Learning Type Extension Trees for Metal Bonding State Prediction

Paolo Frasconi[1] and Manfred Jaeger[2] and Andrea Passerini[3]

[1] Dipartimento di Sistemi e Informatica, Università degli Studi di Firenze, Italy
[2] Department for Computer Science, Aalborg University, Denmark
[3] Dipartimento di Ingegneria e Scienza dell'Informazione, Università degli Studi di Trento, Italy

Abstract. Type Extension Trees (TET) have been recently introduced as an expressive representation language allowing to encode complex combinatorial features of relational entities. They can be efficiently learned with a greedy search strategy driven by a generalized relational information gain and a discriminant function. In predicting the metal bonding state of proteins, TET achieve significant improvements over manually curated motifs, and the expressiveness of combinatorial features significantly contributes to such performance. Preliminary collective classification results seem to indicate it as a promising direction for further research.

1 Learning Type Extension Trees

A TET [1] consists of a tree-structured logic formula where nodes are conjunctions of literals, and edges are labeled with sets of variables. Instead of a simple truth assignment, a TET defines a complex combinatorial feature whose recursive value structure accounts for the number of times each subtree can be satisfied, given the possible bindings of its edge variables. A simple discriminant function [2] can be defined over TET as a kind of pseudo maximum-likelihood ratio, and TET structure learning have been addressed [2] with a recursive top-down strategy. The algorithm basically generates a set of candidate extensions of the current TET, according to some pre-specified language bias as standard in ILP methods, but instead of directly evaluating each of them, it relies on a generalized infogain criterion [2] to select the most promising directions for further expansion. Generalized infogain aims at measuring both *direct* and *potential* informativeness of a certain extension, the latter being conditioned on appropriate refinements further down in the tree structure, thus performing a kind of selective lookahead strategy [3]. The discriminative power of a whole subtree is eventually evaluated once a certain pre-specified depth has been reached, or the score of further expansions falls below a given threshold (see [2] for details).

Supervision on related entities can be introduced in TET by allowing target predicates to be included as candidate extensions (with proper constraints to avoid trivial TET with the target predicate instantiated with target variables). A simple iterative procedure can be also conceived in order to implement a collective classification approach: first, a bootstrap TET is learned without using target predicates, and its predictions are

used to initialize the labels of all instances; second a collective TET is learned by including target predicates in the language bias. During testing, the bootstrap TET initializes predictions, and the collective TET iteratively refines them until no relabeling occurs between two successive iterations, or a maximum number of iterations is reached.

2 Metal Bonding State Prediction

Metal ions play a central role in living organisms, performing structural, catalytical and regulatory functions in the cell. About one third of the known proteins is believed to bind metal ions in their native conformation. Metal binding sites are quite specific in terms of number and type of coordinating residues: CYS and HIS are the most common ligands, followed by GLU and ASP which are however much more frequent in proteins, and each ion has few preferred coordination numbers, ranging from one to eight. Regularities in terms of number, type, and distance of ligands and surrounding residues have been encoded by biologists in motifs [4], either regular expressions or position-specific profiles with amino acid weights and gap costs. Such motifs provide interpretable features characterizing metal binding sites, but their performance are far below those of complex machine learning approaches employing multiple alignment profiles [5]. Preliminary experiments [2] showed that TET are able to significantly improve over manually curated motifs while retaining much of their interpretability, and that counts-of-counts features significantly contribute to such improvements, as shown by comparisons to results obtained learning standard regular-expression like TET and to those achieved by the Tilde ILP system. The dataset and the 5-fold cross validation procedure used in the experiments were taken from [5]. Residue attributes made available to TET learning consist of the binarized evolutionary conservation of either relevant residue types such as CYS, HIS, ASP, GLU or PRO, or relevant residue classes such as *small*, *hydrophobic* or *negative*. Relations have the form *Within_n(p,r1,r2)* and *Plus_n(p,r1,r2)*, and represent pairs of residues *(r1,r2)* in a certain protein *p* whose sequence separation is at most or exactly *n*, respectively. Figure 1 (left) shows a TET branch[4] which proved quite stable in predicting CYS ligands, encoding counts-of-counts features of the candidate residue neighbourhood. Here XXX and YYY can be: HIS or *negative* identifying candidate co-ligands (ASP and GLU are both negatively charged); *polar* or *positive* identifying hydrophilic residues and thus an exposed protein surface; *small* indicating a small residue which favours the local flexibility. Collective classification experiments introduce an additional learning issue, as target predicates available for candidate extensions are predicted and thus subject to predictive errors. Figure 1 (right) shows an ideal TET learned assuming that labels of related residues are given. The left branch considers the number of co-ligands occurring in the protein, while the right one searches for other proteins having ligands in the same position (residues are identified by their position in sequence), and for nearby ligands in the target protein. Such TET fails to generalize to new proteins, where knowledge of the labels of related residues cannot be assumed and must be replaced by predicted labels, as it relies too heavily on the quality of such predictions. More robust TET can be learned assuming predicted

[4] In all reported TETs, inequality constraints forcing newly introduced variables to be different from root ones were skipped for simplicity.

labels (according to a non-collective TET) instead of true ones for related residues during training. However, using the same acceptance threshold for extensions employed in [2] tends to produce too simplified collective TET on some of the folds, worsening performance with respect to the corresponding non-collective TET. The problem can be partially fixed by decreasing the acceptance threshold. Table 1 reports areas under the ROC curve for different folds with non-collective and collective TET, where not conserved residues are considered non-binding by default, in order to focus on ambiguous cases. Collective classification outperforms individual predictions on 2 and 3 out of 5 folds for CYS and HIS respectively. The fact the HIS prediction benefits more from collective classification is not surprising, as CYS predictions tend to be more accurate, and can thus propagate to nearby HIS.

