Learning in Structured Input and Output Spaces

Brijnesh J. Jain and Klaus Obermayer

Berlin University of Technology
Dept. of Electrical Engineering and Computer Science
Sekr. 2-1, Franklinstr. 28/29, 10587 Berlin, Germany
E-mail: {bjj | oby}@cs.tu-berlin.de

In many applications in pattern recognition and machine learning, it is common practice to represent data by feature vectors living in a Banach space, because the Banach space provides powerful analytical techniques for data analysis usually not available for other representations. A standard technique to solve a learning problem in a Banach space is to set up a smooth error function, which is then minimized by using local gradient information.

But often, the data we want to learn about have no natural representation as feature vectors and are more naturally represented in terms of finite structures such as, for example, point patterns, strings, trees, lattices, or graphs. Such pattern recognition and machine learning problems arise in a variety of applications, ranging from predicting the biological activity of a given chemical structure, finding frequent substructures of a data set of chemical compounds, predicting the 3D-fold of a protein given its amino sequence, and natural language parsing, to name just a few. In contrast to feature vectors, the repository of algorithmic tools for data analysis that directly operate on structured representations is sparse. In particular, the concept of a derivative for functions on structures is missing, and therefore standard techniques based on local gradient information are no longer applicable.

Standard approaches fit the data representation with the tools available, for example, they transform structural data to feature vectors. The aim is to *fit practice to theory*. Possible pitfalls are (i) a bias toward the tools in the sense that preprocessing of the data might result in a loss of relevant structural information; and (ii) the reconstruction problem that asks for the data in their "natural" representation given a data point in the space we work with. There are two main classes of methods that apply the principle of assimilating the data to the tools, methods based on pairwise proximity data [1,2,4–6,10] and methods that transform structures into vector spaces [2,3,8,9,11].

In this article, we follow the complementary approach by *fitting theory to practice*. We *accommodate* existing tools with the data representations, i.e. we make the following contributions:

To adopt analytical concepts like continuity and differentiability to finite structures, we develop the theory of $\mathcal{T}$*-spaces*. Given a metric vector space $\mathcal{X}$ and a finite set $\mathcal{T}$ of orthogonal transformations on $\mathcal{X}$, a $\mathcal{T}$-space over $\mathcal{X}$ is defined by the quotient set $\mathcal{X}_\mathcal{T} = \{[\boldsymbol{x}] : \boldsymbol{x} \in \mathcal{X}\}$, where $[\boldsymbol{x}] = \{T\boldsymbol{x} : T \in \mathcal{T}\}$ denotes the equivalence class of $\boldsymbol{x} \in \mathcal{X}$ with respect to $\mathcal{T}$. The first result states that certain classes of finite structures, called k*-structures*, are subsets of $\mathcal{T}$-spaces. A k-structure is a finite structure that can be represented as a k-ary relation $X \subseteq \mathcal{X}^k$. Examples of k-structures are vectors, point patterns, trees, or graphs. Since vectors are also k-structures, the theory of $\mathcal{T}$-spaces generalizes the theory of vector spaces.

Suppose that $\mathcal{X}$ is a Banach space. The key result of this contribution states that the gradient of a smooth function $f : \mathcal{X}_\mathcal{T} \to R$ is a well-defined structure pointing in direction of steepest ascent. This result allows us to optimize functions on structures using local gradient information for which a vast amount of algorithms has been developed.

Based on the theory of $\mathcal{T}$-spaces, we present cost functions for a number of applications problems. These include the sample mean of a set of structures, mining frequent substructures, learning in non-metric distance spaces, and (un)supervised learning with metric $\mathcal{T}$-spaces as input and/or as output space. All proposed cost functions are locally Lipschitz and therefore only smooth almost everywhere. To minimize the cost functions, we can apply techniques from nonsmooth optimization. In a second contribution to this meeting [7], we consider $\mathcal{T}$-linear discriminant functions as a more detailed example of how to apply the theory of $\mathcal{T}$-spaces.

References

1. T. Gärtner. A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter*, 5(1):49–58, 2003.
2. T. Graepel, R. Herbrich, P. Bollmann-Sdorra, and K. Obermayer. Classification on pairwise proximity data. In *Advances in Neural Information Processing*, volume 11, pages 438–444,, 1999.
3. M. Hein, O. Bousquet, and B. Schoelkopf. Maximal margin classification for metric spaces. *Journal of Computer and System Sciences*, 71(3):333–359, 2005.
4. R. Herbrich, T. Graepel, P. Bollmann-Sdorra, and K. Obermayer. Learning a preference relation for information retrieval. In *Workshop Text Categorization and Machine Learning, International Conference on Machine Learning 1998*, page 8084, 1998.
5. S. Hochreiter and K. Obermayer. Support vector machines for dyadic data. *Neural Computation*, 18:1472–1510, 2006.
6. T. Hofmann and J. M. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(1):1–14, 1997.
7. B.J. Jain and K. Obermayer. T-linear discriminant functions. In *The Learning Workshop, Snowbird*, 2007.
8. B. Luo, R. C. Wilson, and E. R. Hancock. Spectral embedding of graphs. *Pattern Recognition*, 36(10):2213–2230, 2003.
9. H. Quang Minh and T. Hofmann. Learning over compact metric spaces. In *Proceedings of the 17th Annual Conference on Learning Theory (COLT 2004)*, 2004.
10. E. Pekalska, P. Paclik, and R.P.W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211, 2001.
11. U. von Luxburg and O. Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–665, 2004.

Topic: learning theory
Preference: poster