The Skew Spectrum of Graphs

Risi Kondor RISI@GATSBY.UCL.AC.UK
Gatsby Computational Neuroscience Unit, UCL, 17 Queen Square, London, WC1N 3AR, U.K.

Karsten M. Borgwardt KMB51@CAM.AC.UK
Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, U.K.

Abstract

The central issue in representing graph-structured data instances in learning algorithms is to design features that are invariant to permuting the numbering of the vertices. We present a new system of invariant graph features, which we call the skew spectrum of graphs. The skew spectrum is based on mapping the adjacency matrix to a function on the symmetric group and computing bispectral invariants. The reduced form of the skew spectrum is computable in $O(n^3)$ time, and experiments show that on several benchmark datasets it can outperform state of the art graph kernels.

1. Introduction

The central challenge in representing unlabeled graphs in learning algorithms is to account for the permutation (relabeling) symmetry of the vertices. Given a graph $\mathcal{G}$, the two main lines of research that have emerged to address the above problem focus respectively on (a) designing an explicit feature mapping $\mathcal{G} \mapsto (q_1, q_2, \ldots, q_k)$; and (b) designing a kernel $k(\mathcal{G}_1, \mathcal{G}_2)$. Proponents of the first approach exploit global invariant properties of $\mathcal{G}$, such as the eigenvalues of its graph Laplacian, or local invariant properties, such as the number of occurrences in $\mathcal{G}$ of a library of small subgraphs. One of the empirically most successful methods counts the lengths of shortest paths between pairs of vertices (Borgwardt & Kriegel, 2005). In constrast, proponents of the kernel approach use various intituitions about simulatneous random walks and diffusion on product graphs (Gärtner, 2002; Gärtner, 2003).

The new method that we present in this paper belongs in the first of the above two categories, but is distinguished from prior work (with the exception of (Shawe-Taylor, 1993)) by its algebraic character. In this regard, it is related to the recent line of papers (Kondor et al., 2007; Huang et al., 2008; Kondor, 2007a) introducing concepts from non-commutative harmonic analysis to machine learning. The mathematical foundations of our work are Kakarala's seminal results on the bispectra of functions on compact groups (Kakarala, 1993; Kakarala, 1992), and the recent discovery of a unitarily equivalent, but computationally more attractive set of invariants called the skew spectrum (Kondor, 2007b). We show how these general theories can be harnessed to construct graph invariants, and examine in detail their computational properties.

Experiments on standard datasets of chemical compounds show that the skew spectrum of graphs is competitive with the state of the art in graph features, and in some cases outperforms all other methods. Another major advantage of the skew spectrum is that since it is an explicit feature mapping, it can be applied as a preprocessing step, and hence its computation time scales linearly with the number of examples. The computational complexity of computing the (reduced) skew spectrum of a single graph of n nodes scales with n^3. Uniquely amongst the graph invariants used in machine learning, the skew spectrum has a fixed number of scalar components (85 for the complete skew spectrum and 49 for its reduced version), resulting in a very compact representation. This does not prevent the skew spectrum form remaining competitive both in speed and representational accuracy up to about $n = 300$.

2. The skew spectrum

Given a finite group G and a subgroup H of G, the **skew spectrum** of a function $f\colon G/H \to \mathbb{C}$ is, as defined in (Kondor, 2007b), the collection of matrices

$$\widehat{q}_\nu(\rho) = \widehat{r}_\nu^\dagger(\rho) \cdot \widehat{f}(\rho), \qquad (1)$$

Appearing in the 6^{th} *International Workshop on Mining and Learning with Graphs*, Helsinki, Finland, 2008.

where $r_\nu : G \to \mathbb{C}$ is the function

$$r_\nu(\sigma) = f(\sigma\nu)f(\sigma),$$

ν extends over a complete set of $H\backslash G/H$ double coset representatives, ρ extends over a complete set $\mathcal{R}$ of inequivalent irreducible representations of G, and the $\widehat{\ }$ symbol stands for the non-commutative Fourier transform

$$\widehat{f}(\rho) = \sum_{\sigma \in G} \rho(\sigma)\, f(\sigma) \qquad \rho \in \mathcal{R}.$$

