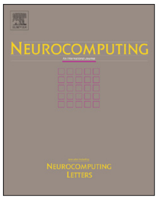
Contents lists available at ScienceDirect

Neurocomputing

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



Adaptive memetic algorithm enhanced with data geometry analysis to select training data for SVMs

Jakub Nalepa*, Michal Kawulok

Institute of Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

ARTICLE INFO

Article history:
Received 12 March 2015
Received in revised form
8 October 2015
Accepted 16 December 2015
Communicated by A. Abraham

Keywords:
Memetic algorithm
Principal component analysis
Adaptation
Parameter control
Support vector machine
Training data

ABSTRACT

Support vector machines (SVMs) are one of the most popular and powerful machine learning techniques, but suffer from a significant drawback of the high time and memory complexities of their training. This issue needs to be endured especially in the case of large and noisy datasets. In this paper, we propose a new adaptive memetic algorithm (PCA^{2}MA) for selecting valuable SVM training data from the entire set. It helps improve the classifier score, and speeds up the classification process by decreasing the number of support vectors. In PCA^{2}MA, a population of reduced training sets undergoes the evolution, which is complemented by the refinement procedures. We propose to exploit both a priori information about the training set—extracted using the data geometry analysis—and the knowledge attained dynamically during the PCA^{2}MA execution to enhance the refined sets. Also, we introduce a new adaptation scheme to control the pivotal algorithm parameters on the fly, based on the current search state. Extensive experimental study performed on benchmark, real-world, and artificial datasets clearly confirms the efficacy and convergence capabilities of the proposed approach. We demonstrate that PCA^{2}MA is highly competitive compared with other state-of-the-art techniques.

1. Introduction

Support vector machine (SVM) [9] is a supervised classifier which has been successfully applied for solving a plethora of pattern recognition problems [52,7,22,55,56]. SVMs classify the incoming (unseen) data based on the decision hyperplane defined by a subset of the training set ($\boldsymbol{T}$) vectors—referred to as *support vectors* (SVs). The process of determining SVs is a constrained quadratic programming (QP) problem of $O(t^3)$ time and $O(t^2)$ memory complexity, where t is the cardinality of $\boldsymbol{T}$. It appears to be a major drawback of SVMs in various domains, e.g., bioinformatics, document categorization, medical image analysis, and many more [40], in which datasets can easily become enormously large. Also, many real-life recognition problems are not linearly solvable and require a non-linear decision function which provides robust classification. The *kernel trick* was therefore introduced to obtain a non-linear decision hyperplane in SVMs [5]. It consists in defining a kernel function (which must satisfy the theorem discussed by Mercer [29]) that computes the inner product of two feature vectors in a derived non-linear feature space.

There are two main streams of research to deal with the challenging SVM training in the case of large datasets. The first approach consists in optimizing the training procedure to speed up solving the underlying QP problem, whereas the second one in determining a valuable subset[1] of $\boldsymbol{T}$ ($\boldsymbol{T}'$ with t' samples, where $t' \ll t$). $\boldsymbol{T}'$ is then fed to the SVM training procedure. In the former case, it is necessary to address the problem of a large memory complexity induced by many approaches [16]. On the other hand, selecting a good subset of $\boldsymbol{T}$ helps decrease the number of determined SVs. It is very beneficial, since the SVM classification time depends linearly on the number of SVs (s) and is of $O(s)$ time—the large number of SVs may significantly slow down the classification. Minimizing s while keeping the classification score high reduces therefore the SVMs response time and makes them suitable for real-time applications even in the case of large datasets. The second group of approaches is thus being actively developed.

In our earlier work, we introduced a genetic algorithm (GA) to select training data for SVMs [20], in which a population of small training sets (individuals), drawn from the entire training set, was evolved with time to maximize the SVM classification score. Although it was shown to be very effective, the GA required lots of parameters (with the size of an individual t' being most important)

* Corresponding author. Tel./fax.: +48 32 237 21 51.
E-mail addresses: jakub.nalepa@polsl.pl (J. Nalepa), michal.kawulok@polsl.pl (M. Kawulok).

[1] The subset from which SVs are likely to be selected.

http://dx.doi.org/10.1016/j.neucom.2015.12.046
0925-2312/© 2015 Elsevier B.V. All rights reserved.

to be given prior to the optimization. Therefore, this GA had to be run multiple times in a time-consuming tuning process to find viable parameter values. We addressed this problem in our adaptive GA (AGA) [34], dynamically adaptive GA (DAGA) [21], and the most efficient—compared with the other evolutionary approaches—memetic algorithm (MASVM) [35]. In all of the mentioned evolutionary algorithms, the initial population of individuals was drawn from $\boldsymbol{T}$ randomly. It may very easily affect the search and convergence capabilities of the algorithm, particularly in the case of noisy datasets, since the “bad” vectors must be adaptively removed from individuals by means of genetic or memetic operators.

1.1. Contribution

In this paper, we propose a new adaptive memetic algorithm (PCA^{2}MA) to select training data for SVMs. Its crucial features include:

- Each dataset is preprocessed using principal component analysis (PCA) to exploit its geometrical properties and to extract potentially valuable vectors before the evolutionary optimization. They are used to (i) create the initial population, and to (ii) guide the evolution effectively towards high-quality refined training sets.
- A new adaptation scheme to control the size of refined sets and the selection scheme in PCA^{2}MA on the fly. The proposed scheme does not have any essential parameters itself (they would have to be tuned prior to the execution), which was a significant drawback of other adaptation procedures.

