Information Sciences 000 (2016) 1–9

Contents lists available at ScienceDirect

Information Sciences

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

Granularity selection for cross-validation of SVM

Yong Liu, Shizhong Liao*

School of Computer Science and Technology, Tianjin University, Tianjin 300072, P. R. China

ARTICLE INFO

Article history:
Received 30 August 2015
Revised 25 April 2016
Accepted 28 June 2016
Available online xxx

Keywords:
Cross-validation
Model selection
Fold of cross-validation
Granular computing
Granularity selection

ABSTRACT

Granularity selection is fundamental to granular computing. Cross-validation (CV) is widely adopted for model selection, where each fold of data set of CV can be considered as an information granule, and the larger the number of the folds is, the smaller the granularity of each fold is. Therefore, for CV, granularity selection is equal to the selection of the number of folds. In this paper, we explore the granularity selection for CV of support vector machine (SVM). We first use the Huber loss to smooth the hinge loss used in SVM, and to approximate CV of SVM. Then, we derive a tight upper bound of the discrepancy between the original and the approximate CV with a high convergence rate. Finally, based on this derived tight bound, we present a granularity selection criterion for trading off the accuracy and time cost. Experimental results demonstrate that the approximate CV with the granularity selection criterion gives the similar accuracies as the traditional CV, and meanwhile significantly improves the efficiency.

1. Introduction

Granular computing has a wide range of applications in data mining, pattern recognition and machine learning [21,28–30]. How to select a proper granularity is one of the fundamental issues in the research and application of granular computing [8,21,28,32,33]. Cross-validation (CV) [20,24] is a tried and tested approach for selecting the optimal model [9,18,19], which is widely used in granular computing. In t-fold CV, the data set is split into t disjoint subset of (approximately) equal size and the algorithm (or model) is trained for t times, each time leaving out one of subsets from training, but using the omitted subset to compute the validation error. The t-fold CV estimate is then the average of the validation errors observed in t iterations, or folds. Each subset of t-fold CV can be considered as the information granule [13,21,23,32], the larger number of folds means the smaller granularity of each subset and the higher cost. Therefore, for CV, granularity selection is equal to the selection of the number of folds, which is a key problem to CV.

Support vector machine (SVM) is an important machine learning method widely adopted in granular computing [22,25,26,31]. The performance of SVM greatly depends on the choice of some hyper-parameters (such as the kernel parameter and regularization parameter), hence how to select the optimal hyper-parameters is important to SVM [1,14,16]. Although the t-fold CV is a commonly used approach for selecting the hyper-parameters for SVM [2,4,17], it requires training t times, which is computationally intensive. For the sake of efficiency, some approximate leave-one-out CV for SVM are given: such as generalized approximate cross-validation (GACV) [27], radius-margin bound [26], span bound [5], support vector count [26]. However, there is few work on the approximation of the general t-fold CV (for all t). Instead of using the full grid, the local search heuristics is used to find local minima in the validation error to speed up the computation of CV

* Corresponding author.
E-mail address: szliao@mail.tju.edu.cn, szliao@tju.edu.cn (S. Liao).

http://dx.doi.org/10.1016/j.ins.2016.06.051
0020-0255/

Please cite this article as: Y. Liu, S. Liao, Granularity selection for cross-validation of SVM, Information Sciences (2016), http://dx.doi.org/10.1016/j.ins.2016.06.051

[10,11]. In [12], an improved CV procedure is proposed, which uses nonparametric testing coupled with sequential analysis to determine the best parameter set on linearly increasing subsets of the data. Different from the above approximate CV methods that speed up the grid-search procedure, in our previous work [15], we present a strategy for approximating the CV error for a class of kernel-based algorithms, in which the loss function must be differentiable. Unfortunately, the hinge loss used in SVM is not differentiable, so the approximate strategy proposed in [15] can not be used for SVM.

In this paper, we present an approximate CV approach for SVM, and further present a novel granularity selection method for it. Specifically, we first use the Huber loss to approximate the hinge loss, and give an approach to approximating the CV of SVM using the Huber loss. Then, we derive a tight upper bound of the discrepancy between the original and approximate CV errors of order $\mathcal{O}\left(\frac{1}{t\cdot r}\right)$, where t is the number of folds and r is the order of Taylor expansion. Finally, based on the derived tight bound, we present a granularity selection criterion to trade off the performance of approximation and the computational cost. The proposed approximate CV requires training on the full data set only once, hence it can significantly improve the efficiency. Experimental results demonstrate that the approximate CV with the granularity selection criterion is sound and efficient.

