Neurocomputing 148 (2015) 310–317

Contents lists available at ScienceDirect

Neurocomputing

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

Large scale multi-class classification with truncated nuclear norm regularization

Yao Hu*, Zhongming Jin, Yi Shi, Debing Zhang, Deng Cai, Xiaofei He

State Key Laboratory of CAD&CG, Zhejiang University, No. 388 Yu Hang Tang Road, Hangzhou 310058, China

ARTICLE INFO

Article history:
Received 27 February 2014
Received in revised form
28 June 2014
Accepted 28 June 2014
Communicated by X. Gao
Available online 7 July 2014

Keywords:
Truncated nuclear norm
Coordinate descent algorithm
Multi-class classification

ABSTRACT

In this paper, we consider the problem of multi-class image classification when the classes behaviour has a low rank structure. That is, classes can be embedded into a low dimensional space. Traditional multi-class classification algorithms usually use nuclear norm to approximate the rank of the weight matrix. Considering the limited ability of the nuclear norm for the accurate approximation, we propose a new scalable large scale multi-class classification algorithm by using the recently proposed *truncated nuclear norm* as a better surrogate of the rank operator of matrices along with multinomial logisitic loss. To solve the non-convex and non-smooth optimization problem, we further develop an efficient iterative procedure. In each iteration, by lifting the non-smooth convex subproblem into an infinite dimensional ℓ_1 norm regularized problem, a simple and efficient accelerated coordinate descent algorithm is applied to find the optimal solution. We conduct a series of evaluations on several public large scale image datasets, where the experimental results show the encouraging improvement of classification accuracy of the proposed algorithm in comparison with the state-of-the-art multi-class classification algorithms.

1. Introduction

Recently there are more and more extensive research efforts on developing the scaling-up image systems with larger labeled image datasets such as ImageNet and Tiny. At the same time, websites, such as Flickr and Facebook, contain thousands of groups. Some groups, such as vehicles and flowers, consist of millions of images, which have been used for learning object classifiers [23].

In this situation, it is necessary to build general purpose object recognizers that are able to recognize many different classes of objects, which can be very useful for image/video retrieval and other vision applications [9,30]. Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ be a set of n training samples labeled into k classes, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathcal{Y} = \{1, 2, \ldots, k\}$, $i = 1, \ldots, n$. The motivation of the multi-class classification is to predict the class label of a new query $\mathbf{x}$ by learning a linear classifier $\hat{y} = \arg\max_{\ell \in \mathcal{Y}} \mathbf{w}_\ell^T \mathbf{x}$. Class-wise weight vectors form the weight matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k] \in \mathbb{R}^{d \times k}$. Currently most of the large-scale image classification approaches adopt the simple strategy to train an independent one-vs-rest (OVR) binary classifier for each class due to its advantage on computational efficiency. As an example, the two top systems at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2010 used such an approach [13,26].

While one-vs-rest based approaches can yield competitive results in practice, they suffer from weak theoretical guarantees. Moreover, one-vs-rest based approaches assume that the classifier of each class is independent of other classifiers. However, this assumption usually does not hold in general since classes are related and built on some underlying common characteristics [1]. In many cases, especially when the number of classes is huge, the classes can be embedded in a much lower dimensional subspace rather than the ambient space. For example, as shown in Fig. 1, by representing each image of Caltech 101 dataset as a 11,200-dimensional Fisher vector [22], the weight matrix $\mathbf{W} \in \mathbb{R}^{11{,}200 \times 102}$ learned by one-vs-rest strategy with SVM indeed has a low rank structure. Therefore, the between-class transfer afforded by implicitly learning shared characteristics reveals much information which greatly benefits the final classification performance. Specifically, when only a limited number of examples are available for some classes of interest, such shared low rank structure is much more helpful.

To learn the underlying low rank structure between the classes and the classifiers simultaneously, the nuclear norm of matrices (*i.e.*, the sum of singular values) is used as a convex surrogate of the rank operator of matrices for regularization penalty [1,31].

* Corresponding author.
E-mail addresses: huyao001@gmail.com (Y. Hu), jinzhongming888@gmail.com (Z. Jin), yishi@zju.edu.cn (Y. Shi), debingzhangchina@gmail.com (D. Zhang), dengcai@gmail.com (D. Cai), xiaofeihe@gmail.com (X. He).

http://dx.doi.org/10.1016/j.neucom.2014.06.073
0925-2312/

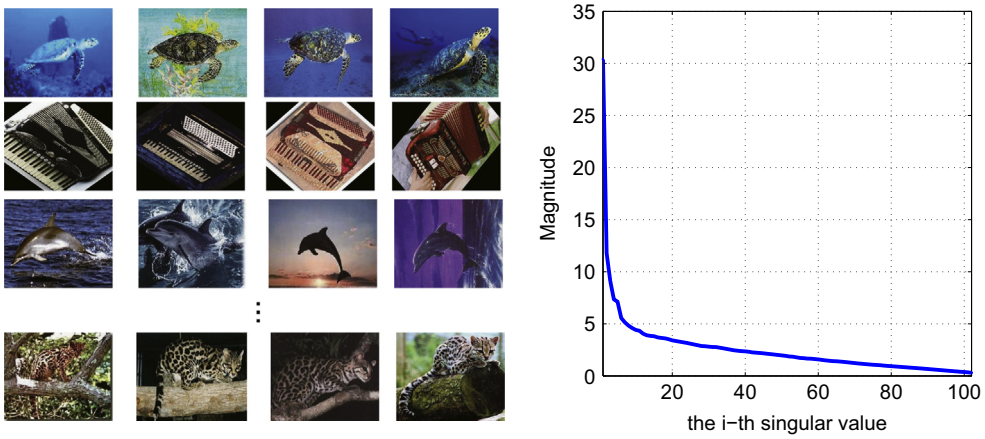

Fig. 1. A illustration to show the low rank nature of the classes behaviour in the multi-class classification. (Left) Samples of multi-class images in the Caltech 101 dataset; (Right) Spectrum of the weight matrix $\mathbf{W} \in \mathbb{R}^{11,200\times 102}$ learned by using one-vs-rest strategy with SVM. It is clear to see that the energy of the obtained weight matrix is dominated by the top 40 singular values.

