The Learning Workshop Cliff Lodge, Snowbird, Utah. April 3-6, 2012

Structure learning and the generalization capacity of algorithms

Joachim M Buhmann
Department of Computer Science, ETH Zurich
jbuhmann@inf.ethz.ch

Pattern recognition addresses the problem to find structure in data. The nature of this structure is defined by an *a priori given hypothesis class* $\mathcal{C}$ of possible data interpretations, e.g. data partitionings for clustering, embedding of relational data in Euclidean spaces or total orders in sorting of objects. The data analyst applies an algorithm $\mathcal{A}$ to select one hypothesis $c(\mathbf{X}) \in \mathcal{C}$ or a subset of hypotheses $\mathcal{C}_\gamma(\mathbf{X}) \subset \mathcal{C}(\mathbf{X})$ that "explain" the data $\mathbf{X}$. The parameter γ specifies the closeness of hypotheses to the target hypothesis $c(\mathbf{X})$. Mathematically, an algorithm establishes a relation between input data $\mathbf{X} \in \mathcal{X}$ out of a data space $\mathcal{X}$ and output hypotheses in a hypothesis class $\mathcal{C}$, i.e., $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{C}$. Often but not always, such an algorithm optimizes a quality measure or, equivalently, minimizes some cost or risk. An exception for clustering is e.g. single linkage which does not optimize a cost function.

The robustness of algorithms poses a key conceptual problem when information processing is affected by noise. Frequently, algorithms return significantly different hypotheses to new (test) data which contain the same signal as previous (training) data but differ in the fluctuations. In a machine learning sense, the algorithm does not generalize well! How should can we define and measure this generalization ability of an algorithm? Following classical statistics, we could try to estimate the probability distribution of the data and with this estimate, we could calculate the statistical risk of the method. However, many information processing problems are characterized by data spaces which are much larger than the hypothesis class of data interpretations. In such a setting, we might never get enough data to estimate the data distribution, but the information in the instances is sufficient to estimate a probability distribution of the hypotheses[1].

We advocate an information theoretic perspective for structure inference in particular in large data set. The fluctuations in the measurements $\mathbf{X}$ quantize the hypothesis space of the pattern recognition algorithm $\mathcal{A}$. This concept has been explicitly developed for data clustering algorithms, which minimize cost functions [2], and for general optimization problems [3]. The information theoretic framework has the following components: a *Problem Generator* $\mathfrak{PG}$ generates data $\mathbf{X} \sim \mathbb{P}(\mathbf{X})$, a *Sender* $\mathfrak{S}$ defines a code and sends one of the code messages, a *Receiver* $\mathfrak{R}$ decodes the sent message. To establish a code, $\mathfrak{PG}$ generates a training data set $\mathbf{X}^{(1)}$. The sender then calculates an approximation set $\mathcal{C}_\gamma(\mathbf{X}^{(1)}) \subset \mathcal{C}$ to the hypothesis $\mathcal{A}(\mathbf{X}^{(1)})$. If $\mathcal{A}$ minimizes costs $\mathcal{R}\left(\mathcal{A}(\mathbf{X}^{(1)})\right)$ then the approximation set is defined by e.g. Boltzmann weights. For pattern recognition algorithms which do not minimize costs with a partial order of all hypotheses, we adapt the concept of smoothed analysis [6] to generate a set of "close" solutions to the hypothesis $\mathcal{A}(\mathbf{X}^{(1)})$. In a second step, the sender selects a set of transformations $\mathbb{T}$ in such a way[2] that the union of the resulting approximation sets $\bigcup_{\tau\in\mathbb{T}} \mathcal{C}_\gamma(\tau \circ \mathbf{X}^{(1)})$ covers the hypothesis space. The set $\mathbb{T}$ is communicated to the receiver as the communication code.

