Neural Networks 71 (2015) 214–224

Contents lists available at ScienceDirect

Neural Networks

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

Budget constrained non-monotonic feature selection

Haiqin Yang [a,b,*], Zenglin Xu [c,d,**], Michael R. Lyu [a,b], Irwin King [a,b]

[a] Shenzhen Key Laboratory of Rich Media Big Data Analytics and Application, Shenzhen Research Institute, The Chinese University of Hong Kong,

[b] Computer Science & Engineering, The Chinese University of Hong Kong, Hong Kong

[c] Big Data Research Center, University of Electronic Science & Technology, Chengdu, Sichuan, China

[d] School of Computer Science and Engineering, University of Electronic Science & Technology, Chengdu, Sichuan, China

ARTICLE INFO

Article history:
Available online 4 September 2015

Keywords:
Feature selection
Multiple kernel learning
Budget constraint
Non-monotonic

ABSTRACT

Feature selection is an important problem in machine learning and data mining. We consider the problem of selecting features under the budget constraint on the feature subset size. Traditional feature selection methods suffer from the "*monotonic*" property. That is, if a feature is selected when the number of specified features is set, it will always be chosen when the number of specified feature is larger than the previous setting. This sacrifices the effectiveness of the non-monotonic feature selection methods. Hence, in this paper, we develop an algorithm for non-monotonic feature selection that approximates the related combinatorial optimization problem by a Multiple Kernel Learning (MKL) problem. We justify the performance guarantee for the derived solution when compared to the global optimal solution for the related combinatorial optimization problem. Finally, we conduct a series of empirical evaluation on both synthetic and real-world benchmark datasets for the classification and regression tasks to demonstrate the promising performance of the proposed framework compared with the baseline feature selection approaches.

1. Introduction

Feature selection is an important task in machine learning and data mining since one is often restricted with budgeted computational resources, such as the memory size, the CPU speed, the communication rate, etc., in a large number of real-world applications. The goal of feature selection is to choose from the input data a subset of informative features (Huang, Yang, King, & Lyu, 2008; Yang, King, & Lyu, 2011). It is often used to reduce the computational cost or save storage space for problems with high dimensional data for problems with either high dimensionality or limited computational power. This is helpful to prevent overfitting for high-dimensional data with relatively small training samples (Tibshirani, 1996; Yang, Lyu, & King, 2013; Yang, Xu, King, & Lyu, 2010). Feature selection has found applications in a number of real-world problems, such as data visualization, natural language processing, computer vision, speech processing, bioinformatics, sensor networks, and group methods of data handling (Ivakhnenko, 1995; Reddy & Ravi, 2012; Tan, Tsang, & Wang, 2014; Thi, Vo, & Dinh, 2014; Wang, Bensmail, & Gao, 2014; Wang, Zhao, Hoi, & Jin, 2014; Wolf & Shashua, 2005). Comprehensive survey papers of feature selection can be found in Blum and Langley (1997), Guyon and Elisseeff (2003) and Kohavi and John (1997). The procedure of feature selection is analogous to pruning approach in neural networks, which aims to trim a network within the assumed initial architecture Augasta and Kathirvalavakumar (2013). Moreover, it is important to note that feature selection is different from feature extraction (He & Niyogi, 2003; Jolliffe, 1986; Kohonen, 2006), which maps the input data into a reduced representation set of features. Comparing with feature extraction, feature selection keeps the same space as the input data and thus has better interpretability for some specific applications.

