Information Fusion 27 (2016) 228–239

Contents lists available at ScienceDirect

Information Fusion

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

Kernel combination via debiased object correspondence analysis

David Windridge [a,b,*], Fei Yan [c]

[a] Department of Computer Science, School of Science and Technology, Middlesex University, The Burroughs, London NW4 4BT, UK
[b] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey GU27XH, UK[1]
[c] Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey GU2 7XH, UK

ARTICLE INFO

Article history:
Received 12 June 2014
Received in revised form 5 January 2015
Accepted 22 February 2015
Available online 11 March 2015

Keywords:
Classifier combination
Support vector machines
Kernel methods
Tomography

ABSTRACT

This paper addresses the problem of combining multi-modal kernels in situations in which object correspondence information is unavailable between modalities, for instance, where missing feature values exist, or when using proprietary databases in multi-modal biometrics. The method thus seeks to recover inter-modality kernel information so as to enable classifiers to be built within a composite embedding space. This is achieved through a principled group-wise identification of objects within differing modal kernel matrices in order to form a composite kernel matrix that retains the full freedom of linear kernel combination existing in multiple kernel learning. The underlying principle is derived from the notion of tomographic reconstruction, which has been applied successfully in conventional pattern recognition.

In setting out this method, we aim to improve upon object-correspondence insensitive methods, such as kernel matrix combination via the Cartesian product of object sets to which the method defaults in the case of no discovered pairwise object identifications. We benchmark the method against the augmented kernel method, an order-insensitive approach derived from the direct sum of constituent kernel matrices, and also against straightforward additive kernel combination where the correspondence information is given *a priori*. We find that the proposed method gives rise to substantial performance improvements.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

The problem of multiple kernel learning (MKL) was identified by Lanckriet et al. [1] and is now well established within the literature [2–11]. It builds on the widespread adoption of kernel-based methods within machine learning for a variety of tasks, in particular regression and classification [12,13]. The latter category includes state-of-the-art methods such as support vector machines (SVMs) [14,12] and kernel Fisher discriminant analysis (kernel FDA) [15,16].

Kernel methods have in common that they map observations into an inner product space, provided that they fulfil the Mercer conditions. A wide choice of kernels is typically available for any given learning problem; each of these kernels can be seen as capturing a different aspect of the data. In classification problems, arbitrarily morphologically-complex (*i.e.*, non-linear) decision boundaries may be obtained within a linear input space via the choice of kernel. Early work on learning the kernel includes [17], where kernel parameters are optimized by minimizing estimates of the generalization error of SVMs, and [18], where the complexity of learning the kernel matrix for SVM classification is analyzed.

Multiple kernel learning seeks to learn an appropriate linear combination of such base kernels, linear combination being chosen because this crucially retains the Mercer properties. Lanckriet et al.'s formulation [1] utilizes linear combination of M $m \times m$ training kernel matrices K_k, $k = 1,\ldots,M$ and m class labels $y_i \in \{1,-1\}$, $i = 1,\ldots,m$, with m the number of training samples, this being equivalent to forming the Cartesian product of the associated feature spaces. The goal of MKL is then to optimize the 'scaling factors' of the feature spaces with respect to the classification. Other MKL formulations address tractability issues when m is large. These include, e.g., the semi-infinite linear programming (SILP) formulation of [3], and the reduced gradient descent algorithm of [6]. The ℓ_1 regularization in [1] can also be generalized to an ℓ_p $(p > 1)$ norm [19] to avoid solution sparsity if required. Other variants of MKL approaches include, to name a few, hyperkernels [20], information theoretic MKL [21], multiple kernel FDA [22,23], multiclass MKL [4], multilabel MKL [24] and nonlinear MKL [25].

A key distinction that may be made between multiple kernel methods is whether they implicitly require object correspondence

[1] Visiting academic.

* Corresponding author at: Department of Computer Science, School of Science and Technology, Middlesex University, The Burroughs, London NW4 4BT, UK. Tel.: +44 (0)1483 686048; fax: +44 (0)1483 686031.
E-mail address: d.windridge@mdx.ac.uk (D. Windridge).

http://dx.doi.org/10.1016/j.inffus.2015.02.002
1566-2535/© 2015 Elsevier B.V. All rights reserved.

information; additive kernel combination such as the method of Lanckriet et al. assumes that this information is present. Thus, the ordering of the objects defining the K_k is assumed to be the same across all modalities. However, methods do exist that are not dependent on this correspondence, the principle such method being *augmented kernel combination* [26]. In augmented kernel combination, the direct sum of kernel matrices is formed, resulting in a block-diagonal kernel matrix (i.e. so that all of the constituent kernel matrices are embedded along the diagonal of the resultant matrix, with all inter-kernel values set to a value of zero); [26] compares the geometric interpretation of linear combination and augmented kernel combination. It is shown in [27] that augmented kernel combination is closely related to classifier fusion.

In general, the problem domain will determine whether object correspondence information is available. For instance, it is not uncommon in multi-modal biometrics to obtain distinct sets of exemplar subjects for each individual biometric measurement (*e.g.*, iris scans, finger prints, photographic images), particularly when employing separate commercial sources [28]. In this case, we would wish to utilize the information collectively contained within each data set for a given test subject, but would lack object correspondences in the collective set of multi-modal data sets. In other words, we have object correspondence in the test set but not the training set. The augmented kernel approach to classification of individual test subjects in this case would be to build a composite kernel matrix via the direct sum of kernel matrices associated with each modality and then utilize this, in combination with a corresponding vector of class labels, for classifier training. (The direct sum kernel matrix is order-insensitive with regard to the training objects within individual modalities provided that the class label vector is correspondingly permuted.)