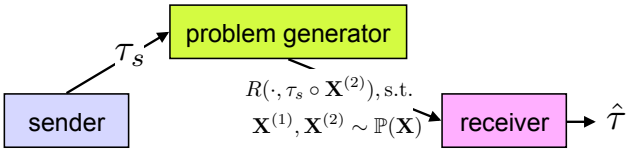

Figure 1. Communication process: (1) the sender selects transformation τ_s, (2) the problem generator draws $\mathbf{X}^{(2)} \sim \mathbb{P}(\mathbf{X})$ and applies τ_s to it, and the receiver has to estimate $\hat{\tau}$ based on $\tilde{\mathbf{X}} = \tau_s \circ \mathbf{X}^{(2)}$.

For communicating messages as depicted in fig. 1, the sender selects a transformation τ_s and send it to the problem generator. $\mathfrak{PG}$ generates a second (test) data set $\mathbf{X}^{(2)}$, applies τ_s to the new data and sends the resulting $\tilde{\mathbf{X}} = \tau \circ \mathbf{X}^{(2)}$ to the receiver. Decoding the message requires to estimate the transformation $\hat{\tau}$ based on the observed data $\tilde{\mathbf{X}}$. The decoding procedure selects the transformation where the respective approximation set has the maximal overlap with one of the approximation sets based on $\mathbf{X}^{(1)}$, i.e.,

$$\hat{\tau} = \arg\max_{\tau\in\mathbb{T}} \mathcal{C}_\gamma(\tau \circ \mathbf{X}^{(1)}) \cap \mathcal{C}_\gamma(\tilde{\mathbf{X}}) \qquad (1)$$

The condition $\mathbf{P}(\hat{\tau} \neq \tau_s | \tau_s, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}) \stackrel{n\to\infty}{\longrightarrow} 0$ of van-

[1] Consider e.g. figure ground segmentation of an image with n pixels which yields a data space of size 256^n. In case that we characterize the image by pairwise comparisons of pixel neighborhoods, we derive a graph with n vertices and $n(n-1)/2$ real valued similarities. The resulting cardinality of the data space amounts to $\mathcal{O}(\mathbb{R}^{n(n-1)/2})$. In contrast to the size of the data space, the cardinality of the hypothesis space is $|\mathcal{C}| = 2^n$.

[2] $\mathbb{T}$ is the set of permutations for combinatorial optimization problems or the set of shifts/rotations for continuous localization/orientation problems.

ishing decoding error defines a design criterion for a code with maximal information content. We have to cover the hypothesis class with a maximal number of distinguishable approximation sets where the messages τ_s can be decoded with vanishing error. A large deviation analysis of the conditional probability of error yields an upper bound on the rate $\rho = \log_2 |\mathbb{T}|$, i.e.,

$$\rho < \mathcal{I}_\gamma(\tau_s, \hat{\tau}) \equiv \frac{1}{n} \log \frac{|\mathcal{C}| \; |\mathcal{C}_\gamma^{(1\&2)}|}{|\mathcal{C}_\gamma^{(1)}| \; |\mathcal{C}_\gamma^{(2)}|} \; . \tag{2}$$

$|\mathcal{C}_\gamma^{(1)}|, |\mathcal{C}_\gamma^{(2)}|$ denote the cardinalities of the approximation sets for data $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$, respectively; $\mathcal{C}_\gamma^{(1\&2)}$ specifies the set of hypotheses which jointly approximate both data sets. $\mathcal{I}_\gamma(\tau_s, \hat{\tau})$ can be interpreted as a mutual information with respective Gibbs distributions as arguments [3]. Model selection is achieved by maximizing $\mathcal{I}_\gamma(\tau_s, \hat{\tau})$ w.r.t. γ. More importantly, we can also select the algorithm $\mathcal{A}$ from a set of algorithms according to the ranking induced by $\mathcal{I}_\gamma$. Algorithms with high generalization capacity are preferable since they are more robust against noise and more sensitive to signal than alternatives.

