Combining near-optimal feature selection with gSpan

Karsten Borgwardt[1], **Xifeng Yan**[2], **Marisa Thoma**[3], **Hong Cheng**[6], **Arthur Gretton**[4], **Le Song**[5], **Alex Smola**[5], **Jiawei Han**[6], **Philip Yu**[2], **Hans-Peter Kriegel**[3]

[1]University of Cambridge, [2]IBM T.J. Watson Research Center New York, [3]Ludwig-Maximilians-Universität München, [4]Max Planck Institute Tübingen, [5]National ICT Australia, Canberra, [6]University of Illinois at Urbana-Champaign

Abstract

Graph classification is an increasingly important step in numerous application domains, such as function prediction of molecules and proteins, computerised scene analysis, and anomaly detection in program flows.

Among the various approaches proposed in the literature, graph classification based on frequent subgraphs is a popular branch: Graphs are represented as (usually binary) vectors, with components indicating whether a graph contains a particular subgraph that is frequent across the dataset.

On large graphs, however, one faces the enormous problem that the number of these frequent subgraphs may grow exponentially with the size of the graphs, but only few of them possess enough discriminative power to make them useful for graph classification. Efficient and discriminative feature selection among frequent subgraphs is hence a key challenge for graph mining.

In this article, we propose an approach to feature selection on frequent subgraphs, called *CORK*, that combines two central advantages. First, it optimises a submodular quality criterion, which means that we can yield a near-optimal solution using greedy feature selection. Second, our submodular quality function criterion can be integrated into gSpan, the state-of-the-art tool for frequent subgraph mining, and help to prune the search space for discriminative frequent subgraphs even *during* frequent subgraph mining.

Preliminary work. Under review by the International Workshop on Mining and Learning with Graphs (MLG). Do not distribute.

1. Introduction

A typical graph classification problem has the following formulation: given a set of training graphs associated with labels $\{G_i, y_i\}_{i=1}^n$, $y_i \in \{\pm 1\}$, the task is to learn a classifier that predicts the labels of unclassified structures. The resulting classification algorithm has a wide variety of real world applications.

As in classification problems on vectorial data, we are interested in *feature selection*: Which *features*, that is which *subgraphs*, are most correlated with the class membership of the graphs in a dataset? In other words, we would like to find *discriminative* sets of subgraphs that allow us to tell apart different classes of graphs. This problem is involved for two reasons: First, the search space grows exponentially with the number of subgraphs we have to consider. Second, the number of subgraphs grows exponentially with the number of nodes in a graph.

In data mining, efficient strategies have been developed to search the space of subgraphs to perform *frequent subgraph mining*, that is, to find those subgraphs that occur in more than $t\%$ of all graphs in our dataset. Despite their efficient mining, these frequent subgraphs are not necessarily *discriminative*. Both a very frequent and a rather rare subgraph can be uniformly distributed over different classes of graphs and may possess little discriminative power. Hence combining the efficiency of frequent subgraph mining with feature selection for discriminative frequent subgraphs is an attractive research goal.

Goal Our goal is to define an efficient near-optimal approach to feature selection among frequent subgraphs generated by gSpan (Yan & Han, 2002). The key idea is to pick frequent subgraphs that greedily maximise a *submodular* quality criterion, thereby guaranteeing that the greedy solution to the feature selection problem is close to the global optimal solution (Nemhauser et al., 1978). To make this approach efficient, we in-

tegrate it into gSpan, the state-of-the-art tool for frequent subgraph mining, and derive pruning criteria that allow us to narrow down the search space when looking for discriminative subgraphs.

Unlike its predecessors that use ad-hoc strategies for feature selection (such as (Cheng et al., 2007)) or do not provide optimality guarantees, we define a principled, near-optimal approach to feature selection on frequent subgraphs that can be integrated into gSpan (Yan & Han, 2002). An excellent wrapper approach to this problem has recently been published by (Tsuda, 2007). Our approach differs from Tsuda's in two ways: Our feature selection method is independent from the choice of classifer (filter method) and we can provide optimality guarantees for our solution.