However, the argument of this paper is that such methods, by omitting the possibility of re-deriving correspondence information, potentially overlook important classification information. To address this, we propose a kernel-based adaptation of a method developed for standard non-kernelized pattern recognition that is capable of bringing about this correspondence.[2] The resulting method for multiple kernel learning gives rise to a kernel matrix that defines an appropriate composite embedding space that, as nearly as possible, approximates the kernel matrix that would exist if all object correspondence information were available. It does so by removing the biasing factors associated with linear methods of kernel combination.

1.1. Linear combination bias in non-kernelized pattern recognition

In conventional (*i.e.*, non-kernelized) pattern recognition, it may be demonstrated [29,30] that linear classifier combination methods impose a bias on the composite decision space formed by decision combination[3] (by "decision space" we here mean the space in which the decision boundary is formed). This bias comes about via the limitations of linear combination in dealing with correlated information in the marginal classifiers (i.e. the feature-selected classifiers constituting the combination), and prevents the optimal decision boundary being constructed, leading to suboptimal overall performance. We thus consider the classifiers within a combination as representing, to some degree of approximation, the marginal distributions of the composite pattern space in which the decision boundary is formed (see Section 2.1 for a pictorial example of this process; in the remainder of the Introduction we give a qualitative account).

This biasing behavior occurs, for instance, when feature selection is applied to an input space of arbitrary dimensionality, S, such that a set of classifiers (indexed by $i \in I$) become associated with non-coincident (*i.e.*, non-overlapping) feature sets that collectively span S (or a subset of it). In such cases, classifier combination effectively acts to combine, in the original input space, the set of orthogonal marginals distributions that are implicitly modeled within the individual classifiers, i (modeling need not be exact, e.g. in the case of discriminative classifiers; see Section 2.1 for an example with artificial neural networks).

A similar situation exists in multi-modal fusion problems, where modalities may equally be regarded as the features of some composite decision space, allocated to specific classifiers associated with the modalities. It was the effort of [29] to demonstrate that this bias is specifically a form of *sampling bias*. The bias attributable to linear combination methods within the composite space is thus due to the mismatch of the very low number of angular samples of the composite decision space (equivalent in magnitude to $|I|$) created by the orthogonal 'marginal' distributions of the feature-selection process in comparison to their linear sampling rate. (The linear sampling rate equates to the total number of distinguishable input vectors.)[4] However, to fully represent arbitrary distributions in the composite space, angular and linear sampling would have to be of the same order (the orthogonal nature of this angular sampling is depicted in Section 2.1).

This mismatch between angular and linear sampling of the composite decision space suggests an analogy with tomography theory, for which the component classifiers of the combination essentially represent Radon-projections (linear integrals) of the composite decision space. Linear combination then acts as the inverse operation to Radon-projection, *i.e.*, *back projection* (essentially a summation over the Radon Projections that intersect at the point of reconstruction). However, in tomography theory back-projection only recovers a *biased* simulacra of the original unprojected composite space (the outcome of back-projection being the original distribution in the space *convolved* with an artefact defined by the angular frequency of the Radon sampling). The process of tomography is thus concerned with the pre- or post-combination filtration of this artefact in order to remove the sampling bias.

Similarly, this bias is represented within tomographic classifier combination theory as a convolution of the true underlying distribution of pattern vectors (denoted F_{true}) in the decision space with an artefact (denoted F_{samp}) deriving from the sampling (F_{true} and F_{samp} are thus density distributions defined over the entirety of S). The 'recovered' density distribution induced by classifier combination is thus $F_{\text{comb}} = F_{\text{true}} \bigstar F_{\text{samp}}$, with $\bigstar$ the convolution operation. F_{samp} is thus defined in the composite space by the response of an origin-centered Dirac delta function, firstly to representation as a series of individual Dirac delta functions in the marginal spaces associated with each classifier, and secondly to the action of the combination rule that reconstructs an 'image', F_{comb}, of the original Dirac delta function within the composite space. That is, F_{samp} is what is obtained if one were to take a single pattern vector from the true underlying distribution of pattern vectors in the decision space, represent it within the individual classifiers via feature-selection, and then 're-project' it back onto in the decision space by applying the combination rule. The resulting entity formalizes the systematic convolutional 'bias' introduced by the

[2] Thus, our method is an MKL method to the extent that it proposes a linear sum of kernels to be optimized. However, the method of generating these kernels is by no means linear.

[3] This applies in situations in which it can be reasonably assumed that there exists no *a priori* restriction on density distributions in the decision space, for instance, prior knowledge of feature independence.

[4] Note this only represents *combination* bias; classifier bias also contributes. *cf.* [30] for a fuller discussion of the bias/variance breakdown under this paradigm. See also [31–33] for a general discussion of bias–variance–covariance decomposition in classifier ensembles.

Download English Version:

https://daneshyari.com/en/article/528238

Download Persian Version:

https://daneshyari.com/article/528238

Daneshyari.com