The Smaller, the Better: Selecting Refined SVM Training Sets Using Adaptive Memetic Algorithm

Jakub Nalepa
Institute of Informatics
Silesian University of Technology
Akademicka 16
44-100 Gliwice, Poland
jakub.nalepa@polsl.pl

Michal Kawulok
Institute of Informatics
Silesian University of Technology
Akademicka 16
44-100 Gliwice, Poland
michal.kawulok@polsl.pl

ABSTRACT

Support vector machine (SVM) is a supervised classifier which has been applied for solving a wide range of pattern recognition problems. However, training of SVMs may easily become a bottleneck, because of its time and memory requirements. Enduring this issue is a vital research topic, especially in the era of big data. In this abstract, we present our adaptive memetic algorithm for selection of refined (significantly smaller) SVM training sets. The algorithm—being a hybrid of an adaptive genetic algorithm and some refinement procedures—exploits the knowledge about the training set vectors extracted before the evolution, and attained dynamically during the search. The results obtained for several real-life, benchmark, and artificial datasets showed that our approach outperforms the other state-of-the-art techniques, and is able to extract very high-quality SVM training sets.

CCS Concepts

•**Computing methodologies** → **Support vector machines; Bio-inspired approaches;** *Artificial intelligence;*

Keywords

SVM; training set selection; memetic algorithm; PCA; adaptation

1. INTRODUCTION

The process of the SVM training involves solving a constrained quadratic programming optimization problem of $O(t^3)$ time and $O(t^2)$ memory complexity, where t denotes the cardinality of the training set $\boldsymbol{T}$. It entails selecting important $\boldsymbol{T}$ vectors, referred to as *support vectors* (SVs). These vectors define the position of the decision hyperplane, which is later used to classify the incoming (unseen) data. Additionally, the classification time linearly depends on the number of SVs (s). Hence, minimizing s (ideally without affecting the classification performance) is of a great importance too, especially in the case of real-time applications.

There are two main streams of research towards dealing with the time-consuming SVM training. The first one encompasses techniques which aim at enhancing the training process itself, whereas the other one includes algorithms which retrieve subsets of $\boldsymbol{T}$ (denoted as $\boldsymbol{T'}$'s) for feeding them into the training procedure. Since SVs constitute a subset of $\boldsymbol{T}$, training SVMs with $\boldsymbol{T'}$'s containing only SVs may be expected to give the same model as if they are trained with the entire $\boldsymbol{T}$. It is beneficial to extract significantly smaller refined sets—this will not only speed up the training, but can also improve the classifier performance (if the "misleading", e.g., mislabeled, vectors are removed from $\boldsymbol{T}$), and decrease the classification time (if s is reduced).

In our very recent work published in the *Neurocomputing* journal [2], we proposed the adaptive memetic algorithm (abbreviated as PCA^{2}MA—**P**rincipal **C**omponent **A**nalysis **A**daptive **M**emetic **A**lgorithm) for selection of refined training sets for SVMs. In PCA^{2}MA, the information about potentially important vectors from $\boldsymbol{T}$ is extracted before the evolution in the pre-processing step, and is coupled with the knowledge attained during the optimization to better guide the search towards high-quality refined sets. Thanks to the efficient, parameter-less adaptation techniques applied in PCA^{2}MA, this algorithm dynamically updates its crucial parameters (the size of refined sets being the most important) on the fly, thus does not require undertaking the time-consuming tuning process. The comprehensive experiments performed on several real-life, benchmark, and artificial datasets revealed that PCA^{2}MA significantly outperforms not only our previous evolutionary approaches (including the memetic algorithm presented at GECCO '14 [1]), but also other state-of-the-art training set selection techniques belonging to various groups of algorithms. The experimental study included visualizations of retrieved $\boldsymbol{T'}$'s along with the annotated SVs, and the statistical tests (two-tailed Wilcoxon tests) to prove the significance of the results.

2. THE ALGORITHM

In PCA^{2}MA, the population of individuals (each representing a refined training set) evolves in time. This evolution is preceded by the pre-processing step, in which the input space is transformed using principal component analysis to extract most discriminating dimensions. Each dimension is then divided into intervals in such a way, that the intervals contain the same number of $\boldsymbol{T}$ vectors. The geometry

GECCO'16 Companion July 20-24, 2016, Denver, CO, USA

ACM ISBN 978-1-4503-4323-7/16/07.

DOI: http://dx.doi.org/10.1145/2908961.2930950

of these vector groups is further analyzed to form the set of *candidate vectors* (CVs), which will contain as heterogeneous $\boldsymbol{T}$ vectors as possible. The CVs are exploited at various PCA^{2}MA steps, to (i) create the initial population, and to (ii) guide the evolutionary process.

During the evolution, the individuals undergo the genetic operations (selection, crossover, and mutation), as well as the memetic ones. The latter operations encompass the education, in which "weak" vectors from an individual are replaced with those residing in the support vectors pool. The pool gathers the vectors which were appended to at least one chromosome, and were selected as SVs at some point during the optimization. Hence, they can be considered as important vectors (they influenced the position of the SVM decision hyperplane). The pool is also utilized to create *super individuals* (chromosomes composed of the so-far-determined SVs only). Determining the fitness of an individual consists in training an SVM using the corresponding $\boldsymbol{T'}$ (containing t' vectors, where $t' \ll t$), and quantifying its performance for the set $\boldsymbol{T}$—the area under the receiver operating characteristic curve (AUC) is used for this purpose.

