Min-Hashing for Probabilistic Frequent Subtree Feature Spaces*

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

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

Abstract. We propose a fast algorithm for approximating graph similarities. Here, the similarity between two graphs is defined by the Jaccard-similarity of their images in a binary feature space spanned by the set of frequent subtrees generated for some training dataset. While being an adequate choice for many similarity based learning tasks, this approach suffers from severe computational limitations. In particular, mining frequent trees in arbitrary graph databases cannot be done in output polynomial time and embedding a graph in the above space is NP-hard.

To overcome these limitations, we represent each graph by k of its spanning trees generated uniformly at random. In this way, we reduce the frequent subgraph mining, as well as the embedding of a graph into the feature space to problems involving only trees and forests. Clearly, the output of this probabilistic technique is always *sound* (any tree found to be frequent by this algorithm is a frequent subtree with respect to the original dataset), but *incomplete* (the algorithm may miss frequent subtrees). Similarly, the embedding of a given graph in the feature space spanned by the above trees is computed with a one-sided error. We improve the speed and space consumption of the above method by applying min-hashing for the embedding step. Each graph is represented by a small sketch vector that can be used to approximate Jaccard-distances. We show that the partial order on the feature set defined by subgraph isomorphism allows for a fast calculation of the min-hash sketch, without explicitly performing the feature space embedding.

Our experimental results demonstrate that the proposed technique can dramatically reduce the number of subtree isomorphism tests, compared to an algorithm performing the embedding explicitly. We also show that even for a few random spanning trees per chemical compound, remarkable precisions of the active molecules can be obtained in a highly imbalanced chemical dataset by taking the i nearest neighbors of an active compound. Finally, we show that the predictive power of support vector machines using our approximate similarities compares favorably to that of state-of-the-art related methods.

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

[1] P. Welke, T. Horváth, and S. Wrobel. Min-Hashing for Probabilistic Frequent Subtree Feature Spaces. To appear in: Proceedings of the 19th International Conference on Discovery Science, DS 2016, Springer LNAI, 2016.