Maximum Common Subgraph Mining: A Fast and Effective Approach towards Feature Generation

Leander Schietgat LEANDER.SCHIETGAT@CS.KULEUVEN.BE
Fabrizio Costa FABRIZIO.COSTA@CS.KULEUVEN.BE
Jan Ramon JAN.RAMON@CS.KULEUVEN.BE
Luc De Raedt LUC.DERAEDT@CS.KULEUVEN.BE
Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium

Keywords: feature generation, subgraph mining, structure-activity learning, chemoinformatics

Abstract

There exists a wide variety of local graph mining approaches that search for frequent, correlated or closed patterns in graphs. These methods typically return very large sets of patterns which can then be used as features to build classifiers. Here we take a different approach: rather than mining for all local patterns, we randomly sample from the set of maximum common subgraphs. The advantages are that maximum common subgraphs are easier to compute than frequent or correlated patterns, and that the resulting features lead to classification models that achieve significantly better predictive performance than models built on the patterns returned by traditional mining approaches.

1. Introduction

During the last decade, a lot of attention has been devoted to mining local patterns in graphs leading to the development of many graph mining systems (Yan & Han, 2002; Bringmann et al., 2006). These systems typically employ constraints to specify the patterns of interest, such as frequency, or top-k according to a correlation measure (e.g., χ^2). Graph mining systems typically perform a complete search through the entire graph space enumerating all subgraphs satisfying these constraints. Usually the resulting patterns are not used directly. Instead, they are used as features in traditional machine learning algorithms. Furthermore, the quality of the generated patterns is measured through the quality of the induced classifiers or models for regression (Wale et al., 2008). While these approaches offer strong guarantees w.r.t. completeness or optimality, they have a high computational cost and require post-processing to deal for example with redundancy issues (Bringmann et al., 2006). In this way, local pattern mining acts as a complex, expensive and *indirect* approach to generate features for graphs.

In this abstract we propose a direct, efficient and simple approach for the generation of interesting graph patterns. This method is related to the work of (Chaoji et al., 2008), who also found that good patterns are obtained by sampling under diversity constraints. Our idea is to compute maximum common subgraphs from randomly selected pairs of examples and directly use them as features. The advantages of this approach are 1) that it is easy to control the number of produced features (while setting the frequency in a pattern mining task yields an unpredictable number of patterns); 2) that patterns can be extracted more efficiently than by frequent or correlated subgraph mining, as no search space has to be traversed; and 3) that on a number of benchmark problems from NCI, the extracted features allow to build SVM classification models that achieve significantly better predictive performance than those built on features returned by traditional local pattern mining approaches.

2. Method

A maximum common subgraph (MCS) of two graphs G and H is a graph I which is subgraph isomorphic to G and H and there exists no other graph J which is also subgraph isomorphic to G and H and $|J| > |I|$. Outerplanar graphs are graphs which can be embedded in the plane such that all of their vertices lie on

Presented at *SRL-MLG-ILP*, Leuven, Belgium, 2009.

the boundary of the outer face. It is known that 95% of the molecules in the NCI datasets are outerplanar. Outerplanar graphs consist of blocks and bridges. A block is a maximum subgraph for which every two vertices are involved in a cycle, while a bridge is an edge that does not belong to a block.

Even though computing the MCS between two general graphs is NP-hard, it is possible to compute the MCS between two outerplanar graphs in polynomial time by using the block-and-bridge preserving (BBP) subgraph isomorphism (Schietgat et al., 2008). This is a variant of the general subgraph isomorphism that only maps blocks to blocks and bridges to bridges.

In this abstract, we use the MCS algorithm of (Schietgat et al., 2008) to generate patterns in graph-based data by computing the MCS between randomly selected pairs of graphs in the input data. The use of this algorithm has a few implications. Firstly, it can only be used to mine the outerplanar graphs. For feature generation, this does not pose a problem since the patterns can still be embedded in non-outerplanar examples by using the general subgraph isomorphism. Secondly, only outerplanar patterns are generated. However, our experiments below have shown that a correlated subgraph mining approach does not find non-outerplanar patterns either. Thirdly, while sampling MCSs, duplicate patterns can be generated. Therefore, we check every time a new MCS is sampled whether it is isomorphic to already found MCSs. Because of the BBP subgraph isomorphism this can also be done in polynomial time. Lastly, the use of the BBP subgraph isomorphism introduces a bias on the learned patterns. For example, ring structures are either entirely included in the patterns or not at all. Keeping ring structures and linear fragments apart seems to make sense from a chemical viewpoint.