This selection concept for algorithms is supported by empirical evidence in model validation problems for relational data clustering [4]. For a correlation matrix of gene expression data gathered from the mussel Mytilus Galloprovincialis [1], pairwise clustering produced a more informative clustering than both normalized cut and correlation clustering. Furthermore, denoising of Boolean matrices guided by the generalization capacity of SVD suggests a cutoff rank for the SVD spectrum [5].

In a recent study, we have measured the sensitivity of sorting algorithms to errors in pairwise comparisons of items. Figure 2 shows the capacity of various sorting algorithms and their bit rate of extracted information per computation step, e.g. per comparison in the case of sorting. The study clearly demonstrates that robust algorithms like `BubbleSort` invest their excess comparisons to compensate for fluctuations. This computational redundancy increases the capacity of the algorithm and yields an improved localization ability in the hypothesis class. Computationally efficient methods like `MergeSort` perform superior in the noiseless case but sacrifice capacity for computational speed in the highly noisy case.

In principle, this concept of measuring the generalization performance of algorithms can be applied to algorithm evaluation and also to robust algorithm design. It endows the space of algorithm with a topology since two algorithms are neighbors if their approximation sets for the same input distributions share a high overlap. We are convinced that the information theoretic analysis of algorithms will shed new light on the relation between computational complexity and statistical complexity.

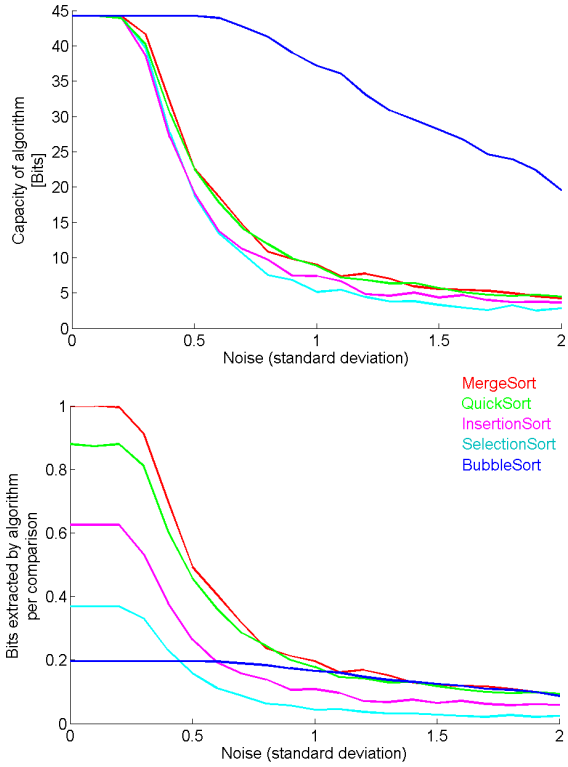

Figure 2.

Acknowledgement: Research on approximate sorting is joint work with L. Busse and M. Chehreghani.

[1] M. Banni, A. Negri, F. Mignone, H. Boussetta, A. Viarengo, and F. Dondero. Gene expression rhythms in the mussel Mytilus galloprovincialis (lam.) across an annual cycle. *PLoS ONE*, 6(5):e18904, 05 2011.

[2] J. M. Buhmann. Information theoretic model validation for clustering. In *International Symposium on Information Theory, Austin Texas*. IEEE, 2010. (http://arxiv.org/abs/1006.0375).

[3] J. M. Buhmann. Context sensitive information: Model validation by information theory. In J.-F. M.-T. et al., editor, *MCPR 2011*, volume 6718 of *LNCS*, pages 21–21. Springer, 2011.

[4] M. H. Chehreghani, A. G. Busetto, and J. M. Buhmann. Information theoretic model validation for spectral clustering. In *AISTATS 2012, Barcelona*, 2012. (in press).

[5] M. Frank and J. M. Buhmann. Selecting the rank of svd by maximum approximation capacity. In *International Symposium on Information Theory, St. Petersburg*, pages 1036 – 1040. IEEE, 2011.

[6] D. Spielman and S.-H. Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM*, 51(3):385 – 463, 2004.