Secondary structure prediction by Inductive Logic Programming

Igor Mozetič
Center for Applied Mathematics and Theoretical Physics
University of Maribor
Krekova 2, 2000 Maribor, Slovenia
igor.mozetic@uni-mb.si

November 1998

Inductive Logic Programming (ILP) is an approach to learning from examples. One specifies learning examples in terms of attributes (continuous or discrete) and the result of learning are rules in human-readable form. We have implemented ALF - an ILP system designed to learn from sequential data. ALF is an extension of the well-known learning program C4.5 [1] which produces rules in the form of decision trees. The main advantage of ALF is its flexibility and the ability to efficiently handle large quantities of data (in the order of 100.000 learning examples and 1000 attributes). Flexibility allows the user to easily extend the set of attributes (e.g., by extending the window size), introduce new attributes (e.g., hydrophobicity, polarity, size of residues), or even use previously learned rules as new attributes. The main difference to the neural network or nearest neighbour learning is that the result of learning are explicit rules. A possible disadvantage is that in the rules, only a relevant, non-redundant subset of attributes is used and therefore redundant information cannot be sufficiently exploited.

While ALF is a general purpose system for learning from sequential data, we have applied it to the protein secondary structure prediction problem. We have taken a subset of the PDB as specified by the June 1998 release of PDBselect ([2], 25% sequence identity threshold) for which DSSP [3] assigned secondary structure without errors. The subset comprises 887 sequence unique protein chains with 189.718 residues. For all the residues we first obtained the PHD server [4, 5] secondary structure assignments with reliability index R (0-9). We formed attributes with values of PHD predictions for H if R $\geq$ 3, L if R $\geq$ 4, E if R $\geq$ 5, and unassigned otherwise. We took the window size of $\pm$ 6 residues, and additionally formed new attributes in terms of the distance to the nearest H, L, or E residue. In order to prevent overfitting the data, the reduced-error pruning [1] with confidence level CF = 1% was used for learning the rules. 7-fold cross validation on the set of 887 chains yielded accuracy Q3 = 72.1 $\pm$ 9.7% and segment overlap SOV [Zemla & Venclovas] = 67 $\pm$ 14%. Incremental learning from the whole dataset was then used to submit predictions for the CASP3 targets.

References

[1] Quinlan,J.R. (1993) C4.5: Programs for machine learning, Morgan Kaufmann.

[2] Hobohm,U. & Sander,C. (1994) Enlarged representative set of protein structures, Protein Science 3, 522.

[3] Kabsch,W. & Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features, Biopolymers 22, 2577-2637.

[4] Rost,B. & Sander,C. (1993) Prediction of protein secondary structure at better then 70

[5] Rost,B. & Sander,C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure, Proteins 19, 55-72.