It is shown in (Kondor, 2007b) that the skew spectrum is unitarily equivalent to the **bispectrum** (Kakarala, 1992), and is hence invariant to the translation action of G on functions given by

$$f \mapsto f^\pi \qquad f^\pi(\sigma) = f(\pi^{-1}\sigma) \qquad \pi \in G. \quad (2)$$

In this paper we take $G = \mathbb{S}_n$ (the symmetric group of permutations over n objects), $H = \mathbb{S}_{n-2}$, and show that if we map the adjacency matrix A of a graph $\mathcal{G}$ to the function

$$f(\sigma) = A_{\sigma(n),\sigma(n-1)}, \quad (3)$$

then the effect of permuting the vertices of $\mathcal{G}$ by $\pi \in \mathbb{S}_n$ is exactly (2). Hence, the skew spectrum of f is a graph invariant.

It is well known that the irreducible representations of $\mathbb{S}_n$ are indexed by integer partitions $\lambda \vdash n$, and the individual rows/columns of the $\rho_\lambda(\sigma)$ representation matrices are indexed by standard tableaux. We show that in Young's orthogonal representation if f is defined as in (3), then $\widehat{f}$ is identically zero except for

1. the single scalar component $\widehat{f}_{(n)}$;
2. the  column of $\widehat{f}_{(n-1,1)}$;
3. the  column of $\widehat{f}_{(n-1,1)}$;
4. the  column of $\widehat{f}_{(n-2,2)}$;
5. the  column of $\widehat{f}_{(n-2,1,1)}$.

Here ■ stands for n, • stands for $n-1$, and we draw tableaux aassuming $n = 8$. Given an appropriate choice of $\{\nu\}$ coset representatives (of which there are 7 in total), the Fourier transform of $\widehat{r}_\nu$ is similarly column-sparse and in the end we have just 85 non-zero scalar components in the (1) matrix products.

Computing $\widehat{r}_\nu$ is too expensive for practical applications, so we instead compute its projection r^*_ν to the space of functions on $\mathbb{S}_n/\mathbb{S}_{n-2}$. This gives $\widehat{r}^*_\nu$ the same sparsity pattern as $\widehat{f}$ has, and reduces the number of scalar graph invariants in the corresponding **reduced skew spectrum** to 49. Despite 49 being a small number (and independent of n), the reduced skew spectrum (computable in $O(n^3)$ operations) seems to be a remarkably powerful tool for capturing the structure of unlabeled graphs. Exhaustive enumeration shows that it can distinguish between almost all pairs of non-isomorphic simple graphs for $n = 1, 2, \ldots, 6$ (beyond $n = 6$ exhaustive enumeration becomes too expensive to try).

3. Experiments

In our experiments we evaluate the performance of the reduced skew spectrum features on four benchmark datasets of chemical structures of molecules: MUTAG, ENZYMES, NCI1, and NCI109. MUTAG (Debnath et al., 1991) is a dataset of 188 mutagenic aromatic and heteroaromatic nitro compounds. The classification task is to predict for each molecule whether it exerts a mutagenic effect on the Gram-negative bacterium *Salmonella typhimurium*. ENZYMES is a dataset which we obtained from (Borgwardt et al., 2005), and which consists of 600 enzymes from the BRENDA enzyme database (Schomburg et al., 2004). In this case the task is to correctly assign each enzyme to one of the 6 EC top level classes. The average number of nodes of the graphs in this dataset is 32.6 and the average number of edges is 124.3. Finally, we also conducted experiments on two balanced subsets of NCI1 and NCI109, which classify compounds based on whether or not they are active in an anti-cancer screen ((Wale & Karypis, 2006) and `http://pubchem.ncbi.nlm.nih.gov`).

Since in these datasets the number of vertices varies from graph to graph, we set n to be the maximum over the entire dataset and augment each of the smaller graphs with the appropriate number of unconnected "phantom" nodes. The experiments consisted of running SVMs on the above data using the reduced skew spectrum features (linear kernel on these features), the random walk kernel (Gärtner et al., 2003), (with λ set to 10^{-3} on MUTAG/ENZYMES, and 10^{-4} on the NCI datasets for optimal performance), and an equal length shortest-path kernel (Borgwardt & Kriegel, 2005). Results are presented in Table 1. In three out of four experiments the skew spectrum beats the other methods, including the shortest path kernel, which is considered state of the art for graphs of this type.

