Efficient Ensemble Learning With Support Vector Machines

http://esat.kuleuven.be/stadius/ensemblesvm/

Marc Claesen[†] MARC.CLAESEN@ESAT.KULEUVEN.BE
Frank De Smet[‡,⋆] FRANK.DESMET@CM.BE
Johan A.K. Suykens[†] JOHAN.SUYKENS@ESAT.KULEUVEN.BE
Bart De Moor[†] BART.DEMOOR@ESAT.KULEUVEN.BE

† KU Leuven, Department of Electrical Engineering, ESAT – STADIUS/iMinds Medical IT
Kasteelpark Arenberg 10 box 2446, 3001 Leuven, Belgium

‡ KU Leuven, Department of Public Health and Primary Care, Environment and Health
Kapucijnenvoer 35 blok d box 7001, 3000 Leuven, Belgium

⋆ National Alliance of Christian Mutualities, Medical Management Department

Keywords: ensemble learning, support vector machine, classification, free software

We present `EnsembleSVM`, a free software package containing efficient routines to perform ensemble classification with support vector machine (SVM) base models (Claesen et al., 2014). The `EnsembleSVM` software is implemented in `C++11` and is licensed under the GNU LGPL. The software is freely available at `http://esat.kuleuven.be/stadius/ensemblesvm/`.

Ensembles of SVM models have merits in several segments of machine learning, including large-scale learning, semi-supervised learning and online learning. In large-scale learning, using ensemble approaches allows base models to be trained on small subsamples, which drastically decreases training time when using nonlinear kernels. Experiments show that nonlinear ensembles can maintain high predictive accuracy while being much faster to train than traditional SVMs. In a supervised context, ensembles are particularly useful for large-scale problems where the predictive performance of linear models is unsatisfactory and standard kernel SVMs take too long to train. This is the case in low-dimensional learning problems with many available training instances.

`EnsembleSVM` provides a set of tools and a full programming API. Base models are trained using the `LIBSVM` software (Chang & Lin, 2011). A range of aggregation schemes is provided, both linear and nonlinear. Duplicate storage and kernel evaluation of support vectors which are shared between constituent models is avoided, resulting in a smaller memory footprint and faster prediction than implementations involving wrappers. Both training and prediction are parallelized in the tools provided by the library.

Acknowledgments

MC is a PhD student at KU Leuven. JS is a professor at KU Leuven. BDM is a full professor at KU Leuven. Research supported by:

- Research Council KU Leuven: GOA/10/09 MaNet, KUL PFV/10/016 SymBioSys, PhD/Postdoc grants,
- Industrial Research fund (IOF): IOF/HB/13/027 Logic Insulin,
- Flemish Government: FWO: projects: G.0871.12N (Neural circuits); PhD/Postdoc grants; IWT: TBM-Logic Insulin(100793), TBM Rectal Cancer(100783), TBM IETA(130256); PhD/Postdoc grants; Hercules Stichting: Hercules 3: PacBio RS, Hercules 1: The C1 single-cell auto prep system, BioMark HD System and IFC controllers (Fluidigm) for single-cell analyses; iMinds Medical Information Technologies SBO 2014; VLK Stichting E. van der Schueren: rectal cancer,
- EU: ERC AdG A-DATADRIVE-B.

References

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*, 27:1–27:27.

Claesen, M., Smet, F. D., Suykens, J. A., & Moor, B. D. (2014). EnsembleSVM: A library for ensemble learning using support vector machines. *Journal of Machine Learning Research*, *15*, 141–145.