Neurocomputing 152 (2015) 322–332

Contents lists available at ScienceDirect

Neurocomputing

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

Network traffic classification via non-convex multi-task feature learning

Dong Li, Guyu Hu, Yibing Wang, Zhisong Pan*

College of Command Information System, PLA University of Science and Technology, Nanjing 210007, China

ARTICLE INFO

Article history:
Received 11 July 2014
Received in revised form
9 September 2014
Accepted 22 October 2014
Communicated by M. Wang
Available online 13 November 2014

Keywords:
Traffic classification
Multi-task feature learning
Non-convex optimization
Capped-ℓ_1, ℓ_1
Iterative shrinkage and thresholding

ABSTRACT

Machine learning has been used in network traffic classification and statistical features are used to represent flows. However, conventional feature selection may work out in face of dynamic and complex traffic data. Multi-Task Learning has obtained quite wide attention nowadays, and one important form of multi-task learning is to exploit the features shared by tasks by sparse models. We propose a fast multi-task sparse feature learning method, using a non-convex Capped-ℓ_1, ℓ_1 as the regularizer to learn a set of shared features in traffic data. Specifically, the non-convex multi-task feature learning model can learn features belonging to each task as well as the common features shared among tasks. We use the iterative shrinkage and thresholding (IST) algorithm to solve the problem, which has a closed-form solution for one of the crucial steps in the whole iteration. Experiment on real traffic data captured from backbone network as well as synthetic data and other popular real-world data show the effectiveness the method, compared with state-of-the-art methods.

1. Introduction

Fast and accurate traffic identification is a fundamental interest for ISP and enterprise networks operators. Traditional techniques including port based or packet-level methods cannot meet requirement due to the continued development of the Internet with various web-based applications. For example, the popularity of P2P application using dynamic ports makes port-based identification invalid. Driven by the practical demands, flow-based traffic identification has become an important issue in the network community. In recent years, machine learning (ML) has been intensively used in network traffic anomaly detection as well as general classification [32,26,45,37,44]. As the ML algorithms are introduced to network traffic flow related problems, how to illustrate a single flow is crucial. Statistical features such as packet length and inter-arrival time as well as other min–max values of key characters have achieved much attention. A popular feature set (or called discriminators) by Moore [33,4] is widely used for machine learning algorithms which contains 248 features derived from packet headers.

The statistical features are based on the experience and background knowledge of experts, and all features are regarded equally and there is no tradeoff between few features for efficiency and more features for sufficiency. Previous literatures usually focus on the application of ML algorithms like SVM, Bayesian Networks, C4.5 Decision Trees etc., and many either use insufficient features or utilize primary feature selection/reduction methods such as forward and backward search [28]. Others adopted consistency-based or correlation-based filtering methods [37,29] to select the approximately optimal subset of features. Some of these methods rely on a certain metric, and only empirical studies have been reported. Moreover, conventional feature selection may just work under certain circumstances, and the selected feature cannot cope with dynamic traffic from multiple periods. Other feature reduction techniques like PCA could make the model less interpretable. The traffic data from a dynamic network actually involves multiple tasks instead of a single one. Furthermore, it may be hard to cover different features in certain period, i.e a single task when training data are needed. Conventional learning or feature selection may fail to capture some essence features of traffic in this case, thus we need other models. Roughly two techniques can cope with this stream data: incremental (online) learning and multi-task learning. This paper focus on the latter one.

Multi-task learning (MTL) [8] aims to improve learning ability by using multiple related tasks which may share common information. Here multiple tasks may have something in common while differ from each other due to various reasons. Thus the multi-task assumptions may meet the dynamic network traffic flow in practice. The common information contained in multiple tasks were exploited as hierarchical prior of Bayesian models [48], hidden nodes in neural networks [8], relational knowledge [41], predictive structure [1] and a subset of relevant features [3,24,31,11,15,19]. There are successful applications adopting multi-task learning in a wide range of areas

* Corresponding author.
E-mail address: hotpzs@hotmail.com (Z. Pan).

http://dx.doi.org/10.1016/j.neucom.2014.10.061
0925-2312/© 2014 Elsevier B.V. All rights reserved.

such as rank for web search [10], diagnosis of disease [58,53], audiology [6] and near infrared spectroscopy [9].

Among these methods, learning a subset of relevant features has received a lot of interests, and the so-called multi-task feature learning is exactly defined to learn a few common features across the tasks [2,3], and the well known ℓ_1 norm regularizer controls the number of features shared by all tasks. For example, the $\ell_{2,1}-$norm penalty [30] makes the entries in a column of a feature matrix either zeros or non-zeros, thus selecting features shared by all the tasks. Moreover, different from simply feature selection, the multi-task feature learning acquires the learning model at the same time. Furthermore, the number of tasks where each feature is shared is also an criterion of the importance of the feature. The general paradigm for classification and regression which minimizes the combination of empirical loss and the regularizer is still suitable for multi-task feature learning, and kinds of novel regularization help to work out the relationships among tasks. Thanks to the mature techniques for convex optimization, the convex multi-task feature learning models have been intensively explored.