3. Experiments

In order to evaluate the properties and the quality of the extracted subgraphs as features for predictive tasks, we give an experimental answer to the following questions: **Q1** What is the difference in predictive performance between the features generated by maximum common subgraph mining and a correlated subgraph mining system? **Q2** Are the patterns returned by the MCS procedure less redundant than those returned by a correlated pattern mining procedure? **Q3** How does the predictive performance vary w.r.t. the number of sampled MCSs? **Q4** Is there a significant difference between sampling randomly and selecting the top-k correlated MCSs or the top-k frequent MCSs? **Q5** How many pairs of molecules need to be sampled in order to obtain k unique MCSs and how much time does this take?

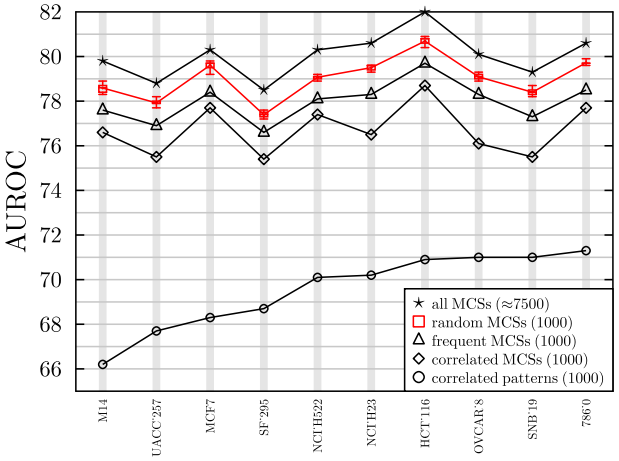

Figure 1. Predictive performance of the different sampling approaches on 10 NCI datasets.

Datasets We use 10 randomly selected datasets from the NCI cancer dataset collection (Swamidass et al., 2005), which contain molecules and their ability to suppress or inhibit the growth of human tumour cell lines. Each dataset has on average 3,500 examples with a balanced binary class distribution.

Methodology Each example is propositionalized into a bitvector encoding to represent the occurrence of the mined patterns. SVMs in combination with the Tanimoto-kernel (Swamidass et al., 2005) were used as classification model. The area under the ROC-curve (AUROC) score is reported for all experiments using a stratified 10-fold cross-validation. Patterns are mined only from training data. The regularization parameter of the SVM is tuned running an internal 5-fold cross-validation over the training data. Statistical significance is assessed either using a sign test from the win/loss-ratio or noticing that generalization to other molecules from the same population is significantly better at the 1% level for samples of ≈3,500 molecules when an increase of 2.5% in AUROC is measured.

Results

Q1 We compare 1000 randomly selected MCSs (results are averaged over 10 runs) to the 1000 most χ^2 correlated patterns (extracted with the system from (Bringmann et al., 2006)) over the 10 datasets. Figure 1 shows a clear advantage for the MCS approach.

Q2 We compute the percentage of examples from the test set (averaged over the 10 datasets) with a unique encoding (uniqueness) and the percentage of examples belonging to the largest cluster of examples that have

Table 1. Redundancy evaluation of 4 pattern sets.

Mining approach	Uniqueness	Redundancy
1000 randomly selected MCSs	98.328±0.003	0.355±0.014
1000 most frequent MCSs	96.846±0.004	0.820±0.055
1000 most correlated MCSs	91.115±0.018	1.358±0.211
(Bringmann et al., 2006)	56.590±4.351	17.748±4.807