2. Near-optimal feature selection on frequent subgraphs

2.1. Feature Selection and Submodularity

Assume that we are measuring the discriminative power $q(\mathcal{S})$ of a set of frequent subgraphs $\mathcal{S}$ in terms of a quality function q. A near-optimality solution is reached for a *submodular* quality function q when used in combination with greedy feature selection. Greedy forward feature selection consists in iteratively picking the feature that – in union with the features selected so far - maximises the quality function q. In general, this strategy will not yield an optimal solution, but it can be shown to yield a near-optimal solution if q is submodular:

Definition 1 (Submodular set function) *A quality function q is said to be* **submodular** *on a set $\mathcal{D}$ if for $\mathcal{T}' \subset \mathcal{T} \subseteq \mathcal{D}$ and $X \in \mathcal{D}$*

$$q(\mathcal{T}' \cup \{X\}) - q(\mathcal{T}') \geq q(\mathcal{T} \cup \{X\}) - q(\mathcal{T}) \quad (1)$$

If q is submodular and we employ greedy forward feature selection, then we can exploit the following theorem from (Nemhauser et al., 1978):

Theorem 2 *If q is a submodular, nondecreasing set function on a set $\mathcal{D}$ and $q(\emptyset) = 0$, then greedy forward feature selection is guaranteed to find a set of features $\mathcal{T} \subseteq \mathcal{D}$ such that*

$$q(\mathcal{T}) \geq \left(1 - \frac{1}{e}\right) \max_{\mathcal{U} \subseteq \mathcal{D}:\ |\mathcal{U}|=t} q(\mathcal{U}) \ , \quad (2)$$

where t is the number of features to be selected.

As a direct consequence, the result from greedy feature selection achieves at least $\left(1 - \frac{1}{e}\right) \approx 63\%$ of the score of the optimal solution to the feature selection problem. This is referred to as being *near-optimal* in the literature (e.g. (Guestrin et al., 2005)).

2.2. Feature selection on frequent subgraphs

Once we have determined the frequent subgraphs using gSpan, a natural way of representing each graph G is in terms of a binary indicator vector of length $|\mathcal{S}|$:

Definition 3 (Indicator vector) *Given a graph G_j from a dataset $\mathcal{G}$ and a set of frequent subgraph features $\mathcal{S}$ discovered by gSpan. We then define an indicator vector $v^{(j)}$ for G_j as*

$$v_i^{(j)} = \begin{cases} 1 & \text{if } \mathcal{S}_i \sqsubseteq G_j \text{ ($\mathcal{S}_i$ is a subgraph of G_j)} \\ 0 & \text{otherwise} \end{cases} , \quad (3)$$

where $v_i^{(j)}$ is the i-th component of $v^{(j)}$ and $\mathcal{S}_i$ is the i-th graph in $\mathcal{S}$.

2.3. Definition of CORK

Definition 4 *Let $\mathcal{G}$ be a dataset of binary vectors, consisting of two classes $\mathcal{G} = \{\mathcal{A}, \mathcal{B}\}$. Let $\mathcal{D}$ denote the set of features of the data objects in $\mathcal{G}$, and let X be a single feature from $\mathcal{D}$,* i.e., *$X \in \mathcal{D}$.*

Definition 5 (Correspondence) *A pair of data objects $(v^{(i)}, v^{(j)})$ is called a* **correspondence** *in a set of features indicated by indices $\mathcal{U} \subseteq \{1, \ldots, |\mathcal{D}|\}$ (or, w.r.t. to a set of features $\mathcal{U}$) iff*

$$(v^{(i)} \in \mathcal{A}) \wedge (v^{(j)} \in \mathcal{B}) \wedge \forall d \in \mathcal{U} : (v_d^{(i)} = v_d^{(j)}), \quad (4)$$

Definition 6 (CORK) *We define a quality criterion q, called* **CORK** *(*Correspondence-based Quality Criterion*), for a subset of features $\mathcal{U}$ as*

$$q(\mathcal{U}) = (-1) * \textit{number of correspondences in } \mathcal{U} \quad (5)$$

This criterion is submodular, and can be turned (by adding a constant) into a submodular set function that fulfills the conditions of Theorem 2 (Proof in extended version of this abstract).

2.4. Pruning gSpan's search space via CORK

gSpan exploits the fact that the frequency of a subgraph $S \in \mathcal{S}$ is an upper bound for the frequency of all of its super-graphs $S' \sqsupseteq S$ (all subgraphs containing S) when pruning the search space for frequent subgraphs. In a similar way, we show that from the CORK-value of a subgraph S, we can derive an upper bound for the CORK-values of all of its supergraphs, that allows us to further prune the search space.