4. Conclusions

We have presented a new system of graph invariants, called the skew spectrum of graphs, based on purely

	MUTAG	ENZYME	NCI1	NCI109
Number of instances/classes	188/2	600/6	4110/2	4127/2
Max. number of nodes	28	126	111	111
Reduced skew spectrum	**88.61** (0.21)	25.83 (0.34)	**62.72** (0.05)	**62.62** (0.03)
Random walk kernel	71.89 (0.66)	14.97 (0.28)	51.30 (0.23)	53.11 (0.11)
Shortest path kernel	81.28 (0.45)	**27.53** (0.29)	61.66 (0.10)	62.35 (0.13)

Table 1. Prediction accuracy in percent of the (reduced) skew spectrum features and state of the art graph kernels on four classification benchmarks in 10 repetitions of 10-fold cross-validation. Standard errors are indicated in parentheses. Best results for each datasets are in bold.

algebraic considerations. From a mathematical point of view the skew spectrum is interesting because it brings a fundamentally new technique to constructing graph invariants. From a practical machine learning point of view the skew spectrum is interesting because it provides a powerful, yet efficiently computable representation for graph structured data instances. The skew spectrum also has natural extensions to hypergraphs and partially labeled graphs, where the invariance group is not the full symmetric group, but only a subgroup of the form $\mathbb{S}_{k_1} \times \mathbb{S}_{k_2} \times \ldots \times \mathbb{S}_{k_l}$. A longer version of this paper will appear in (Kondor & Borgwardt, 2008), and a compreshinse technical report will appear in (Kondor, 2008).

References

Borgwardt, K. M., & Kriegel, H.-P. (2005). Shortest-path kernels on graphs. *Proc. Intl. Conf. Data Mining* (pp. 74–81).

Borgwardt, K. M., Ong, C. S., Schonauer, S., Vishwanathan, S. V. N., Smola, A. J., & Kriegel, H. P. (2005). Protein function prediction via graph kernels. *Bioinformatics*, *21*, i47–i56.

Debnath, A. K., Lopez de Compadre, R. L., Debnath, G., Shusterman, A. J., & Hansch, C. (1991). Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *J Med Chem*, *34*, 786–797.

Gärtner, T. (2002). Exponential and geometric kernels for graphs. *NIPS*02 workshop on unreal data*. `http://mlg.anu.edu.au/unrealdata`.

Gärtner, T. (2003). A survey of kernels for structured data. *SIGKDD Explorations*, *5*, 49–58.

Gärtner, T., Flach, P., & Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. *COLT* (pp. 129–143). Springer.

Huang, J., Guestrin, C., & Guibas, L. (2008). Efficient inference for distributions on permutations. *Advances in Neural Information Processing Systems 20*.

Kakarala, R. (1992). *Triple corelation on groups*. Doctoral dissertation, Department of Mathematics, UC Irvine.

Kakarala, R. (1993). A group theoretic approach to the triple correlation. *IEEE Workshop on higher order statistics* (pp. 28–32).

Kondor, R. (2007a). A novel set of rotationally and translationally invariant features for images based on the non-commutative bispectrum. `http://arxiv.org/abs/cs.CV/0701127`.

Kondor, R. (2007b). The skew spectrum of functions on finite groups and their homogeneous spaces. `http://arxiv.org/abs/0712.4259`.

Kondor, R. (2008). The skew spectrum of graphs. To appear at `http://arxiv.org/`.

Kondor, R., & Borgwardt, K. M. (2008). The skew spectrum of graphs. *Proceedings of the International Conference on Machine Learning*.

Kondor, R., Howard, A., & Jebara, T. (2007). Multi-object tracking with representations of the symmetric group. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., & Schomburg, D. (2004). Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Research*, *32D*, 431–433.

Shawe-Taylor, J. (1993). Symmetries and discriminability in feedforward network architectures. *IEEE Transactions on Neural Networks*, *4*, 816–826.

Wale, N., & Karypis, G. (2006). Comparison of descriptor spaces for chemical compound retrieval and classification. *Proc. of ICDM* (pp. 678–689). Hong Kong.