elements is $N(N+1)/2$ due to the symmetry. Then the observed matching weights are defined as

$$w_{ij} = z_{ij}\bar{w}_{ij}, \qquad P(z_{ij} = 1) = \epsilon. \tag{1}$$

The true matching weight matrix $\bar{\boldsymbol{W}} = (\bar{w}_{ij}) \in \mathbb{R}^{N\times N}$ is treated as an unknown constant matrix with elements $\bar{w}_{ij} = \bar{w}_{ji} \geq 0$. This setting will be appropriate for a large-scale data, such as those obtained automatically from the web, where only a small portion $\boldsymbol{W}$ of the true association $\bar{\boldsymbol{W}}$ may be obtained as our knowledge.

In Section 4.2, we will consider a resampling scheme corresponding to (1). For the cross-validation, we resample $\boldsymbol{W}^*$ from $\boldsymbol{W}$ with small probability $\kappa > 0$, whereas $\boldsymbol{X}$ is left untouched. Our sampling/resampling scheme is very unique in the sense that the source of randomness is $\boldsymbol{W}$ instead of $\boldsymbol{X}$, and existing results of cross-validation for resampling from $\boldsymbol{X}$ such as Stone (1977) and Golub, Heath, and Wahba (1979) are not applicable here. The traditional method of resampling data vectors is discussed in Section 4.3.

The true error is defined with respect to the unknown $\bar{\boldsymbol{W}}$, and the fitting error is defined with respect to the observed $\boldsymbol{W}$. We would like to look at the true error for finding appropriate values of the regularization terms (regularization parameters are generally denoted as γ throughout) and the dimension K of the transformed vectors. However, the true error is unavailable, and the fitting error is biased for estimating the true error. The main thrust of this paper is to show asymptotically that the cv error, if rescaled properly, is an unbiased estimator of the true error. The value of ϵ is unnecessary for computing the cv error, but $\boldsymbol{W}$ should be a sparse matrix. The unbiasedness of the cv error is illustrated by a simulation study in Section 5, and it is shown theoretically by the asymptotic theory of $N \to \infty$ in Section 6.

2. Illustrative example

Let us see an example of CDMCA applied to the MNIST database of handwritten digits (see Appendix B.1 for the experimental details). The number of domains is $D = 3$ with the number of vectors $n_1 = 60{,}000$, $n_2 = 10$, $n_3 = 3$, and dimensions $p_1 = 2784$, $p_2 = 100$, $p_3 = 50$. The handwritten digit images are stored in domain $d = 1$, while domain $d = 2$ is for the digit labels "zero", "one", …, "nine", and domain $d = 3$ is for attribute labels "even", "odd", "prime". This CDMCA is also interpreted as MCA with $N = 60{,}013$ and $P = 2934$.

The elements of $\bar{\boldsymbol{W}}$ are simply the indicator variables (called dummy variables in statistics) of image labels. Instead of working on $\bar{\boldsymbol{W}}$, here we made $\boldsymbol{W}$ by sampling 20% of the elements from $\bar{\boldsymbol{W}}$ for illustrating how CDMCA works. The optimal $\boldsymbol{A}$ is computed from $\boldsymbol{W}$ using the method described in Section 3.3 with regularization parameter $\gamma_M = 0.1$. The data matrix $\boldsymbol{X}$ is centered, and the transformed matrix $\boldsymbol{Y}$ is rescaled. The first and second elements of $\boldsymbol{y}_i$, namely, (y_{i1}, y_{i2}), $i = 1, \ldots, N$, are shown in Fig. 1. For the computation of $\boldsymbol{A}$, we do not have to specify the value of K in advance. Similar to PCA, we first solve the optimal $\boldsymbol{A} = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^P) \in \mathbb{R}^{P\times P}$ for the case of $K = P$, then take the first K columns to get the optimal $\boldsymbol{A} = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^K) \in \mathbb{R}^{P\times K}$ for any $K \leq P$. We observe that images and labels are placed in the common space so that they represent both $\boldsymbol{X}$ and $\boldsymbol{W}$. Given a digit image, we may find the nearest digit label or attribute label to tell what the image represents.

The optimal $\boldsymbol{A}$ of $K = 9$ is then computed for several γ_M values. For each $\boldsymbol{A}$, the 10 000 images of test dataset are projected to the common space and the digit labels and attribute labels are predicted. We observe in Fig. 2(a) that the classification errors become small when the regularization parameter is around $\gamma_M = 0.1$. Since $\boldsymbol{x}_i$ does not contribute to $\boldsymbol{A}$ if $\sum_{j=1}^{N} w_{ij} = 0$, these error rates are computed using only 20% of $\boldsymbol{X}$; they improve to 0.0359 ($d = 2$)