Probabilistic Frequent Subtree Kernels

Pascal Welke[1], Tamás Horváth[1,2], and Stefan Wrobel[2,1]

[1] Dept. of Computer Science, University of Bonn, Germany
[2] Fraunhofer IAIS, Schloss Birlinghoven, Sankt Augustin, Germany

Abstract. Graph kernels have become a well-established approach in graph mining. One of the early graph kernels, the *frequent subgraph kernel*, is based on embedding the graphs into a feature space spanned by the set of all frequent connected subgraphs in the input graph database. A drawback of this graph kernel is that the preprocessing step of generating *all* frequent connected subgraphs is computationally intractable. Many practical approaches ignore this limitation, implying that such systems can be infeasible even for small datasets. Approaches that do not disregard this aspect either restrict the feature space or restrict the class of the input graphs to guarantee correctness and efficiency.
We propose a frequent subgraph kernel that is not restricted to any particular graph class, but still efficiently computable. All such kernels can only be achieved by relaxing the correctness condition on mining frequent connected subgraphs. We give up the demand on completeness and represent each input graph by a polynomial size random sample of its spanning trees. Such a random sample is a forest and can be generated in polynomial time. Thus, as frequent subtrees in forests can be listed with polynomial delay, we arrive at an efficient frequent subgraph mining algorithm. Our approach is sound, but incomplete: (i) it is only able to identify frequent subtrees, and not arbitrary graph patterns, and (ii) even if a tree pattern is frequent, it might not be identified as such. Calculating a representation in this feature space for any unseeng graph is done by the same incomplete procedure.
Our empirical evaluation on two chemical datasets shows that a considerable fraction of all frequent subtrees can be recovered even from *one* random spanning tree per graph. Regarding the expressive power of probabilistic frequent subtrees, we have observed a marginal loss in predictive performance. However, we have achieved a three time speed-up against the ordinary frequent subgraph kernel. Furthermore, our method is able to process significantly larger datasets and generates a much smaller feature set than the original algorithm.

A long version of this extended abstract appeared in [1].

[1] P. Welke, T. Horváth, and S. Wrobel. Probabilistic Subtree Kernels. To appear in: New Frontiers in Mining Complex Patterns, Springer, 2016.

 In: R. Bergmann, S. Görg, G. Müller (Eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. Trier, Germany, 7.-9. October 2015, published at http://ceur-ws.org