The rest of the paper is organized as follows. We start by introducing some preliminaries and notations in Section 2. We then propose a novel strategy for approximating the CV of SVM in Section 3. In Section 4, we present a granularity selection criterion to choose the number of folds. We empirically analyze the performance of our approximate CV with the granularity selection criterion in Section 5. We end in Section 6 with conclusion. All the proofs are given in Appendix.

2. Preliminaries and notations

We consider the supervised learning where a learning algorithm receives a sample of n labeled points

$$S = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^n, z_i \in \mathcal{Z} = (\mathcal{X} \times \mathcal{Y}),$$

where $\mathcal{X}$ denotes the input space and $\mathcal{Y} = \{-1, +1\}$ the output space. We assume S is drawn identically and independently from a fixed, but unknown probability distribution P on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel [3], and assume $K(\mathbf{x}, \mathbf{x}) \le 1$, $\forall \mathbf{x} \in \mathcal{X}$[1] . The reproducing kernel Hilbert space (RKHS) associated with K is defined to be the completion of the linear span of the set of functions $\mathcal{H}_K = \text{span}\{\Phi(\mathbf{x}) = K(\mathbf{x}, \cdot) : \mathbf{x} \in \mathcal{X}\}$ with the inner product denoted as $\langle \cdot, \cdot \rangle_K$ satisfying $\left\langle K(\mathbf{x}, \cdot), K(\mathbf{x}', \cdot) \right\rangle_K = K(\mathbf{x}, \mathbf{x}')$. The learning algorithms we study is SVM [7,26]:

$$f_{P_S}^{\text{svm}} := \arg\min_{f \in \mathcal{H}_K} \frac{1}{|S|} \sum_{z_i \in S} \ell(y_i f(\mathbf{x}_i)) + \lambda \|f\|_K^2,$$

where $\ell(\cdot)$ is the hinge loss $\ell(t) = \max(0, 1 - t)$, λ is the regularization parameter, and $|S|$ is the size of S.

Let $S_1, \ldots, S_t$ be a random equipartition of S into t parts, called folds. For simplicity, assume that n mod t, and hence, $|S_i| = \frac{n}{t} =: l$, $i = 1, \ldots, t$. Each S_i can be considered as an information granule [13,23,32]. Note that the larger t, the smaller size of S_i, which implies the smaller granularity of S_i. Thus, the selection of fold can be regarded as the granularity selection in CV.

Let $P_{S\setminus S_i}$ be the empirical distribution of the sample S without the observations S_i, that is

$$P_{S\setminus S_i} = \frac{1}{n-l} \sum_{z_i \in S\setminus S_i} \delta_{z_i}, \tag{1}$$

where δ_{z_i} is the Dirac distribution in z_i. The hypothesis learned on all of the data excluding S_i can be written as:

$$f_{P_{S\setminus S_i}}^{\text{svm}} := \arg\min_{f \in \mathcal{H}_K} \frac{1}{n-l} \sum_{z_i \in S\setminus S_i} \ell(y_i f(\mathbf{x}_i)) + \lambda \|f\|_K^2.$$

Then, the t-fold CV error can be written as

$$t\text{-CV} := \frac{1}{n} \sum_{i=1}^{t} \sum_{z_j \in S_i} I\left(y_j f_{P_{S\setminus S_i}}^{\text{svm}}(\mathbf{x}_j)\right),$$

where, $I(c) = 1$ if $c < 0$, otherwise 0. Although the t-CV is wildly used for model selection, it requires training t times, which is computationally expensive. In our previous work [15], we present a strategy to approximate the t-CV based on Bouligand influence function (BIF) [6] for some kernel-based algorithms, in which the loss function must be differentiable. This approximate CV needs to be trained only once, hence it can significantly improve the efficiency. Unfortunately, the hinge loss used in SVM is not differentiable so that the approximate strategy proposed in [15] can not be used for SVM, directly. To address this problem, in the next section, we will propose to use a differentiable approximation of the hinge loss, inspired by the Huber loss.

[1] $K(\mathbf{x}, \mathbf{x}) \le 1$ is a common assumption, for example, met for the popular Gaussian kernel and Laplace kernel.

Please cite this article as: Y. Liu, S. Liao, Granularity selection for cross-validation of SVM, Information Sciences (2016), http://dx.doi.org/10.1016/j.ins.2016.06.051