Ensembles of support vector machines with evolutionarily optimized hyperparameters and training sets

Silesian University of Technology
Faculty of Automatic Control, Electronics and Computer Science

mgr inż. Wojciech Dudzik

Supervisor: **dr hab. inż. Michał Kawulok, prof. PŚ**

Assistant supervisor: **dr hab. inż. Jakub Nalepa, prof. PŚ**

Abstract

Support vector machines (SVMs) are widely used for binary classification, with a proven track record of high accuracy and robustness in various domains, including bioinformatics, economy, and software engineering. They work by finding the hyperplane that separates two classes of data points while maximizing the margin between the hyperplane and the closest data points. These closest data points are often selected as support vectors that determine the decision boundary (hyperplane) of the SVM classifier and are crucial for the classification process of new data points. This approach leads to a robust and well-generalizing classifier, which can handle high-dimensional and noisy data. However, the major drawback of SVMs is that they are computationally expensive to train. The training process has time and memory complexities of $O(t^3)$ and $O(t^2)$, respectively, where t is the size of the training set. Additionally, the choice of kernel function can have a significant impact on the performance of the classifier.

Despite these challenges, SVMs continue to play an important role in the field of machine learning. Many researchers have proposed various methods to improve the scaling properties of SVMs, such as accelerating the training process by better-utilizing hardware or limiting the size of the training set by selecting the most promising vectors. However, as shown in the literature review the problems of hyperparameters optimization and the selection of training and feature sets are in majority of cases addressed independently of each other, which can negatively affect the performance of the SVM classifier.

This dissertation proposes effective solutions for optimizing SVM models in the context of binary classification. Algorithms that utilize the mutual dependence of training set selection and hyperparameter optimization are proposed and validated. Furthermore, the dissertation introduces new methods for constructing ensembles of SVM models in

a form of cascades to enhance their classification performance and extend their capabilities while keeping the training process fast. To achieve these goals, the proposed methods are based on evolutionary computations, which have already proved to be a robust solution to many problems.

The analyzed methods include the existing evolutionary techniques for training set selection which have the drawback of requiring the SVM hyperparameters to be tuned beforehand. To mitigate this issue the methods that combine hyperparameter optimization in an alternating scheme are introduced. What is more, a method for simultaneous optimization of hyperparameters, training set, and feature set called SE-SVM is proposed. This algorithm is also extended to utilize a new adaptive RBF kernel. Those algorithms were proposed to verify the first hypothesis that: Simultaneous optimization of the training set and the SVM hyperparameters improve training and classification time compared to other state-of-the-art methods proposed for this purpose without affecting the classification quality. The ensemble scheme is building a cascade of SVMs and should provide the best classification quality with Pareto-optimal training and classification times. These cascades are being built by utilizing a previously presented algorithm SE-SVM. They aim to verify the second hypothesis that: SVM ensembles created using evolutionary algorithms provide improved classification performance compared to other well-established methods, including existing algorithms for building SVM ensembles.

All of these techniques are experimentally analyzed and compared to other popular classifiers and state-of-the-art SVM optimization techniques using artificial 2D datasets and benchmark datasets. The analysis includes a qualitative analysis of visualized decision boundaries on a 2D dataset. Using those datasets the new concept of certain regions in ensemble classifiers (in a form of cascades) is introduced. Moreover, the quantitative analysis is performed on both 2D datasets and a set of 31 benchmark datasets. All of those results are also analyzed using statistical tests.

The results show that the simultaneous optimization approach (SE-SVM method) provides a great reduction in training and classification times compared to other methods. This positively verifies the first hypothesis stated in the work. What is more, in both cases of 2D datasets and benchmark datasets the presented ensemble methods (ECE-SVM) outperform all other well-established methods, including existing algorithms for building SVM ensembles. These results positively verify the second hypothesis. Although the ECE-SVM approach may require more computational resources (compared to SE-SVM), the potential gains in classification quality and confidence make it a valuable technique for further usage. It is shown that both SE-SVM and ECE-SVM lie on Pareto-front when classification quality is plotted against either classification or training time.