Table 1. Area under the ROC curves for CYS and HIS metal bonding state prediction for different folds. Comparison between single TET as in [2] and collective TET with different thresholds for extension acceptance.

	CYS			HIS		
fold	TET_{single}	TET_{coll_simple}	$\text{TET}_{coll\_complex}$	TET_{single}	TET_{coll_simple}	$\text{TET}_{coll\_complex}$
1	89.6 ± 1.6	91.6 ± 1.4	91.7 ± 1.4	83.1 ± 2.1	80.5 ± 2.2	80.0 ± 2.2
2	92.0 ± 1.4	91.1 ± 1.4	91.0 ± 1.5	84.1 ± 2.4	84.8 ± 2.3	85.4 ± 2.3
3	84.6 ± 1.8	85.6 ± 1.7	87.6 ± 1.6	78.2 ± 2.2	80.0 ± 2.2	80.3 ± 2.2
4	85.8 ± 1.7	85.1 ± 1.8	85.7 ± 1.7	79.6 ± 2.4	76.8 ± 2.5	79.5 ± 2.4
5	89.8 ± 1.6	81.0 ± 2.0	83.9 ± 1.9	82.1 ± 2.2	86.2 ± 2.0	87.3 ± 2.0

Figure 2 shows the best conserved fragments which were learned across different folds in CYS prediction. The TET combines target information from the relational neighbourhood with unsupervised features in order to account for possible estimation errors. The first branch considers predicted co-ligands in the protein as in the corresponding branch of the ideal TET, but it further refines such information considering their distance to the target residue, and the presence of additional conserved CYS nearby as an indication of possible ligands. The second branch considers unsupervised features only, looking for candidate co-ligands (CYS or HIS) or small residues improving flexibility in the neighbourhood.

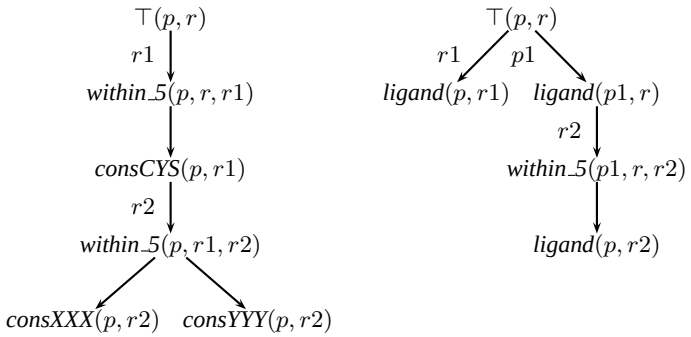

Fig. 1. TETs for metal bonding state prediction: (left) TET fragment from [2]; (right) TET learned assuming that labels of related residues are given

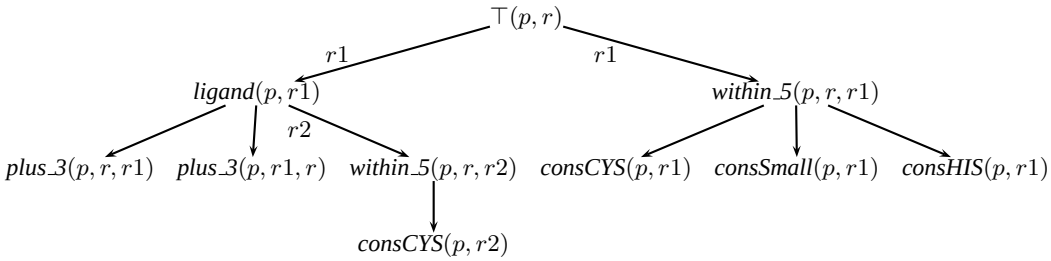

Fig. 2. TET fragments learned assuming a collective classification setting

References

1. M. Jaeger. Type extension trees: a unified framework for relational feature construction. In *Proceedings of Mining and Learning with Graphs (MLG-06)*, 2006.
2. P. Frasconi, M. Jaeger and A. Passerini. Feature Discovery with Type Extension Trees. In *Proceedings of the 18th Int. Conf. on Inductive Logic Programming (ILP-08)*, 2008.
3. L. P. Castillo and S. Wrobel. A comparative study on methods for reducing myopia of hill-climbing search in multirelational learning. In *Proc. of the 21st Int. Conf. on Machine Learning*, 2004.
4. N. Hulo, C. J. A. Sigrist, V. Le Saux, P. S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. De Castro, P. Bucher, and A. Bairoch. Recent improvements to the prosite database. *Nucleic Acids Research*, 32(Database-Issue):134–137, 2004.
5. A. Passerini, M. Punta, A. Ceroni, B. Rost, and P. Frasconi. Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. *Proteins*, 65(2):305–316, 2006.