

QSAR Analysis with Support Vector Machines and Graph Kernels

Pierre Mahé¹ Nobuhisa Ueda² Tatsuya Akutsu²
pierre.mahé@ensmp.fr ueda@kuicr.kyoto-u.ac.jp takutsu@kuicr.kyoto-u.ac.jp

Jean-Luc Perret² Jean-Philippe Vert¹
luc@kuicr.kyoto-u.ac.jp Jean-Philippe.Vert@ensmp.fr

¹ Ecole des Mines de Paris, 35 rue Saint Honor, 77300 Fontainebleau, France

² Bioinformatics Center, Kyoto University, Uji, Kyoto 611-0011, Japan

Keywords: chemical compounds, graph kernel, support vector machine

1 Introduction

Kernel methods, such as support vector machines, have been applied to solving various problems in bioinformatics. Recently, marginalized kernels between labeled graphs have been proposed [2, 3], which enable the application of kernel methods to the analysis and classification of chemical compounds such as QSAR (quantitative structure-activity relationship). These graph kernels are based on the detection of common paths between different graphs. These correspond to a dot product between the graphs mapped to an infinite-dimensional feature space, but can be computed in polynomial time with respect to the graph sizes. Encouraging experimental results suggest that this approach might be useful in analysis of chemical compounds.

These graph kernels, however, are subject to several limitations. First, the choice of representing implicitly each graph by the set of path probabilities under a simple random walk model might be questioned. In chemoinformatics, subgraphs are believed to be more relevant features than paths. Moreover, the random walk model used is subject to “tottering”, in the sense that it can move to one direction and instantly come back to its original position, resulting in redundant paths which might decrease the characterization of a given graph once mapped to the feature space of these graph kernels. Second, the graph kernel has a computational complexity roughly proportional to the product of the sizes of the two graphs to be compared, which results in slow implementation for real-world problem.

2 Two Extensions of the Marginalized Graph Kernels

We propose two extensions of marginalized graph kernels, with the double goal to reduce their computation time and increase their relevance as measure of similarity between graphs. The first extension is to relabel each vertex automatically in order to insert information about the environment of each vertex in its label with the use of the Morgan algorithm [7]. This has both an effect in terms of feature relevance, because label paths contain information about their environment as well, and computation time, because the number of identical labeled paths significantly decreases. Second, we modify the random walk model in order to remove totters, without increasing the complexity of the implementation. Details of these extensions can be found in [6].

3 Computational Experiments

We applied this family of graph kernels for QSAR analysis. The dataset (MUTAG [1]) contains 188 molecules (aromatic and heteroaromatic nitro compounds) tested for mutagenicity on *Salmonella typhimurium* (another dataset (PTC) was tested [6] but omitted). The problem consists in predicting a low or high mutagenic activity, and we use the standard SVM algorithm for binary classification.

Table 1 shows success rates and corresponding computation times obtained with the introduction of the label enrichment methods for 20 iterations of the Morgan index process, where the original random walk model was used. Removing tottering paths reaches the same performance, but introduces a huge increase in the complexity of the kernel.

The extended graph kernels presented here compare favorably to state of the art algorithms using only molecular descriptors [4], but with the key advantage not to require a molecular descriptor selection. The two extensions presented here only slightly improved the mutagenicity prediction model, but the label enrichment method allowed a much faster computation speed, opening the way to use graph kernels to quickly compute similarities between thousands of molecules.

Future work involves combining this exclusively 2D-structure based approach with classical descriptors, as did for example Kramer et al. [5] who showed that better results can be obtained by combining molecular descriptors with fragment analysis.

Table 1: Classification results and computation times for the MUTAG dataset introducing different Morgan indices in the labeling of atoms.

#Iterations	None	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10-20th
Success rate (%)	89.9	91	91	88.3	86.7	86.2	81.4	76.6	75.5	75	74.5
Computation (sec)	7032	5718	874	230	33	7	4	4	3	3	3

References

- [1] Debnath, A.K., Lopez de Compadre, R.L., Debnath, G., Schusterman, A.J., and Hansch, C., Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity, *Journal of Medicinal Chemistry*, 34:786-797, 1991.
- [2] Gärtner, T., Flach, P., and Wrobel, S., On graph kernels: hardness results and efficient alternatives, *Proc. 16th Conf. Computational Learning Theory*, 129-143, 2003.
- [3] Kashima, H., Tsuda, K., and Inokuchi, A., Marginalized kernels between labeled graphs, *Proc. 20th Int. Conf. Machine Learning*, 321-328, 2003.
- [4] King, R.D., Muggleton, S.H., Srinivasan, A., and Sternberg, M.J.E., Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming, *Proc. Natl. Acad. Sci. USA*, 93:438-442, 1996.
- [5] Kramer, S. and De Raedt, L., Feature construction with version spaces for biochemical applications, *Proc. 18th Int. Conf. Machine Learning*, 258-265, 2001.
- [6] Mahé, P., Ueda, N., Akutsu, T., Perret, J-L., and Vert, J-P., Extensions of marginalized graph kernels, *Proc. 21st Int. Conf. Machine Learning*, 552-559, 2004.
- [7] Morgan, H.L., The generation of unique machine description for chemical structures - a technique developed at chemical abstracts service, *Journal of Chemical Documentation*, 5:107-113, 1965.