Applied Mathematics and Computation 219 (2013) 8876–8880

Contents lists available at SciVerse ScienceDirect

Applied Mathematics and Computation

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

Kernel-improved Support Vector Machine for semanteme data

Zhihua Li [a,b,c,*], Xue Yang [b,c], Wenqiang Gu [b,c], Haitao Zhang [a,b]

[a] Key Laboratory of Advanced Process Control for Light Industry Ministry of Education, JiangSu, WuXi 214122, PR China
[b] Engineering Research Center of Internet of Things Technology Application Ministry of Education, JiangSu, WuXi 214122, PR China
[c] Engineering School of Internet of Things, JiangNan University, JiangSu, WuXi 214122, PR China

ARTICLE INFO

Keywords:
Semanteme-based Support Vector Machine
Dissimilarity measure
Inner production computation
Improved kernel
Semanteme data classification

ABSTRACT

The computation for the inner production of semanteme data and the Support Vector Machine (SVM) classification of semantic data are very difficult. In this paper, a novel kernel-based semanteme data classification method is proposed, and the (SVM) is extended to semantic SVM, called the Semanteme-based Support Vector Machine (SSVM). A new dissimilarity definition and a simple inner production computation method for semanteme data are presented, and the parameters for optimization selection in SSVM are also discussed. In the proposed algorithm, the solving support vector process based on the new inner production computation method remains a quadratic program problem but has a lower computation complexity. The SSVM is insensitive to outliers, and its classification capability for unbalanced data in actual datasets is analyzed. The experimental results demonstrate the average advantage of SSVM over algorithm C4.5, adaptive dissimilarity metric, and value difference metric in terms of classification and robust capability, indicating that the proposed method has promising performance.

1. Introduction

A large number of classification algorithms depend on the similarity or dissimilarity of exemplars, such as Euclidean distance and inner production, among others. However, majority of these algorithms only process continuous-attributed data, not semantic data. Semantic data are extreme, having disordered and unbalanced distribution [1]. The dissimilarity of different exemplars among data items is very weak, and such exemplars even intersect with each other [2,4]. Therefore, the diversity of exemplars is very difficult to measure. The primary metric methods for semantic data at present are the overlap metric [2], value difference metric (VDM) [3], adaptive dissimilarity metric (ADM) [4], and hamming distance metric [5]. However, these methods have a number of limitations [6]. The metric methods [3,5,6] are based on the hypothesis that if the entities are different, the degree of their dissimilarity is the same, that is, the dissimilarity value is either "0" or "1". However, such methods do not show the degree of dissimilarity between different exemplars. The metric methods [4] are also based on the hypothesis that the attributes of exemplars are mutually independent, which does not coincide with actual cases. Additionally, numerous kernel-based clustering algorithms and methods, such as Support Vector Machine (SVM), map data items from the input space into the feature space to expand the distance of exemplars in the input space. The classification margin is maximized to improve the capability of classifiers and provide excellent performance in small datasets [7–9]. However, such methods fail to compute the inner production of the semantic data [6,10]. Semantic data are found in various applications, such as in credit data and attributes of blood indices, among others. However, most classification methods are unsuitable for classifying semantic datasets.

* Corresponding author at: Engineering Research Center of Internet of Things Technology Application Ministry of Education, JiangSu, WuXi 214122, PR China.
E-mail address: ezhli@yahoo.com.cn (Z. Li).

0096-3003/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.amc.2013.03.069

This paper presents a simple method for computing the inner production of semanteme data by giving a new definition of dissimilarity for this type of data through further studies on kernel function, kernel methods, and SVM. A kernel-based semanteme data classification method rooted at the SVM is proposed, called semanteme-based Support Vector Machine (SSVM), and this approach exploits new application aspects of SVM, semantic data, and heterogeneous data. SSVM can accomplish the classification computation of both semantic and heterogeneous data. Experiments show that SSVM has promising performance.

The remainder of this paper is structured as follows: Section 2 reviews the kernel-based method and function in SVM. Section 3 discusses the new dissimilarity and computation method of inner products based on semantic data and describes a kernel-based classification method for semantic data. Section 4 presents several examples on standard datasets to demonstrate the proposed method and assess its performance. Finally, Sections 5 states a number of concluding remarks, as well as possible avenues for further research.

2. Kernel methodologies and kernel functions in SVM

Based on the Mercer kernel theory, the kernel method is an operator that can be used as inner production computation function only if an operator satisfies the Mercer kernel conditions [8]. This theorem can solve complex, nonlinear classification problems without increasing computational complexity. The famous SVM is a successful application of kernel. The kernel function in SVM is summarized into two main types [10], namely, distance-based kernel and inner production-based kernel. The corresponding styles of such kernels can be formulated as (1) and (2), respectively.

$$K(x_i, x_j) = f(d_U(x_i, x_j)), \tag{1}$$

$$K(x_i, x_j) = f(I(x_i, x_j)), \tag{2}$$

where $d_U(x_i, x_j)$ and $I(x_i, x_j)$ are the Euclidean distance and Euclidean inner production in input space, respectively; and the $f(\cdot)$ is a mapping function from the input space to the feature space.

3. Discontinuous-based inner production and the proposed approach

No similarity metric or order relationship [11] exists in semantic datasets. The "distance" in pattern recognition is difficult to identify, thus making the measurement of the similarity or dissimilarity of data difficult. Given the existing non-metric characteristics in semantic data [11,12], a problem on the classification exists. This paper presents a generalized metric distance and studies the classification algorithms for semanteme data considering the kernel method.

3.1. Dissimilarity

Given a dataset $x_i \in X$, $x_i = (x_{i1}, x_{i2}, \ldots, x_{id}, x_{i(d+1)}, \ldots, x_{i(d+m)})^T$, and $1 \leqslant i \leqslant n$, the former d dimensions are semantic attributes, and the latter m dimensions are continuous attributes, that is, X is a heterogeneous dataset. The dissimilarity metric $d_{symbolic}$ between the semantic data exemplars is also defined.

Definition 1. The dissimilarity measure $d_{symbolic}(x_i, x_j)$ between the semantic entity x_i and x_j is defined as follows:

$$s(x_i, x_j) = \sum_{p=1}^{l} \theta(x_{ip}, x_{jp}), \tag{3}$$

$$\theta(x_{ip}, x_{jp}) = \begin{cases} 1 (x_{ip} = x_{jp}) \\ 0 (x_{ip} \neq x_{jp}) \end{cases} (i \neq j, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n, 1 \leqslant p \leqslant l, 1 \leqslant l \leqslant d),$$

$$d_{symbolic}(x_i, x_j) = [\dim s - s(x_i, x_j)]/\dim s, \tag{4}$$

where dim s is the total number of dimensions for the semantic attributes. By comparing Definition 1 with the metric method of [2–4], the obvious advantages of Definition 1 are summarized as follows:

- The new metric overcomes the absolute metric value of either "0" or "1" in [2] by transforming the dissimilarity of the semantic data into a fuzzy value belonging to the interval [0, 1]. The new metric method is evidently more objective and actual than the others.
- The domain [0,1] of $d_{symbolic}(x_i, x_j)$ plays an important role in normalization and benefits from being used in combination with other metric methods.

Given these definitions, if the Euclidean distance is used to measure the dissimilarity between the continuous-attributed exemplars, then the dissimilarity between the heterogeneous x_i and x_j can also be defined.