		Filter						Wrapper	
		SC		PC		CORK		LAR	
Dataset	# of Features	Accuracy	Std	Accuracy	Std	Accuracy	Std	Accuracy	Std
NCI1	57	66.98	2.31	65.43	3.82	70.98	2.31	73.08	2.06
NCI33	53	66.50	2.57	64.15	3.46	70.08	2.76	72.81	2.51
NCI41	49	70.20	3.23	65.37	4.27	70.38	2.72	72.39	2.58
NCI47	56	67.04	2.35	67.00	3.45	71.42	2.22	72.62	2.07
NCI81	64	69.04	2.17	64.27	5.01	70.76	2.21	72.58	1.88

Table 1. Classification accuracy (and standard deviation (std)) in % on the 5 NCI graph datasets (SC = Sequential Cover, PC = Pearson's Correlation Coefficient, CORK = Correspondence-based Quality Criterion, LAR = LAR-LASSO used for feature selection). Frequency threshold for frequent subgraphs is 10%.

Theorem 7 *Let $S, S' \in \mathcal{S}$ be frequent subgraphs, and S' be a supergraph of S. Let $\mathcal{A}_{S_1}$ denote the number of graphs in class $\mathcal{A}$ that contain S, $\mathcal{A}_{S_0}$ the number of graphs in $\mathcal{A}$ that do not contain S (and define $\mathcal{B}_{S_0}$, $\mathcal{B}_{S_1}$ analogously). Then*

$$q(\{S\}) = -\left(\mathcal{A}_{S_0} * \mathcal{B}_{S_0} + \mathcal{A}_{S_1} * \mathcal{B}_{S_1}\right) \tag{6}$$

and

$$q(\{S'\}) \leq q(\{S\}) + \max(\{\mathcal{A}_{S_1}(\mathcal{B}_{S_1} - \mathcal{B}_{S_0}), (\mathcal{A}_{S_1} - \mathcal{A}_{S_0})\,\mathcal{B}_{S_1}, 0\}) \tag{7}$$

Proof See extended version of this abstract. ■

We can now use Inequality (7) to provide an upper bound for the CORK values of supergraphs of a given subgraph S and exploit this information for pruning the search space in a branch-and-bound fashion.

Inequality (7) can be directly applied in the first iteration of greedy selection. For later iterations of greedy selection, we can define a similar bound for pruning (see extended version). The main difference to (7) is that in later iterations of greedy selection, we only have to consider those graphs which are part of a correspondence (rather than all graphs).

3. Results and Discussion

To assess the quality of the features selected by CORK, we use it on five (balanced) NCI anti-cancer-screen graph classification benchmarks. We let CORK pick informative frequent subgraphs, as determined by gSpan (frequency threshold 10%), on these datasets, and then perform 10-fold cross-validation using a C-SVM. All SVM-parameters are determined on the training set only. We compare CORK to two state-of-the-art filter methods, namely Pearson's Correlation Coefficient and the Sequential Cover method (Deshpande et al., 2005), and to the LAR-LASSO wrapper approach by (Tsuda, 2007).

Among the filter methods, CORK performs best, and among all methods, it is second to the LAR-LASSO wrapper approach by (Tsuda, 2007). We are currently exploring other submodular criteria (Boros et al., 2003) for subgraph feature selection that might lead to even higher levels of prediction accuracy.

References

Boros, E., Horiyama, T., Ibaraki, T., Makino, K., & Yagiura, M. (2003). Finding essential attributes from binary data. *Ann. Math. Artif. Intell.*, *39*, 223–257.

Cheng, H., Yan, X., Han, J., & Hsu, C. (2007). Discriminative frequent pattern analysis for effective classification. *Proc. of ICDE*. Istanbul, Turkey.

Deshpande, M., Kuramochi, M., Wale, N., & Karypis, G. (2005). Frequent substructure-based approaches for classifying chemical compounds. *IEEE Transactions on Knowledge and Data Engineering*, *17*, 1036–1050.

Guestrin, C., Krause, A., & Singh, A. (2005). Near-optimal sensor placements in gaussian processes. *Proc. Intl. Conf. Machine Learning*. Bonn, Germany.

Nemhauser, G., Wolsey, L., & Fisher, M. (1978). An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, *14*, 265–294.

Tsuda, K. (2007). Entire regularization paths for graph data. *Proc. Intl. Conf. Machine Learning* (pp. 919–926).

Yan, X., & Han, J. (2002). gSpan: Graph-based substructure pattern mining. *Proc. 2002 Int. Conf. on Data Mining (ICDM'02)* (pp. 721–724).