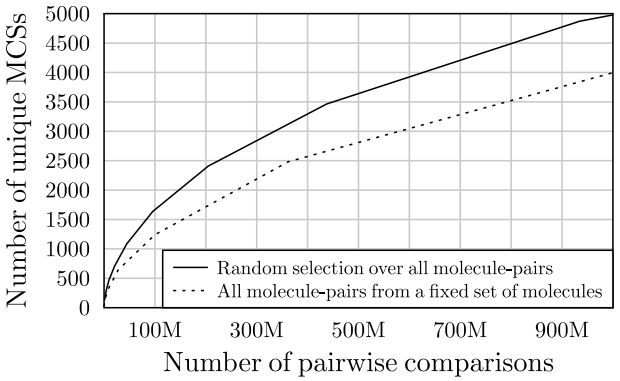

Figure 2. Relationship between size of molecule-pairs set and size of unique MCS set.

the same encoding (redundancy). Table 1 shows that MCSs provide more discriminative encodings.

Q3 We measure predictive performance as we increase the number of randomly sampled MCSs: for 100, 200, 400, 800, 1600, 3200 and 6400 patterns, an AUROC of respectively 75.0, 76.7, 78.0, 79.0, 79.8, 80.3 and 80.5 is obtained (averaged over 5 datasets).

Q4 We measure predictive performance over 10 datasets using all unique MCSs ($\approx$ 7500), 1K most χ^2 correlated MCSs, 1K most frequent MCSs and 1K randomly selected MCSs (averaged over 10 runs). Figure 1 shows that the random sampling approach is significantly better than the correlated mining one.

Q5 We have experimentally determined (Fig. 2) that to obtain 1K different MCSs we need 45K pairs of randomly sampled molecules or a random sample of 400 molecules out of which to consider all possible pairs[1]. We have observed an almost perfect linear relationship (with coefficient 2.6) between the number of molecules and the number of different MCSs, that is: given a set of 1K molecules, extracting the MCSs from all pairs gives 2.6K unique MCSs.

Runtimes Mining the 1000 most χ^2 correlated graphs with correlated pattern mining takes on average 2 hours, while sampling 1000 unique MCSs takes on average 28 minutes, achieving a 400% speedup.

[1]Note that 400 molecules generate 160K pairs, a larger number due to the reduced diversity of the molecules involved in the pairs.

4. Conclusions and Future Work

We have shown that sampling from the set of pairwise maximum common subgraphs is a competitive strategy w.r.t. enumerating all correlated subgraphs (which in turn is a better strategy than enumerating all frequent subgraphs). We conjecture that the reason is that the sampling strategy achieves a greater diversity in the set of returned solutions thus better controlling the redundancy issue. Moreover, the employed bias (i.e. BBP subgraph isomorphism and pairwise maximal commonality) guarantees that, at least in the case of small molecules, the returned subgraphs are not only non-redundant but also interesting features for biological activity prediction tasks. We note that this approach has a wider applicability than the specific molecular context.

Possible extensions include the exploration of different sampling strategies to further reduce computational costs and the use of different language bias (e.g., maximum common subtrees or subsequences).

Acknowledgments IWT-Vlaanderen to LS. GOA Prob. Logic Learning to FC. FWO-Vlaanderen to JR.

References

Bringmann, B., Zimmermann, A., Raedt, L. D., & Nijssen, S. (2006). Don't be afraid of simpler patterns. *Proc. of the 10th ECML-PKDD* (pp. 55–66).

Chaoji, V., Al Hasan, M., Salem, S., Besson, J., & J. Zaki, M. (2008). Origami: A novel and effective approach for mining representative orthogonal graph patterns. *Stat. Anal. Data Min.*, *1*, 67–84.

Schietgat, L., Ramon, J., Bruynooghe, M., & Blockeel, H. (2008). An efficiently computable graph-based metric for the classification of small molecules. *Proc. of the Int. Conf. on Disc. Science* (pp. 197–209).

Swamidass, S. J., Chen, J., Bruand, J., Phung, P., Ralaivola, L., & Baldi, P. (2005). Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinf.*, *21*, i359–368.

Wale, N., Watson, I., & Karypis, G. (2008). Comparison of descriptor spaces for chemical compound retrieval and classification. *Kn. In. Sy.*, *14*, 347–375.

Yan, X., & Han, J. (2002). gSpan: Graph-based substructure pattern mining. *Proc. of the 2002 IEEE Int. Conf. on Data Mining* (pp. 721–724).