Training Multi-class Support Vector Machines

Ürün Dogan - Institut für Neuroinformatik / Ruhr University Bochum

Abstract

Support Vector Machine (SVM) algorithms are state-of-the-art approaches in pattern recognition. Originally proposed for binary classification tasks, they do not imply a single, unequivocal extension to multiple classes, and several multi-class SVMs have been proposed in the literature. This thesis focuses on stringent theoretical, algorithmic and experimental treatment of multi-class SVMs. Postulating as a common desirable property an argmax decision rule, I include five existing multi-class SVM formulations in my analysis, four of which are all-together methods in the sense that they solve a single single quadratic optimization program. For experimental comparison, I additionally consider the well-known one-versus-all approach, which solves a series of binary classification problems. More importantly, I provide a detailed theoretical analysis of all-together multi-class formulations, in particular the variants proposed by Lee, Lin & Wahba (LLW); by Weston & Watkins (WW); by Crammer & Singer (CS); as well as Multi-Class Classification with Maximum Margin Regression proposed by Szedmak, Shawe-Taylor, & Parado-Hernandez (MC-MMR). I present a unified view on these four approaches, in which they differ along three conceptual dimensions. Rather than merely leading to a simplified taxonomy, this unified view allows us to make substantial progress on several theoretical, algorithmic, and implementation-related aspects in multi-class SVM optimization. For example, I propose a unifying template for the primal and dual optimization problems arising when training the aforementioned all-in-one machines. It moreover becomes apparent that one possible combination of concepts has to date been missing in the literature so far. I derive this novel multi-class SVM variant (DGI), which shares concepts used in the CS and LLW formulations. Like the CS machine, DGI only uses a single slack variable per training example, but adopts the margin concept of the LLW formulation. In practice, the WW and LLW machines have been implemented and used less often than their CS counterpart, despite desirable theoretical properties. This can be ascribed to the fact that no fast, efficient training algorithms have so far been proposed in the literature. I fill this gap by presenting a new decomposition algorithm for training multi-class SVMs, in particular for the LLW, WW and DGI machines. The new decomposition method accelerates SVM training by considering hypotheses without bias term. Further, I propose to use working sets of size two instead of following the sequential minimal optimization (SMO) paradigm. This is complemented by a second order working set selection scheme. Consequently, I for the first time provide the prerequisites necessary for competitive implementations of the WW and LLW machines, in the sense of also being fast enough for carrying out SVM model selection. On that basis, I conduct extensive empirical comparison of the five different multi-class SVMs mentioned above, plus the DGI machine. The experimental evidence confirmed that the two-variable decomposition algorithm outperforms standard SMO. The LLW SVM performed best in terms of accuracy, at the cost of longer training times. Compared to the CS machine, WW yielded better generalizing hypotheses and did not require longer training times. This holds direct implications for practitioners, who should, depending on the problem size, prefer either LLW or WW over the CS machine. Motivated by these experimental results, I conduct further theoretical analysis of the generalization risk in the WW and CS machines. I am able to show that an equally derived risk bound is lower for the WW formulation.