In this paper, we consider the problem of feature selection under the budget constraint on the feature subset size. This setting is important for two reasons. On the one hand, budgeted learning is a new research aspect of machine learning since people are often facing a fixed budget in the presence of non-uniform cost functions for the acquisition of feature values, labels, or entire instances, and

* Correspondence to: Shenzhen Key Laboratory of Rich Media Big Data Analytics and Applications, Shenzhen Research Institute; Department of Computer Science and Engineering, The Chinese University of Hong Kong. Tel.: +852 31634251.
** Corresponding author at: Big Data Research Center, University of Electronic Science & Technology, Chengdu, Sichuan, China.
E-mail addresses: hqyang@ieee.org (H. Yang), zlxu@uestc.edu.cn (Z. Xu).

http://dx.doi.org/10.1016/j.neunet.2015.08.004
0893-6080/

for prediction errors (Dekel & Singer, 2006; Margineantu, Greiner, Singliar, & Melville, 2010). On the other hand, the number of required features also depends on the objective of the task, and there is no single number of features that are optimal for all tasks. For example, for data visualization, only two or three features are necessary. In this work, we assume that an external oracle decides the number of selected features.

Given the budget of the feature subset size, denoted by m, the goal of feature selection is to choose a subset of m features, denoted by $\mathcal{S}$, that maximizes a generalized performance criterion $\mathcal{Q}$. It is cast into the following combinatorial optimization problem:

$$\mathcal{S}^* = \arg\max_{\mathcal{S}} \mathcal{Q}(\mathcal{S}) \quad \text{s.t. } |\mathcal{S}| = m. \tag{1}$$

A number of performance criteria have been proposed for feature selection, including mutual information (Koller & Sahami, 1996), maximum margin (Guyon, Weston, Barnhill, & Vapnik, 2002; Weston et al., 2000), kernel alignment (Cristianini, Shawe-Taylor, Elisseeff, & Kandola, 2001; Neumann, Schnörr, & Steidl, 2005), worst case classification bounds (Bhattacharyya, 2004; Xu, King, & Lyu, 2007), graph-spectrum based measures (Zhao & Liu, 2007), Parzen window (Yu, Ding, & Loscalzo, 2008), clustering-based measures (Boutsidis, Mahoney, & Drineas, 2009; Fisher, 1996), PCA-based measures (Malhi & Gao, 2004), and the Hilbert Schmidt independence criterion (Song, Smola, Gretton, Borgwardt, & Bedo, 2007), etc. Among them, due to the effectiveness, the maximum-margin-based criterion is probably one of the most widely used criteria for feature selection.

The computational challenge in solving the optimization problem in Eq. (1) arises from its combinatorial nature, i.e., a binary selection of features that maximizes the performance criterion $\mathcal{Q}$ given the number of required features. A number of feature selection algorithms have been proposed to approximately solve Eq. (1). Most of them first compute a score or weight for each feature, and then select the features with the largest scores. For instance, a common approach is to first learn an SVM model, and select m features with the largest absolute weights. This idea was justified in Vapnik (1998) by sensitivity analysis and was also utilized for feature selection. A similar idea was used in SVM-Recursive Feature Elimination (SVM-RFE) (Guyon et al., 2002), where features with smallest weights were removed iteratively. In Fung and Mangasarian (2000) and Ng (2004), regularization on the L_1-norm of weights was suggested to replace the L_2-norm for feature selection when learning an SVM model. Another feature selection model related to the L_1-norm is lasso (Tibshirani, 1996), which selects features by constraining the L_1-norm of weights. By varying the L_1-norm of weights, a regularization path of selected features can be tracked. A similar model is LARS (Efron, Hastie, Johnstone, & Tibshirani, 2004), which can be regarded as unconstrained version of lasso. Other models related to the L_1-norm regularization include the direct optimization over the L_1-norm of the feature indicator (Sonnenburg, Rätsch, Schäfer, & Schölkopf, 2006; Xu, King, Lyu, & Jin, 2010). In addition to the optimization on the L_2-norm and the L_1-norm, several studies (Bradley & Mangasarian, 1998; Chan, Vasconcelos, & Lanckriet, 2007; Huang, King, & Lyu, 2008; Neumann et al., 2005; Weston, Elisseeff, Schölkopf, & Tipping, 2003) explored the L_0-norm when computing the weights of features. In Bradley and Mangasarian (1998), the authors proposed Feature Selection Concave method (FSV) that uses an approximate of the L_0-norm of the weights. It was improved in Neumann et al. (2005) and Weston et al. (2003) via an additional regularizer or a different approximation of the L_0-norm. In addition to selecting features by weights, in Rakotomamonjy (2003), Vapnik (1998) and Weston et al. (2000), the authors proposed to select features based on $R^2\|\mathbf{w}\|^2$, where R is the radius of the smallest sphere that contains all the data points.

