SHOGUN - A Large Scale Machine Learning Toolbox

Sören Sonnenburg[†], Fabio De Bona[♮],Gunnar Rätsch[♮]

[†] Fraunhofer Institut FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany

[♮] Friedrich Miescher Laboratory, Spemannstr. 35, 72076 Tübingen, Germany

Soeren.Sonnenburg@first.fraunhofer.de,

{Gunnar.Raetsch,Fabio.De.Bona}@tuebingen.mpg.de

Abstract

We have developed an R Interface for our Machine Learning Toolbox SHOGUN. It features algorithms to train hidden markov models and learn regression and 2-class classification problems. While the toolbox's focus is on kernel methods such as Support Vector Machines, it also implements a number of linear methods like Linear Discriminant Analysis, Linear Programming Machines and Perceptrons.

It provides a generic SVM object interfacing to *seven* different SVM implementations, among them the state of the art LibSVM[1] and SVM^{light}[2]. Each of these can be combined with a variety of kernels. The toolbox not only provides efficient implementations of the most common kernels, like the Linear, Polynomial, Gaussian and Sigmoid Kernel but also comes with a number of recent string kernels as e.g. the Spectrum or Weighted Degree Kernel (with shifts). For the latter the efficient linadd[4] optimizations are implemented. Also SHOGUN offers the freedom of working with custom pre-computed kernels.

One of its key features is the "combined kernel" which can be constructed by a weighted linear combination of a number of sub-kernels, each of which not necessarily working on the same domain. An optimal sub-kernel weighting can be learned using Multiple Kernel Learning.[3]

The input feature-objects can be dense, sparse or strings and of type int/short/double/char and can be converted into different feature types. Chains of "preprocessors" (e.g. substracting the mean) can be attached to each feature object allowing for on-the-fly pre-processing.
SHOGUN also supports Matlab™, Octave and Python-numarray. The Source Code is freely available for academic non commercial use under http://www.fml.mpg.de/raetsch/shogun.

References

[1] C.-C. Chang and C.-J. Lin. Libsvm: Introduction and benchmarks. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2000.

[2] T. Joachims. Making large–scale SVM learning practical. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 169–184, Cambridge, MA, 1999. MIT Press.

[3] S. Sonnenburg, G. Rätsch, S. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 2006. accepted.

[4] Sören Sonnenburg, Gunnar Rätsch, and Bernhard Schölkopf. Large scale genomic sequence SVM classifiers. In *Proceedings of the 22nd International Machine Learning Conference*. ACM Press, 2005.