ICML 2010 Travel Scholarship: Cover Sheet

Name:

Nathanael Fillmore.

Institution:

University of Wisconsin, Madison, Computer Sciences Department.

Address:

1210 W. Dayton St, Madison, WI 53706.

Email:

nathanae@cs.wisc.edu.

Phone number:

+1 (651) 769-3651.

Type of degree and expected completion date:

Ph.D., 2013.

Name of advisor:

Michael H. Coen.

Paper accepted at ICML or one of the workshops:

Accepted at ICML: Michael H. Coen, M. Hidayath Ansari, Nathanael Fillmore. Comparing Clusterings in Space.

To be submitted to Learning in Non-(geo)metric Spaces: Nathanael Fillmore, A topological approach to protein remote homology detection. (Acceptance notification on June 6.)[1]

Amount of support expected from the student's home institution, if any:

None.

Expected travel costs:

Registration: \$300 + \$35 = \$335.
Flight: \$1,800.
Hotel: \$150 × 6 (June 20–26) = \$900.
Bus to/from airport: \$10 × 2 = \$20.

[1] Mike: This is my course project for CS776. Are we also submitting a paper to this workshop on simdist? I assume it is okay to submit more than one paper to a workshop so long as I am only first author on one of them?

Total: \$3055.[2]

[2] Mike: Do these seem right to you? I found these prices on Orbitz.com, but \$1,800 seems high. Also, it would seem extremely reasonable to share a room and reduce hotel costs by half or more. Also, I don't know whether June 20–26 is a reasonable length of time to specify? The conference starts June 21, and the workshops take place on June 25.

ICML 2010 Travel Scholarship: Research Statement

Nathanael Fillmore

Two results from my research will be presented at ICML 2010. These are summarized below.

Comparing clusterings (ICML; joint with Michael H. Coen and M. Hidayath Ansari)**.**

Combining diverse metrics for remote homology detection (workshop submission)**.**

Two proteins are said to exhibit a remote homology if they have a moderately distant biological relation. Finding remote homologies based only on amino acid sequences is an important problem in computational biology, and numerous approaches exist. Many of the most successful approaches specify a biologically meaningful sequence similarity or dissimilarity measure and then apply a standard discriminative learning algorithm, e.g., [1, 2, 3, 4]. These similarity measures rely on diverse properties of the sequences, e.g., [4] compares local sequence alignments, while [2] compares k-mer count vectors. This diversity provides a reason to believe that detection can be improved by combining these similarity measures.

The problem can be formalized in a general setting, which encompases a wide variety of nonbiological problems, as follows:

We generalize our setting as follows. This general setting encompases numerous learning problems beyond our specific biological application. We assume our

Thus one can hope that detection could be improved

Many of the most successful approaches use (a) a notion of sequence similarity together with (b) a discriminative learning algorithm. Diverse similarity measures have been used, e.g., Fisher kernels [cite], string mismatch kernels,

When proteins are distantly related they are said to exhibit a remotoe

An important problem in computational biology is to determine whether two proteins are biologically related, i.e., are homologous, based only on their sequence of amino acids.

A topological approach to protein remote homology detection.

An important problem in computational biology is to determine whether two proteins are related or not, based only on their sequence of amino acids. The most successful approaches in this area have been kernel methods. Several kernels have been used, including

The idea: Let X denote a finite collection $\{x_1, \ldots, x_n\}$ of points in some Riemannian manifold $\mathcal{M}$. Note $\mathcal{M}$ comes with a distance function d. We assume no direct knowledge of X, $\mathcal{M}$, or d other than the cardinality of X. In fact our goal is to learn the restriction of d to $X \times X$. We do not need to have any particular representation of X (other than, say, an index for each element of X). We also assume there exist m Riemannian manifolds $\mathcal{M}_k$ and smooth maps $f_k : \mathcal{M} \to \mathcal{M}_k$ for $k = 1, \ldots, m$. We do not know what any of the $\mathcal{M}_k$ or f_k are. Being Riemannian, each manifold $\mathcal{M}_k$ comes with a distance function $d_k : \mathcal{M}_k \times \mathcal{M}_k \to [0, \infty)$. We do not know what any of the d_k are in full, but we do know their restrictions to the images $f_k(\mathcal{M}) \times f_k(\mathcal{M})$. That is, we are given

$$\{d_k(f_k(x_i), f_k(x_j)) : i, j = 1, \ldots, n,\ k = 1, \ldots, m\}.$$

Our goal is to learn a function $\hat{d} : X \times X \to [0, \infty)$ such that $\hat{d}$ agrees with d on X, i.e., $\hat{d}(x_i) = d(x_i)$ for all $x_i \in X$.

Clearly in general more than one function $\hat{d}$ is possible. We decide to choose the one that requires the least change. To be precise, we want all the f_k to be C-Lipschitz continuous for C as small as possible, i.e.,

$$d_k(f_k(x), f_k(y)) \leq C d(x, y) \quad \text{for all } x, y \in \mathcal{M}.$$

This is still not identifiable. We also impose the condition that d be an *intrinsic* metric, i.e., that $d(x, y)$ be equal to the length of the shortest path between x and y for any x, y:

$$d(x, y) = \inf\{L(\gamma) : \gamma(0) = x, \gamma(1) = y, \gamma : [0, 1] \to \mathcal{M} \text{ is continuous}\},$$

where $L(\gamma)$ is the length of the path γ, defined by

$$L(\gamma) = \sup_J \sup_{0=t_0 \leq t_1 \leq \cdots \leq t_J = 1} \sum_{j=1}^{J} d(\gamma(t_{j-1}), \gamma(t_j)).$$

Of course we can only

So our overall objective is:

$$\inf_{d, f_1, \ldots, f_m} \quad C$$

References

[1] Tommi Jaakkola, Mark Diekhans, and David Haussler. A discriminative framework for detecting remote protein homologies, 1999.

[2] Christina S. Leslie, Eleazar Eskin, Adiel Cohen, Jason Weston, and William Stafford Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–476, 2004.

[3] Li Liao and William Stafford Noble. Combining pairwise sequence similarity and support vector machines for remote protein homology detection. In *RECOMB '02: Proceedings of the sixth annual international conference on Computational biology*, pages 225–232, New York, NY, USA, 2002. ACM.

[4] Hiroto Saigo, Jean-Philippe Vert, Nobuhisa Ueda, and Tatsuya Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689, 2004.