Although the above approximate approaches have been successfully applied to a number of applications of feature selection, they are limited by the **monotonic** property of feature selection that is defined below:

Definition 1 (*Non-Monotonic Feature Selection*)**.** A feature selection algorithm $\mathcal{A}$ is monotonic if and only if it satisfies the following property: for any two different numbers of selected features, i.e., k and m, we always have $\mathcal{S}_k \subseteq \mathcal{S}_m$ if $k \leq m$, where $\mathcal{S}_m$ stands for the subset of m features selected by $\mathcal{A}$. Otherwise, it is called non-monotonic feature selection.

To see the monotonic property of most existing algorithms for feature selection, first note that these algorithms rank features according to their weights/scores that are computed independently from the number of selected features m. Hence, if a feature f is chosen when the number of selected features is k, it will also be chosen if the number of selected features m is larger than k. In other words, $f \in \mathcal{S}_k \rightarrow f \in \mathcal{S}_m$ if $k < m$, and therefore $\mathcal{S}_k \subseteq \mathcal{S}_m$. As argued in Guyon and Elisseeff (2003), since variables that are less informative by themselves can be informative together, a monotonic feature selection algorithm may be suboptimal in identifying the set of most informative features. To further motivate the need of non-monotonic feature selection, we consider a binary classification problem with three features X, Y, Z. Fig. 1(a)–(c) show the projection of data points on individual features X, Y and Z, respectively. We clearly see that Z is the most informative feature to the two classes. Fig. 1(d)–(f) show the projection of data distribution on the plane of two joint features XY, XZ, and YZ, respectively. We observe that XY are the two most informative features. Note that although Z is the single most informative feature, its combinations with any other feature are not as informative as XY, which justifies the need of non-monotonic feature selection.

In this paper, we propose a **non-monotonic** feature selection method that solves the optimization problem in Eq. (1) approximately. In particular, we alleviate the monotonic property by computing scores for individual features that depend on the number of selected features m. We first convert the combinatorial optimization problem in Eq. (1) into a formulation that is closely related to multiple kernel learning (MKL) (Lanckriet, Cristianini, Bartlett, Ghaoui, & Jordan, 2004; Sonnenburg et al., 2006; Xu, Jin, Ye, Lyu, & King, 2009; Yang, Xu, King, & Lyu, 2014; Yang, Xu, Ye, King, & Lyu, 2011). The key idea is to first construct a separate kernel matrix for each feature, and then find the binary combination of kernels that minimizes the margin classification error. We relax the original combinatorial optimization problem into a convex optimization problem that can be solved efficiently by expressing it as a Quadratically Constrained Quadratic Programming (QCQP) problem. We present a strategy that selects a subset of features based on the solution of the relaxed problem, which can still maintain the non-monotonic property. This is different from the recent work in Tan et al. (2014). We furthermore show the **performance guarantee**, which bounds the difference in the value of objective function between using the features selected by the proposed strategy and using the global optimal subset of features found by exhaustive search. Our empirical study shows that the proposed approach performs better than the baseline methods for feature selection. Finally, we would like to clarify that although our work involves the employment of MKL, the focus of our work is not to develop a new algorithm for MKL, but an efficient algorithm for non-monotonic feature selection.

The rest of this paper is organized as follows. We present the non-monotonic feature selection for classification and regression in Sections 2 and 3, respectively. Sections 4 and 5 present experimental results with a number of benchmark datasets for classification and regression, respectively. We conclude our work in Section 6.