Information Sciences 232 (2013) 419–436

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

# An efficient classification approach for large-scale mobile ubiquitous computing

Feilong Tang [a,*], Ilsun You [b], Can Tang [c], Minyi Guo [d]

[a] School of Software, Shanghai Jiao Tong University, Shanghai 200240, China
[b] School of Information Science, Korean Bible University, Seoul, South Korea
[c] Department of Finance, Heilongjiang University, Harbin 150080, China
[d] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

*Article history:*
Received 31 August 2011
Received in revised form 30 August 2012
Accepted 20 September 2012
Available online 16 October 2012

*Keywords:*
Classification
Semi-sparse algorithm
Multi-class multi-label classification
Parallelization
Sequential minimal optimization (SMO)
Mobile ubiquitous computing

ABSTRACT

Context classification is at the center of user-centric ubiquitous computing that targets the provision of personalized services based on expressed preferences and interests. Classification of context for Mobile Ubiquitous Computing (MUC), where there are high volumes of data and users place large demands on a context-aware system, must be effective and efficient in computational terms. The Sequential Minimal Optimization (SMO) based SVMTorch is widely used for text classification; it is however inefficient for MUC-oriented context analysis due to: (1) a low classification speed caused by inefficient matrix multiplication, and (2) the inability to classify multi-label data.

In this paper, we propose an efficient classification approach to improve and extend the SVMTorch. Firstly, we propose a semi-sparse algorithm to speed up vector/matrix multiplication which lies at the core of the SVMTorch-based classification approaches. Theoretically, to multiply two vectors (i.e., a selected vector and a trained vector) with $m$ and $n$ non-zero elements respectively, the traditional SVMTorch needs $O(m + n)$ time while our semi-sparse algorithm requires only $O(n)$ time, where $n$ is the number of non-zero elements in the trained vector. Secondly, we extend the functions of the traditional SVMTorch approach which is limited to the classification of single-label data, to support multi-class multi-label classification. Finally, we parallelize the improved SVMTorch which incorporates the semi-spares algorithm and function extensions to access multi-core processor and cluster systems to further improve the effectiveness and efficiency of the classification process. The experimental results demonstrate that our proposed solution significantly improves the performance and capability of the traditional SVMTorch. The results support the conclusion that the larger training and testing data sets are, the more improvement our solution brings to the effectiveness and efficiency of the context classification. This conclusion is verified in a Chinese web page classifier developed based on the solution presented in this paper.



## 1. Introduction

Ubiquitous computing is a user-centric mobile computing paradigm in which users enjoy service provision in a dynamic mobile situation [24,46]. To provide mobile users with preferred services adaptively classification based on context is a prerequisite in many ubiquitous applications; for example, in recommender systems the goal is targeted service provision of resources to specific users based on h/her context. However, in *Mobile Ubiquitous Computing* (MUC) environments there

* Corresponding author. Tel.: +86 2134202949; fax: +86 2134204728.
*E-mail address:* tang-fl@cs.sjtu.edu.cn (F. Tang).

0020-0255/$ - see front matter 
http://dx.doi.org/10.1016/j.ins.2012.09.050

are serious challenges to the classification of text due to: (1) *large-scale and high-dimensional contextual data*, and (2) *real-time implementation where users have highly dynamic spatial change*. Consequently, context classification approaches for MUC applications should be sufficiently fast and versatile to meet the highly dynamic environments and provide support for multi-class multi-label context analysis [11,33,44].

Among existing classification models, the machine learning based support vector machine (SVM) outperforms other methods especially in high-dimensional nonlinear classification [2,34,39]. SVMTorch is a classic implementation of the SVM. It exhibits excellent performance on text classification, and has been widely used in face recognition, image processing and other areas. However, for MUC environments with the characteristics identified in points 1 and 2 above the SVMTorch fails to perform effectively due to a low classification speed caused by inefficient matrix multiplication and the inability to implement multi-label classification.

The *Sequential Minimal Optimization* (SMO) algorithm forms the kernel of the SVMTorch. Different from other classification methods that solve the largest possible optimization problem at each step, the SMO solves the smallest possible optimization problem in each iteration. Specifically, the SMO chooses two Lagrange multipliers to jointly optimize for these multipliers and updates the SVM to reflect new optimal values at each step [9,25]. In traditional implementations (e.g., SVMTorch) of the SMO, there is a large number of sparse matrix multiplication for not only the linear kernel but also the Gaussian kernel. We observe that during the matrix multiplication, the selected sparse vector is thoroughly traversed in each iteration, which inevitably results in time-consuming training and testing processes, especially to large-scale sparse matrices. On the other hand, SVMTorch only supports single-label classification, therefore it is incapable of classifying MUC-oriented contextual data with multiple labels.

The motivation of this paper is set to investigate and solve the identified challenging issues. The solution proposed in this paper improves matrix multiplication efficiency of the SVMTorch and extends its functions to support multi-class multi-label data classification in multi-core processor and cluster systems. The main contributions of this paper are summarized as follows.

- We propose a semi-sparse algorithm for vector/matrix multiplication to improve the traditional SMO algorithm. Traversing a selected vector at each iteration in the SMO wastes a large amount of training and testing time during the sparse matrix multiplication. Our semi-sparse algorithm can speed up the training and testing processes of the SVMTorch classifier through *densifying* the selected sparse vector to avoid redundant traverses. A theoretical analysis indicates that the traditional SMO requires $O(m + n)$ time on traversing two vectors (a selected vector and a trained vector) with $m$ and $n$ non-zero elements in a vector multiplication. Our semi-sparse algorithm improves efficiency in that it only needs $O(n)$ time for vector multiplication without reducing classification accuracy of the SVMTorch. The experimental results also show that the two implementations of our semi-sparse algorithm can reduce the training time to 53.7% and 74.95% of that in the traditional SVMTorch respectively.
- We extend the traditional SVMTorch classifier to support multi-label data classification. The traditional SVMTorch has a good behavioral characteristics in two-class and multi-class single-label classifications; however, it cannot support multi-class multi-label classification tasks. This solution proposed in this paper transforms a multi-label problem into multiple single-label sub-problems based on the 1-vs-many policy and proposes a comprehensive evaluation metrics. Experimental results demonstrate that our approach exhibits excellent precision and performance for multi-class multi-label classification tasks.
- We parallelize the SVMTorch classifier. This demonstrates enhanced performance in both the semi-sparse matrix multiplication and multi-label classification in multi-core processor and cluster systems. Existing SVMTorch implementations are restricted to single-core processor systems; this limits its classification speed. We use MPI (Message Passing Interface) to simultaneously distribute multiple sub-tasks in a training or testing process to different cores; this enables a significant increase in the speed of the classification process significantly.
- A Chinese web page classifier has been developed predicated on our novel semi-sparse algorithm, parallelization technology and multi-label classification approach. It has been shown to work very well in large-scale Chinese web page classification.

The remainder of this paper is organized as follows. In Section 2, we introduce preliminaries with a discussion on the background. In Section 3, we present our novel semi-sparse algorithm for vector/matrix multiplication with an analysis. Section 4 presents our proposed approach to multi-class multi-label context classification and the parallelization scheme. Experimental results and comprehensive performance evaluations are reported in Section 5. The paper closes with conclusions set out in Section 6.

## 2. Preliminaries and background

In this section we initially introduce some preliminaries followed by a review related research.

### 2.1. Preliminaries

The SVM, developed by Vapnik [43], exhibits many attractive features including high classification precision. The SVM solves classification problems generally through a quadratic optimization. Let there be a training set with $\ell$ samples