Since the appropriate size of the refined sets (thus the size of the individuals) is not known *a priori*, it is dynamically updated during the evolution (along with the selection scheme). The process of updating t' is parameter-less, and is based on investigating the current (and desired) population characteristics only. This allowed for responding to the current search state and status very efficiently.

3. THE RESULTS

The experiments presented in [2] involved investigating the PCA^{2}MA performance for real-life, benchmark, and artificial datasets[1], and were divided into three groups: (i) the sensitivity analysis, the comparison with (ii) our previous evolutionary algorithms, and (iii) state-of-the-art techniques.

The results of the sensitivity analysis (in which various PCA^{2}MA variants were examined) revealed that the analysis of the $\boldsymbol{T}$ geometry (in the pre-processing step) is pivotal, and helps extract important training set vectors. Utilizing them allowed for improving the convergence capabilities of PCA^{2}MA. The new adaptation was shown much more efficient than our previous technique [1]. The PCA^{2}MA variant with both enhancements (pre-processing and the new adaptation) notably outperformed the other algorithm variants—the retrieved refined sets were of a higher quality.

The comparison with other evolutionary techniques proved that PCA^{2}MA is the best evolutionary algorithm for selection of $\boldsymbol{T'}$'s. SVMs trained with refined sets delivered by PCA^{2}MA gave the largest AUC values for the validation set. At the same time, the number of SVs was kept very small without affecting the classifier performance. It indicates that PCA^{2}MA selects most important $\boldsymbol{T}$ vectors, and refrains from adding other vectors unnecessarily. Importantly, the adaptive memetic algorithm was able to effectively balance the exploitation of smaller refined sets with the exploration of the larger ones. This balance is crucial in the case of challenging datasets, in which exploiting small $\boldsymbol{T'}$'s may not boost the classification performance due to a too small number of SVs. Finally, PCA^{2}MA appeared quite stable and led to very similar results in numerous independent algorithm runs for the investigated datasets.

The performance of PCA^{2}MA was also compared with the selected state-of-the-art techniques, belonging to various groups of algorithms (randomized approaches and the one analyzing the geometrical properties of $\boldsymbol{T}$). Additionally, SVMs were trained using the entire $\boldsymbol{T}$'s too. In most cases, the average AUC values obtained using SVMs trained with refined sets (the AUC values were calculated for the validation sets) retrieved using PCA^{2}MA were larger than the best AUC values elaborated using other techniques. The sets obtained using PCA^{2}MA contained significantly less samples compared with other approaches. It allowed for decreasing the number of SVs without lowering the classification score. We analyzed the convergence time of the proposed algorithm (i.e., the time after which the fitness of the best individual in the population could not be further improved) and showed, that it could be safely terminated faster. It is not possible in "one-pass" algorithms, which investigate the entire set $\boldsymbol{T}$ in order to extract its valuable subset. The results proved the stability of PCA^{2}MA. It is worth mentioning that the SVM training appeared impossible for massively-large real-life dataset (when the entire $\boldsymbol{T}$ was fed into the training procedure) due to its memory requirements.

Finally, the refined sets (along with the selected SVs) were visualized for all artificial sets. The investigation revealed that the refined sets obtained using PCA^{2}MA do not necessarily follow geometrical patterns. Albeit the pre-processing examines the geometry of $\boldsymbol{T}$, the vectors which were not annotated as CVs in this process can still be accessed and appended to individuals during the evolution. In many cases, it turned out that utilizing these vectors (which appeared not important at the first glance) allowed for boosting the SVM classification performance. This would not be possible for other methods, in which the rejected $\boldsymbol{T}$ vectors cannot be later re-analyzed and added to $\boldsymbol{T'}$'s.

4. CONCLUSIONS

In our recent paper [2], we proposed an adaptive memetic algorithm for selection of refined training sets for SVMs. This approach enhanced the evolutionary optimization of the refined sets using additional procedures exploiting the knowledge concerning the $\boldsymbol{T}$ vectors extracted before the execution (in the pre-processing step), and attained dynamically during the search. The crucial PCA^{2}MA parameters were dynamically controlled on the fly to adapt to the search progress using the parameter-less adaptation schemes. The extensive experimental study revealed that PCA^{2}MA is superb compared with other evolutionary techniques, and several algorithms from the literature. It allowed for (i) retrieving very high-quality $\boldsymbol{T'}$'s, (ii) decreasing s, and (iii) improving the process of selecting $\boldsymbol{T'}$'s. Finally, we showed that smaller training sets can be notably better than the larger ones, if they exclude mislabeled and noisy vectors from $\boldsymbol{T}$.

5. REFERENCES

[1] J. Nalepa and M. Kawulok. A memetic algorithm to select training data for support vector machines. In *Proceedings of GECCO '14*, pages 573–580. ACM, 2014.

[2] J. Nalepa and M. Kawulok. Adaptive memetic algorithm enhanced with data geometry analysis to select training data for SVMs. *Neurocomputing*, 185:113 – 132, 2016.

[1]The real-life and artificial datasets can be found at: http://sun.aei.polsl.pl/~jnalepa/SVM