In order to exactly reflect the commonness as well as the individuality among multiple tasks in a dynamic network, we need to learn features beyond a common set. From this perspective, some non-convex regularizers may help to uncover more important information compared with widely used convex regularizers. According to an intuitive explanation, a class of non-convex regularizers sometimes can approximate ℓ_0 norm better thus could capture the sparse structure. In other words, non-convex penalties are more flexible in terms of extracting features among multiple tasks. In this paper, we exploit the non-convex capped-ℓ_1,ℓ_1 regularizer model [17] learn the features specific to each task as well as the common features shared among tasks.

Solving a general non-convex optimization problem can be challenging. Gong et al. [18] recently proposed a general iterative shrinkage and thresholding (GIST) algorithm for a class of non-convex penalties, which has a closed-form solution for many used penalties including Capped-ℓ_1 that is very similar to the Capped-ℓ_1,ℓ_1 we used in this paper. Inspired by the GIST framework, we propose an IST algorithm to solve multi-task feature learning (ISTMTFL) problems with the non-convex regularizer Capped-ℓ_1,ℓ_1. The key idea is to solve a group of independent non-convex optimization problems by taking the advantage of the separability of original problem; these subproblems have the same form and one of the crucial step of the iterations turns to have a closed-form solution. And because of the separability, the method is easy to be parallelized thus can cope with large scale data. We use the proposed methods to learn a set of important features of real traffic data captured from backbone network. Other experiments on both synthetic data and real-world data show the efficiency of the proposed algorithm compared with state-of-the-art methods. In this paper, our main contributions can be summarized below:

- Propose an iterative shrinkage and thresholding for a non-convex multi-task feature learning problem.
- Apply the proposed model to learn common features among multiple traffic classification tasks.
- Point out the parallelization prospect and analyze the computational complexity of the algorithm.

The remainder of this paper is organized as follows. Section 2 introduces some work on multi-task feature learning and the GIST framework. We describe the proposed ISTMTFL in Section 3. We evaluate the methods on traffic flow data and other datasets in Section 4. Section 5 concludes this paper and some details for describing the algorithm are provided in the appendix. Table 1 lists the notations we use in this paper.

Table 1
Notations used in this paper.

Notations	Descriptions
$\mathbb{R}$	The set of real numbers
a	A scalar
$\mathbf{a}$	A vector
A	A matrix
w_i	The i-th entry of vector $\mathbf{w}$
a_{ij}	The entry i-th row of vector $\mathbf{w}$
$\Vert\cdot\Vert_1$	The ℓ_1-norm of a vector
$\Vert\cdot\Vert_2$	The ℓ_2-norm of a vector
$\Vert A\Vert_F$	The Frobenius norm of matrix A, $\sqrt{\sum_{i=1}^m\sum_{j=1}^n\vert a_{ij}\vert^2}$
$\Vert A\Vert_{p,q}$	The $\ell_{p,q}-$norm of matrix A, $\left(\sum_i\left((\sum_j\vert a_{ij}\vert^q)^{1/q}\right)^p\right)^{1/p}$
$.*$	The pairwise operator, $\mathbf{w}=\mathbf{u}.*\mathbf{v}$, then $w_i=u_iv_i$

2. Related works

The ℓ_1 norm regularization technique has widely used in compressed sensing [13] and sparse coding [40,27]. The geometrical sparse property of ℓ_1 norm also makes sparse coding a fundamental technique for image classification [46,51,50] as well as image clustering [49] in recent years. Although most of these works are carried on single task, it is nature to extend these ideas to multi-task learning settings [21].

2.1. Multi-task feature learning

For multi-task feature learning, conventional models may assume that some common features are shared by all tasks [38,30,22], where the $\ell_{p,q}-$norm penalty is used to address joint feature selection. However, many factual situations need a weaker condition or assumption, for a consideration of robustness. There probably exists two main ways to decouple the strong relationship posted on tasks and features. The first class focus on tasks. Robust multi-task learning tries to distinguish different tasks or find out outlier tasks by task clustering [23], Gaussian processes [52], structural regularization [11,16]. The second method starts from the features and uses various regularizer to control the common features. Both [24] and [17] have investigated the situation that certain features can be shared by some tasks but not all tasks. The convex $\ell_1+\ell_{1,\infty}$ and non-convex Capped-ℓ_1,ℓ_1 regularizers were used to meet the demands. Theocratical analysis shows that the non-convex regularizer in the latter case ([17, Remark 11]) could obtain a better bound under weaker conditions regarding parameters estimate error.

Multi-Stage (MS) convex relaxation has been used to solve non-convex penalties [54,55], which relax the original non-convex problem to a sequence of convex problems. Recent literature MSMTFL [17] applies this paradigm to multi-task feature learning by refining a non-convex problem into a convex one, and the error bound of MSMTFL could be improved during the multi-stage iteration. However, these kind of algorithms may involve high computational costs in the procedure of a sequence of stages, thus may not be suitable for large scale problems.

In terms of a class of optimization problems consist of convex penalties and smooth convex loss functions (always required with Lipschitz continuous gradient), the iterative shrinkage and thresholding (IST) and its accelerated version FISTA [5] or Nesterov's accelerated gradient method [35,36] have been widely applied to solve them. Examples include the classical ℓ_1-norm regularizer problem [43] and other variants based on it. These methods iteratively use the first order Taylor expansion of loss function plus the original regularizer to approximate the object function, and then generate the current solution by minimizing a proximal