Pattern Recognition Letters 43 (2014) 127–135

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Adaptive and Weighted Collaborative Representations for image classification

Radu Timofte [a,*], Luc Van Gool [a,b]

[a] VISICS, ESAT-PSI/iMinds, University of Leuven, Belgium
[b] Computer Vision Lab, D-ITET, ETH Zurich, Switzerland

ARTICLE INFO

Article history:
Available online 29 August 2013

Communicated by A. Del Bimbo

Keywords:
Ridge Regression
Collaborative Representation
Sparse Representation
Classification
Kernel

ABSTRACT

Recently, Zhang et al. (2011) proposed a classifier based on Collaborative Representations (CR) with Regularized Least Squares (CRC-RLS) for image face recognition. CRC-RLS can replace Sparse Representation (SR) based Classification (SRC) as a simple and fast alternative. With SR resulting from an l_1-Regularized Least Squares decomposition, CR starts from an l_2-Regularized Least Squares formulation. Moreover, it has an algebraic solution.

We extend CRC-RLS to the case where the samples or features are weighted. Particularly, we consider weights based on the classification confidence for samples and the variance of feature channels. The Weighted Collaborative Representation Classifier (WCRC) improves the classification performance over that of the original formulation, while keeping the simplicity and the speed of the original CRC-RLS formulation. Moreover we investigate into query-adaptive WCRC formulations and kernelized extensions that show further performance improvements but come at the expense of increased computation time.

1. Introduction

For face recognition, Wright et al. (2009) pointed out the effectiveness of representing newly observed faces as linear combinations of previously observed ones. Usually, the problem is formulated as one of finding the coefficients that minimize the residual between a new sample and its linear reconstruction. The residual is commonly measured as the least squares difference, which allows for an algebraic solution. When all previous faces or samples contribute to the optimal linear combination, one has a so-called Collaborative Representation (CR) (Zhang et al., 2011). Aside from the basic least squares criterion for the creation of such optima, other constraints have been considered as well. For stabilizing the coefficients of the least squares representation, one could add a term that tries to minimize the l_2-norm of the coefficients, while still admitting an algebraic solution. Enforcing sparsity on the solution leads to an l_0-regularization, i.e. an l_1-Regularized Least Squares problem in practice, known as *lasso* (Tibshirani, 1996). Such adaptations yields a Sparse Representation (SR) (Wright et al., 2009), a key element in compressed sensing. Indeed, most signals admit a decomposition over a reduced set of signals from the same class (Bruckstein et al., 2009). Unfortunately, there no longer exists a known algebraic solution in that case. A combined l_1/l_2 regularization renders the coefficients more robust and enforces group sparsity (Zou and Hastie, 2005).

The resulting coefficients carry a meaning in that they reflect the importance of each sample. The coefficients and resulting residuals are used in the classification. The newly incoming sample or 'query' is assigned to the class that has the minimum residual error or the largest sum of coefficient magnitudes.

Here, we investigate the influence of weighting the samples and features in the aforementioned representations, in particular in the l_2-Regularized Least Squares formulations with algebraic solutions. In real classification tasks the training samples are not equally discriminative. Moreover, the coefficients of some samples correlate more closely with the correct prediction by the classifier. The same holds for the feature channels, as some are more informative than others for the classification process. One may also let depend the weights on the particular query that is to be classified. We extend our previous work (Timofte and Van Gool, 2012b) by addressing such adaptive weightings and by investigating the corresponding kernel extensions, as a means for further gains in classification performance.

The remainder is organized as follows. Section 2 briefly reviews the least squares based formulations. Section 3 presents weighted variants, Section 4 adapts to query-specific weighting, and Section 5 discusses the kernel trick. In Section 6 we describe the classifiers based on these representations. Section 7 shows experimental results, while Section 8 concludes the paper.

* Corresponding author. Tel.: +32 16321718, mobile: +32 494358850.
E-mail addresses: radu.timofte@esat.kuleuven.be (R. Timofte), vangool@vision.ee.ethz.ch (L. Van Gool).
URL: http://homes.esat.kuleuven.be/~rtimofte/ (R. Timofte).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.08.010

Notations and assumptions. The training data is stacked columnwise forming a matrix $X = [x_1, x_2, \ldots, x_m] \in \mathbb{R}^{n\times m}$ with $m \in \mathbb{N}$ n-dimensional training samples. $\beta \in \mathbb{R}^m$ denotes the vector of coefficients from the linear representation of a query sample $y \in \mathbb{R}^n$ over the (training) samples (X). We assume each sample to be zero mean and to have unit length (the l_2-norm is 1). By $\|x\|_p$ we denote the l_p-norm of a vector x, $\|x\|$ is the Euclidean norm (or l_2-norm), X^T is the transpose of matrix X, and X^{-1} the inverse. $diag(x)$ is the diagonal matrix with the vector x on the diagonal. I is the identity matrix.

2. Least Squares formulations

We shortly review three of the best-known Regularized Least Squares formulations.

The Ordinary Least Squares (OLS) solves:

$$\hat{\beta}_{OLS} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|^2 \tag{1}$$

If $(X^TX)^{-1}$ exists we have as algebraic solution:

$$\hat{\beta}_{OLS} = (X^TX)^{-1}X^Ty \tag{2}$$

The Collaborative Representation with Regularized Least Squares (CR) (Zhang et al., 2011), solves:

$$\hat{\beta}_{CR} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|^2 + \lambda_{CR}\|\beta\|^2 \tag{3}$$

where $\lambda_{CR} \in \mathbb{R}$ is a regulatory parameter. The algebraic solution becomes:

$$\hat{\beta}_{CR} = (X^TX + \lambda_{CR}I)^{-1}X^Ty \tag{4}$$

By adding the l_2 regularization we cope with the case where X^TX is singular and, moreover, we stabilize/robustify the solution and make it less noise dependent.