The main objective of PCA^{2}MA is to find the refined training sets which (i) will maximize the classification performance of SVMs trained using these sets, and (ii) will minimize the number of determined SVs, therefore will contain important vectors (as already mentioned, the SVM classification time depends linearly on the number of SVs established during the training). Additionally, for very large datasets, performing the SVM training is often impossible (due to its time and memory complexity). This issue is tackled in PCA^{2}MA, which adaptively evolves significantly smaller subsets of these datasets. It allows for finding the SVM decision hyperplane even for massively large real-life sets, for which the SVM training cannot be proceeded due to its memory complexity.

The following sections discuss the novel techniques proposed in this paper in more detail.

1.1.1. Data geometry analysis for preprocessing of datasets

In the proposed algorithm, the initial assessment of samples in $\boldsymbol{T}$ is executed as a preprocessing step before the evolutionary optimization. This procedure, which exploits PCA to investigate the data geometry, creates a set of potentially valuable (and “useful”) vectors, termed *candidate vectors* (CVs). They are later used to educate existing individuals and to introduce new ones in PCA^{2}MA. Also, we create the initial population by sampling CVs rather than the entire $\boldsymbol{T}$. We thus utilize both the *extracted* knowledge attained before the execution as well as the *historical* information (those samples from $\boldsymbol{T}$ that are determined as SVs during the evolution) gained on the fly for better exploration and exploitation of the search space.

1.1.2. Adaptation in PCA^{2}MA

In PCA^{2}MA, we establish a new adaptation scheme for controlling the algorithm parameters, including the size of an individual and the selection scheme, during the optimization. This scheme does not require any additional parameters, which was a drawback of MASVM. Also, it was unclear how to determine the initial individual size. In this paper, we propose to include 4 samples in each refined set at first—it is the minimum feasible value which avoids biasing (at least 3 vectors are necessary to define the hyperplane for two classes in the 2D linear case). The adaptation scheme is then applied to increase the size only if necessary. It exploits the current population characteristics to guide the search as best as possible. We aim at not only determining valuable subsets of $\boldsymbol{T}$ (SVMs trained using these sets perform extremely well for the unseen validation data), but also at decreasing the sizes of these sets as much as possible. Then, the number of SVs can be kept low (without affecting the classifier performance), which significantly decreases the SVM classification time. It is particularly important for massive-scale real-life problems embracing extremely large (and possibly noisy) datasets.

An extensive experimental study performed on a number of datasets, including benchmark, real-life, and artificially generated sets clearly confirms the efficacy of the proposed MA and its competitiveness compared with other state-of-the-art algorithms. We conducted a detailed sensitivity analysis to verify how various algorithm components influence its convergence capabilities and performance. Finally, we present the results of the two-tailed Wilcoxon tests applied to verify the statistical significance of the results.

1.2. Paper outline

The remainder of the paper is organized as follows. Section 2 reviews the state-of-the-art for handling large training sets, and highlights the current advances in adaptive evolutionary algorithms. In Section 3, we discuss in detail the proposed adaptive memetic algorithm. The results of an extensive experimental study, along with the sensitivity analysis on various method components, are analyzed in Section 4. Section 5 concludes the paper and summarizes directions of our ongoing work.

2. Related literature

2.1. Training SVMs from large datasets

In real-life classification problems, the cardinality of $\boldsymbol{T}$ can easily excess millions (or even more) samples, and training SVMs using the entire set becomes impossible due to its time and memory complexities ($O(t^3)$ and $O(t^2)$ respectively) in the case of a standard QP solver [37]. Similarly, if $\boldsymbol{T}$ is of a poor quality, contains lots of noisy (and/or mislabeled) samples, or is imbalanced, then feeding it to the SVM learning engine can not only result in a very time-consuming process, but also can worsen the SVM classification performance. Selecting a valuable subset from the entire $\boldsymbol{T}$ appeared therefore as a vital research topic due to its practical applicability, and a plethora of various approaches to tackle this problem emerged during the recent years.

The initial efforts towards dealing with large datasets were aimed at reducing the time complexity of the QP optimization by splitting it into subproblems [18]. Although it reduces the training time, the size of many real-world datasets is still too large to make this method applicable in practice. It concerns various domains, including bioinformatics, genomics, document categorization, and more [40]. Also, this approach does not solve the problem of a potential $\boldsymbol{T}$ noisiness.

There exist techniques approximating the answer of a non-linear kernel machine [38,39]. In such approaches, the input data are mapped into a low-dimensional randomized feature space. Then, linear learning methods are applied to estimate the answer of the corresponding non-linear kernel machine. Approximation of the decision function by means of multiplying the input with a Gaussian random matrix and applying non-linearity was the basis of the recently proposed *Fastfood* algorithm [24].

Only a small subset of vectors belonging to $\boldsymbol{T}$ (i.e., SVs) contributes to the position of the decision hyperplane. Thus, training SVMs with a subset of $\boldsymbol{T}$ containing only SVs will give at least as good results as training with the entire $\boldsymbol{T}$ (it may improve the SVM performance, if the noisy data from $\boldsymbol{T}$ are discarded). A simple—yet

Please cite this article as: J. Nalepa, M. Kawulok, Adaptive memetic algorithm enhanced with data geometry analysis to select training data for SVMs, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2015.12.046