AI Magazine Volume 22 Number 2 (2001) (© AAAI)



# *Unsupervised Learning: Foundations of Neural Computation*

## A Review

*DeLiang Wang*

The resurgence of the field of neural networks in the 1980s was primarily fueled by supervised learning, exemplified by the backpropagation algorithm. In supervised learning, a desired output signal is provided to the learner together with an input signal, and the system adjusts parameters so that its response in the future will be closer to the desired signal.

Although supervised learning has been dominant in machine learning, much of our intelligence, in particular, perception, is acquired without a teacher. Through mere exposure, humans and animals learn how to analyze their environments and recognize relevant objects and events. For example, consider our experience of sorting out apples from oranges by their appearances, an ability that can be gained before naming them. This analysis calls for unsupervised learning—learning without a teacher, also known as *self-organization*. Unsupervised learning has been studied in neural networks since the early days. However, in recent years, there has been a steady shift in the research focus from supervised learning to unsupervised learning, and the latter now becomes a predominant subject in neural networks. *Unsupervised Learning: Foundations of Neural Computation* is a collection of 21 papers published in the journal *Neural Computation* in the 10-year period since its founding in 1989 by Terrence Sejnowski. *Neural Computation* has become the leading journal of its kind. The editors of the book are Geoffrey Hinton and Terrence Sejnowski, two pioneers in neural networks. The selected papers include some of the most influential titles of late, for example, "What Is the Goal of Sensory Coding" by David Field and "An Information-Maximization Approach to Blind Separation and Blind Deconvolution" by Anthony Bell and Terrence Sejnowski. The edited volume provides a sample of important works on unsupervised learning, which cut across the fields of AI, neuroscience, and psychology.

The central issue in unsupervised learning concerns its goal. What do we want the system to learn without giving external instruction? There is no simple answer to this critical question. In fact, many different objectives have been proposed, including to discover clusters in the input data, extract features that characterize the input data more compactly, and uncover nonaccidental coincidences within the input data.

*Unsupervised Learning: Foundations of Neural Computation, eds. Geoffrey Hinton and Terrence J. Sejnowski, The MIT Press, Cambridge, Massachusetts, 1999, 398 pp., ISBN 0-262-58168-X.*

Beneath these objectives is the fundamental task of representation: Unsupervised learning attempts to derive hidden structure from the raw data. This endeavor is meaningful because input data are far from random; they are produced by physical processes. For example, a picture taken by a camera reflects the luminance of physical objects that constitute the visual scene, and an audio recording reflects acoustic events in the auditory scene. Physical processes tend to be coherent; an object occupies a connected region of the space, has a smooth surface, moves continuously, and so on. From the information theory standpoint, physical objects and events tend to have limited complexity and can be described in a small number of bits. This observation is, in my view, the foundation of unsupervised learning. Because perception is concerned with recovering the physical causes of the input data, a better representation should reveal more of the underlying physical causes.

Physical causes are hidden in the data, and they could, in principle, be revealed by unsupervised learning. However, there is an enormous variety of physical causes; trees have different colors, have textures, leave patterns, and so on, and they all look very different from animals. Without external supervision, the best unsupervised learning can achieve is to uncover generic structure that exists in a variety of physical causes. Fortunately, guided by some general assumptions or principles, there are plenty of interesting problems to solve.

One general principle for unsupervised learning is minimum entropy proposed in Barlow's article. The idea is that the derived representation should minimize redundancy (correlation) contained in the input data. The goal is similar to that pursued in communication theory: to minimize the bandwidth needed for signal transmission. Closely associated is the minimum–description length principle advocated in the Zemel and Hinton article on learning population codes. Another principle, put forward in Field's article, is sparse coding: The goal of the representation is to minimize the number of units in a distributed network that are activated by

any given image. In the article, Field argues systematically for such a representation in the mammalian visual system. Other general principles include maximizing mutual information between the input and the output of the system and deriving mutually independent feature vectors.

In a less obvious way, one can view unsupervised learning as supervised learning with no input, treating the data as the output of the system. The representation to be derived is then viewed as a model for the input data. This is the generative approach embodied in the Helmholtz machine introduced in the article by Dayan et al. According to this approach, the goal of unsupervised learning is to model the probability density of the input data. The generative approach can be traced back to the Boltzmann machine (Ackley, Hinton, and Sejnowski 1985).