However, many researchers have shown that nuclear norm cannot approximate [11,32] the rank function accurately. So nuclear norm based approaches usually fail to get low rank solutions in the real applications.

In this paper, we propose a new scalable large scale classification algorithm called *truncated nuclear norm regularization based multi-class* (TNNRMC) classification. The recently proposed *truncated nuclear norm regularization* [32] is used for penalization of the final obtained weight matrix of rank r instead of the nuclear norm. Different from the traditional nuclear norm which minimizes the summation of all the singular values, TNNR only minimizes the sum of the smallest $\min\{d,k\}-r$ singular values since the rank of a matrix only corresponds to the top r non-zero singular values. On the other hand, we consider the multinomial logistic loss function enjoying active theoretical guarantees for multi-class classification [2]. In this way, we further relax the final nonconvex and nonsmooth optimization problem through a simple and efficient iterative scheme and an accelerated coordinate descent scheme is adopted for subproblem in each iteration. Experimental results on several large scale datasets show that the proposed algorithm can obtain an improvement on the accuracy of classification.

The rest of the paper is organized as follows. In the next section, we provide a brief introduction of the related work. In Section 3, we detail our proposed approach for multi-class classification with low rank penalty. We introduce a simple and efficient coordinate descent algorithm to solve the optimization problem in Section 4. A variety of experimental results are presented in Section 5. Finally, we provide some concluding remarks in Section 6.

Notions: For a given matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$, the nuclear norm can be defined as $\|\mathbf{W}\|_* = \sum_{i=1}^{\min(d,k)} \sigma_i(\mathbf{W})$, where $\sigma_1(\mathbf{W}) \geq \sigma_2(\mathbf{W}) \geq \cdots \geq \sigma_{\min(d,k)}(\mathbf{W}) \geq 0$ are the singular values of the matrix $\mathbf{W}$. The spectral norm of the matrix $\mathbf{W}$ is its largest singular value, *i.e.*, $\|\mathbf{W}\|_2 = \sigma_1(\mathbf{W})$. For any $\theta \in \mathbb{R}^{\mathcal{I}}$ and $\mathcal{I}$ is an index set, its support can be defined as the set of indices which are nonzero, *i.e.*, $supp(\theta) = \{i \in \mathcal{I} | \theta_i \neq 0\}$.

2. Related work

The challenge of the accurate classification of an instance into one of a large number of target classes' surfaces exists in many domains, such as object recognition [9], face identification [30], and textual topic classification [18]. Traditionally researchers firstly try to find the compact representations image from different perspectives [14–17]. And then most of the previous approaches for large scale classification work in a similar way: one binary SVM is learned per class in a one-vs-rest fashion [25,26]. The reason why one-vs-rest strategy is popular is because of its scalability on the number of classes. Meanwhile, many researchers also propose to view the large scale classification as a ranking problem [29], where the goal is to rank the labels according to their relevance when given a query.

To uncover the latent structure between the classes, a variety of regularizations are proposed by considering multi-class classification problem from the different perspective. Crammer and Singer [6] propose to learn the weight vectors penalized by the Frobenius norm regularization, *i.e.* $\|W\|_F^2$, to obtain a stable solution when using an SVM to learn classifiers for each class independently. However, just as we referred previously, the classifiers are never independent in a dataset with a large number of categories. Many empirical evaluations have shown that the weight matrix in the multi-class setting has a low rank structure. So the rank operator of matrices, $rank(\cdot)$, is used as a regularization to control the rank of the finally obtained weight matrix $\mathbf{W}$. However, considering the nonconvex and discontinuous nature of the rank operator, the rank operator regularized problem is NP-hard in general. Inspired by compressed sensing, the nuclear norm $\|\mathbf{W}\|_*$ regularization [6] is proposed to approximate the nonconvex rank operator since nuclear norm is the tightest convex lower bound of the rank operator of matrices [24]. Following Fazel [8], the nuclear norm regularized problem can be formulated as a Semi-Definite Programming (SDP) problem. However, even the current best SDP solvers such as SDPT3 [28] and SeDuMi [27] can only handle the matrices whose size are less than 500×500 efficiently. This limits the usage of nuclear norm regularization in the large scale multi-class classification problem. Amit et al. [1] propose to use gradient method for the smooth relaxation of the nonsmooth nuclear norm regularized maximal hinge loss function by replacing the nuclear norm with a smooth proxy. However, this approach does not accurately capture the low rank structure between the classes because of the smooth processing of the objective function.

3. Multi-class classification with low rank penalty

Generally, the multi-class classification model can be formulated as an empirical risk minimization problem as follows:

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times k}} \lambda\Omega(\mathbf{W}) + \phi(\mathbf{W}), \tag{1}$$

where $\Omega(\mathbf{W})$ is the regularization penalty and $\phi(\mathbf{W})$ is the loss function. Specifically, in this paper, we focus on the widely used