SHOGUN - A Large Scale Machine Learning Toolbox

Sören Sonnenburg[1,2,*], Gunnar Rätsch[2], Sebastian Henschel[2], Christian Widmer[2], Jonas Behr[2], Alexander Zien[2,5], Fabio de Bona[2], Christian Gehl[3], Alexander Binder[3], Vojtech Franc[4]

1. Machine Learning Group, TU-Berlin, Franklinstr. 28/29, 10587 Berlin, Germany
2. Friedrich Miescher Laboratory, Spemannstr. 35, 72076 Tübingen, Germany
3. Fraunhofer Institut FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany
4. Center for Machine Perception, Technicka 2, 166 27 Praha 6, Czech Republic
5. LIFE Biosystems GmbH, Poststr. 34, 69115 Heidelberg, Germany

*Contact author: Soeren.Sonnenburg@tu-berlin.de

Keywords: Machine Learning, Large Scale, Support Vector Machines, Kernels, Open-Source

We have developed R-bindings for our machine learning toolbox SHOGUN, which features algorithms for hidden markov models, regression and classification problems. SHOGUN's focus is on Support Vector Machines, but also implements a number of linear methods like Linear Discriminant Analysis, Linear Programming Machines and Perceptrons. It provides a generic SVM interface enabling the choice between *fifteen* different SVM optimizers as back-ends, among them the state of the art LibSVM[1] and SVM^{light}[4], SVMOcas [3] or Liblinear [2]. The SVMs can be easily combined with more than 35 different kernel functions (see http://www.shogun-toolbox.org/doc). Moreover, it offers options for using precomputed kernels and easily allows the integration of custom kernels. One of SHOGUN's key features is the *combined kernel* to construct weighted linear combinations of multiple kernels that may even be defined on different input domains and learned using Multiple Kernel Learning (MKL) algorithms, e.g., [5, 7]. Input feature objects can be dense or sparse vectors of strings, integers (8, 16, 32 or 64 bit; signed or unsigned), or floating point numbers (32 or 64 bit), and can be converted into different feature types. Chains of "pre-processors" (e.g., subtracting the mean) can be attached to each feature object allowing on-the-fly pre-processing. Finally, several commonly used performance measures for evaluation (e.g., area under ROC) are implemented.

A central aspect in the design of SHOGUN was to enable large-scale learning. We implemented auxiliary routines that allow faster computation of combinations of kernel elements that lead to significant speedups during training and evaluation [6] enabling us to solve several large-scale learning problems in biological sequence analysis [3, 6, 8] involving millions of sequences. Furthermore, linear SVMs can be efficiently trained by computing feature spaces on-the-fly, even allowing to mix sparse, dense and other data types.

All of SHOGUN's core functions are encapsulated in a library (libshogun) and are easily accessible and extendible by C++ application developers. Built around SHOGUN's core we provide two types of R interfaces: A modular interface created with SWIG[1], and a static interface. However, note that interfaces are available from C++ to other languages, such as *Python*, *Octave* and *Matlab*^{TM}. SHOGUN's source code is freely available under the GNU General Public License at http://www.shogun-toolbox.org.

References

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

[2] R. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin, *LIBLINEAR: A library for large linear classification*, Journal of Machine Learning Research **9** (2008), 1871–1874.

[3] V. Franc and S. Sonnenburg, *Optimized cutting plane algorithm for large-scale risk minimization*, Journal of Machine Learning Research (2009).

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

[5] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien, *Efficient and accurate l_p-norm multiple kernel learning*, Advances in Neural Information Processing Systems 21, MIT Press, Cambridge, MA, 2010.

[6] S. Sonnenburg, G. Rätsch, and K. Rieck, *Large-scale learning with string kernels*, Large-Scale Kernel Machines (L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, eds.), MIT Press, 2007, pp. 73–103.

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

[8] S. Sonnenburg, A. Zien, and G. Rätsch, *ARTS: Accurate recognition of transcription starts in human*, Bioinformatics **22** (2006), no. 14, e472–480.

[1] The Simplified Wrapper and Interface Generator, cf. http://www.swig.org.