The Sparse Representation (SR) (Wright et al., 2009) is obtained by enforcing sparsity by means of l_1-regularization instead of l_2 as for CR:

$$\hat{\beta}_{SR} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|^2 + \lambda_{SR}\|\beta\|_1 \tag{5}$$

where $\lambda_{SR} \in \mathbb{R}$ is the Lagrangian regulatory parameter. For this problem, also known as *lasso* (Tibshirani, 1996), one does not have an algebraic solution, but efficient optimization solvers are available, like LARS (Efron et al., 2004), Feature Sign (Lee et al., 2006) or L1LS (Kim et al., 2007).

By combining sparsity and robustness by means of joint l_1 and l_2 regularization we pursue group sparsity:

$$\hat{\beta}_{EN} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1 \tag{6}$$

a problem known as (naive) Elastic Net (EN) (Zou and Hastie, 2005), where $\lambda_{1,2}$ are regulatory scalar parameters. $(1+\lambda_2)\hat{\beta}_{EN}$ gives the compensated EN solution.

3. Weighted representations

In this section we review weighted extensions of the aforementioned least squares formulations.

Generalized Least Squares (GLS) generalizes the Ordinary Least Squares (OLS) for cases with unequal variances or correlations between the observations. If $y = X\beta_{GLS} + \epsilon$ with zero mean residuals, i.e. $E[\epsilon|X] = 0$, and their variance is $Var[\epsilon|X] = \Omega$, GLS minimizes their squared Mahalanobis length to estimate β_{GLS}:

$$\hat{\beta}_{GLS} = \underset{\beta}{\operatorname{argmin}} (y - X\beta)^T\Omega^{-1}(y - X\beta) \tag{7}$$

$$\hat{\beta}_{GLS} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y \tag{8}$$

This simplifies to Weighted Least Squares (WLS) in case Ω is diagonal.

Ridge Regression (RR), also known as Tikhonov regularization (Tikhonov and Arsenin, 1977), solves:

$$\hat{\beta}_{RR} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|^2 + \|\Gamma_{RR}\beta\|^2$$

$$\hat{\beta}_{RR} = (X^TX + \Gamma_{RR}^T\Gamma_{RR})^{-1}X^Ty \tag{9}$$

where to alleviate ill-posed problems, $\Gamma_{RR} \in \mathbb{R}^{m\times m}$ is suitably chosen. $\Gamma_{RR} \in \mathbb{R}^{m\times m}$ is the Tikhonov matrix and enables to weight samples differently. RR simplifies to OLS or CR if Γ_{RR} is null or a scaled identity matrix, respectively.

A Generalized Weighted Collaborative Representation (WCR) can have the following formulation:

$$\hat{\beta}_{WCR} = \underset{\beta}{\operatorname{argmin}} (y - X\beta)^T\Omega_{WCR}^{-1}(y - X\beta) + \|\Gamma_{WCR}\beta\|^2 \tag{10}$$

$$\hat{\beta}_{WCR} = (X^T\Omega_{WCR}^{-1}X + \Gamma_{WCR}^T\Gamma_{WCR})^{-1}X^T\Omega_{WCR}^{-1}y \tag{11}$$

Similar to GLS, Ω_{WCR} gives the importance of each dimension, and similar to RR, Γ_{WCR} modulates the importance of each sample in the solution.

Adding sparsity regularization (l_1-norm) to WCR leads to a Generalized Weighted Elastic Net (WEN) formulation:

$$\hat{\beta}_{WEN} = \underset{\beta}{\operatorname{argmin}} \quad \{(y - X\beta)^T\Omega_{WEN}^{-1}(y - X\beta) + \|\Gamma_{WEN}\beta\|^2 + \|\Lambda_{WEN}|\beta|\|_1\} \tag{12}$$

where $\Lambda_{WEN} \in \mathbb{R}^{m\times m}$ expresses the importance of each sample for the l_1 term. By adding the sparsity regularization we lose the advantage of having a clean algebraic solution.

Other recent l_1-regularized methods weight the coefficients in relation to the covariances of the training samples, as in Weight Fused Lasso (Daye and Jeng, 2009) or Weight Fused Elastic Net (Fu and Xu, 2012). Other pairwise constraints are used in Group Lasso (Yuan and Lin, 2006), Pairwise Elastic Net (Lorbert et al., 2010), or Trace Lasso (Grave et al., 2011).

4. Adaptive weighted representations

In the previous section we formulated weighted representations. We can make a distinction between the way the weights are set in the representation formulation:

(i) independent of the query,
(ii) adaptive to (dependent on) the query.

Independence of the query is the default case for weighted representations. It means that the weights are learned (estimated) from the training data or set without considering the specificity of an arbitrary query. The dependent or adaptive approach uses the specificity of the input query in the computation of the weights, and thus adapts the weighted representation to the nature of the input.

When the weighting matrix (or residual variance in WLS) Ω is not directly known, it can be estimated, as in Feasible Generalized Least Squares (FGLS) (Little and Rubin, 2002), adapted to the query. Here we consider the case of GLS. First, we can use OLS and obtain the residuals u, and take for Ω the diagonal matrix of squared residuals, $diag(\hat{u}_{OLS})^2$, and estimate $\hat{\beta}_{FGLS_1}$:

$$\hat{u}_{OLS} = y - X\hat{\beta}_{OLS}, \quad \Omega_{OLS} = diag(\hat{u}_{OLS})^2 \tag{13}$$