Unsupervised learning algorithms commonly use two techniques: (1) optimization and (2) Hebbian learning. The previous discussion on the goal of unsupervised learning makes it clear that learning algorithms almost invariably boil down to an optimization problem, whether to minimize entropy or maximize mutual information. The Hebbian learning rule states that the connection between two neurons is strengthened if they fire at the same time (Hebb 1949), which is supported by strong biological evidence. The anti-Hebbian rule, which weakens the connection when two neurons fire simultaneously, also proves useful. The utility of the Hebbian (anti-Hebbian) rule in unsupervised learning should not come as a surprise because the Hebbian rule is about correlation (anticorrelation), the detection of which is a central task for unsupervised learning.

The method of *independent component analysis* (ICA), which attempts to identify statistically independent causes from their mixtures, has recently generated considerable excitement in the broad area of signal processing. The idea of ICA is equivalent to the minimum entropy principle, and unsupervised learning produces algorithms for deriving independent components through training with mixture samples (the articles by Bell and Sejnowski, Amari, and Hyvörinen and Oja). In the last few years, ICA has been applied with impressive success to an array of real-world problems, including medical data analysis (for example, EEG) and the cocktail party problem for decomposing acoustic mixtures.

A related success is the development of model neurons whose response properties resemble the receptive fields of simple cells in the mammalian visual cortex. Simple cells possess receptive fields that can be characterized as oriented and spatially localized bandpass filters, best described by Gabor filters. It is remarkable that such receptive fields can emerge as a result of applying unsupervised learning to an ensemble of natural images (as in the Atick and Redlich article) (Bell and Sejnowski 1997; Olshausen and Field 1996). These results provide a computational basis for reasoning about general strategies used by the brain for sensory processing.

Most unsupervised learning algorithms are based on statistical estimation of the input data. As pointed out in the Konen and von der Malsburg article, such algorithms generally suffer from the problem of combinatorial explosion when dealing with realistically large patterns. They proposed incorporating structure, specifically the prior principle of conservation of topological structure, into their self-organization network for symmetry detection (see also the Gold et al. article). Their article emphasizes geometric principles, rather than statistical principles, for unsupervised learning. It is revealing to consider the old Minsky-Papert connectedness problem (Minsky and Papert 1969) in this context. This problem is one of telling connected patterns from disconnected ones. On a two-dimensional grid, there are exponentially many connected patterns. In theory, one could get a multilayer network to learn the connectedness predicate. However, as pointed out by Minsky and Papert (1988), it is practically infeasible because it requires far too many training samples and too much learning time. Not until recently was a neural network solution found, and the solution to the problem is based on a simple architecture with primarily nearest-neighbor coupling and an oscillatory correlation representation that labels pixels by synchrony and desynchrony (Wang 2000). This solution echoes the point of Konen and von der Malsburg on the importance of prior structure. From the philosophical point of view, the brain of a newborn possesses genetic knowledge resulting from millions of years of evolution. Although, in theory, all is learnable, including connectivity and representation, computational complexity has to be an important consideration. Hence, future investigation on unsupervised learning needs to incorporate appropriate prior structure.

In summary, this book is essential reading for professionals and graduate students who work on sensory encoding, perceptual processing, and machine learning. It is also a valuable source for engineers working in the areas of computer vision, speech processing, and communication.

## References

Ackley, D. H.; Hinton, G. E.; and Sejnowski, T. J. 1985. A Learning Algorithm for Boltzmann Machines. *Cognitive Science* 9(2): 147–169.

Bell, A. J., and Sejnowski, T. J. 1997. The "Independent Components" of Natural Scenes Are Edge Filters. *Vision Research* 37(23): 3327–3338.

Hebb, D. O. 1949. *The Organization of Behavior.* New York: Wiley.

Minsky, M. L., and Papert, S. A. 1988. *Perceptrons.* Expanded ed. Cambridge, Mass.: MIT Press.

Minsky, M. L., and Papert, S. A. 1969. *Perceptrons.* Cambridge, Mass.: MIT Press.

Olshausen, B. A., and Field, D. J. 1996. Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature* 381:607–609.

Wang, D. L. 2000. On Connectedness: A Solution Based on Oscillatory Correlation. *Neural Computation* 12:131–139.

**DeLiang Wang** is an associate professor of computer and information science and cognitive science at The Ohio State University. His research interests include neural networks for perception, neurodynamics, and computational neuroscience. His e-mail address is dwang@cis.ohio-state.edu.