Fisher Kernels and Logical Sequences with an Application to Protein Fold Recognition

Kristian Kersting
Machine Learning Lab, Computer Science Department,
University of Freiburg
Georges-Koehler-Alle 079, 79112 Freiburg, Germany
kersting@informatik.uni-freiburg.de

Thomas Gärtner
Department of Computer Science
University of Bristol
The Merchant Venturers Building
Woodland Road, Bristol BS8 1UB
United Kingdom

Knowledge Discovery
Fraunhofer Institut
Autonome Intelligente Systeme
Schloss Birlinghoven
53754 Sankt Augustin, Germany

thomas.gaertner@ais.fraunhofer.de

Generative models in general and hidden Markov models (HMM) [10] in particular are widely used in computational biology. One of their application areas there is protein fold recognition [2] where one tries to understand how proteins fold up in nature. Despite of their success HMMs have some weaknesses: (1) their predictive accuracy is usually lower than that of discriminative classifiers, and (2) they are able to handle sequences of flat, i.e., unstructured, symbols only.

Fisher Kernels [6, 5] were developed to improve the classification accuracy of generative models. The key idea is to use the gradient of the log likelihood with respect to the parameters of a generative Model as the features in a discriminative classifiers. Usually, for the generative model a HMM is used and for the discriminative classifier a support vector machine [1] is used. This has led to improved classification results in several domains were instances are flat sequences.

Many sequences occurring in real-world problems, however, exhibit internal structure. The elements of such sequences can be seen as atoms in a first order logic (see e.g. [9] for an introduction to logic programming). For example, the secondary structure of the Ribosomal protein L4 can be represented as

$$\begin{gathered} st(null,2).he(h(right,alpha),6).st(plus,2).he(h(right,alpha),4).st(plus,2). \\ he(h(right,alpha),4).st(plus,3).he(h(right,alpha),4).st(plus,1).he(h(right,alpha),6). \end{gathered}$$

Here, helices of a certain type and length, i.e., $he(HelixType, Length)$, and strands of a certain orientation and length, i.e., $st(Orientation, Length)$ are essentially structured symbols, namely atoms over logical predicates. It has been argued that using such secondary structure information is likely to improve fold recognition results, see e.g. [3]. The application of HMMs to such sequences requires one to either ignore the structure of helices and strands, which results in a loss of information, or to take all possible combinations (of arguments such as orientation and length) into account, which leads to a combinatorial explosion in the number of parameters.

Logical hidden Markov models (LOHMMs) [7] have recently been introduced as an extension of HMMs that allows for *logical sequences*, i.e., sequences of atoms in a first order logic. The main idea of LOHMMs is to use abstract symbols and states instead of specific symbols and states. These abstractions, obtained by summarizing sets of states, are represented as logical atoms and allow for the use of (logical) variables as well as unification. In [8], LOHMMs have been applied to the problem of discovering structural signatures of protein folds and led to more compact models. The trained LOHMM consisted of 120 parameters corresponding to an HMM with more than 62000 parameters.

The present work investigates whether the predictive accuracy of LOHMMs can be improved using Fisher kernels and support vector machines. For that, we devise a method to compute the gradient of the log likelihood with respect to the parameters of a LOHMM. Our empirical evaluation compares the accuracy achieved in [8] with the accuracy achieved by Fisher kernels and support vector machines using the same LOHMMs and datasets as in [8]. The data consists of logical sequences of the secondary structure of protein domains. The task is to predict one of five SCOP [4] folds for 2187 test sequences given a model trained on 200 sequences per fold. For that, we first computed the gradients for each fold, and then trained support vector machines with Fisher kernels. In our empirical evaluation, we obtained a substantial improvement in the identification of protein folds over the accuracy achieved by LOHMMs alone. The predictive accuracy increased from 74% to 82.7% where the precision and recall values were well balanced within each fold.

Summing up, we believe that *Fisher kernels of logical hidden Markov models* will become a useful tool for the classification of structured sequence data in general and for protein fold recognition in particular.

References

[1] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifier. In *Proceedings of the fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.

[2] J. Gough, K. Karplus, R. Hughey, and C. Chothia. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *JMB*, 313(4):903–919, 2001.

[3] J. Hargbo and A. Elofsson. Hidden markov models that use predictied secondary structure for fold recognition. *Proteins: Structure, Function, and Genetics*, 36:68–76, 1999.

[4] T. Hubbard, A. Murzin, S. Brenner, and C. Chotia. *SCOP*: a structural classification of proteins database. *NAR*, 27(1):236–239, 1997.

[5] T. Jaakkola, M. Diekhans, , and D. Haussler. Using the fisher kernel method to detect remote protein homologies. In *Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 149–158, 1999.

[6] T.S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*. MIT Press, 1999.

[7] K. Kersting, T. Raiko, and L. De Raedt. Logical Hidden Markov Models (Extended Abstract). In *Proceedings of the First European Workshop on Probabilistic Graphical Models (PGM-02)*, Spain, November 2002.

[8] K. Kersting, T. Raiko, S. Kramer, and L. De Raedt. Towards discovering structural signatures of protein folds based on logical hidden markov models. In *Proceedings of the Pacific Symposium on Biocomputing*, 2003. (to appear).

[9] J. W. Lloyd. *Foundations of Logic Programming*. Springer, 2. edition, 1989.

[10] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.