Abstracta SPECIAL ISSUE III, pp. 42 – 46, 2009

COMMENTARY ON "RELIABLE REASONING"

Stephen José Hanson

This book is a wonderful redux to a time in the early half of the 20th century when statistical learning theory was just developing, and when new methods and concepts were being discovered. By the end of the 1970s learning theory had been sidetracked by the need for practical results that were consistent with the prevailing dogma of AI (artificial intelligence). Proponents of AI had posited that human thought was propositional in fact, specifically related to some unknown logical form. Consequently, much of the technology and computer science was devoted to developing the LISP or Prolog language that would somehow be more consistent the "programming language" of the mind. Early work in Machine Learning (1970s-1980s) focused on algorithms that had some functional relationship to some fundamental human task: categorization, perceptual object recognition, language understanding, reasoning, for example. But the algorithms were often too narrowly focused on these types of tasks and were termed as "brittle", in that small variations in the input conditions caused the algorithm to completely fail. One of the often repeated stories that was associated with the abrupt decline of AI during the late 1980s, was a DARPA project involving autonomous navigation of a tank in off-road environments. With the assembled VIPs of DARPA project Managers, Directors and military brass, the tank performed flawlessly through a set of standard obstacles and off-road variations, until, the sun went behind the clouds, causing the tank to immediately take a right turn into a tree continually ram the tree over and over again. After this event and others of a similar vein, much of the AI funding at major centers (MIT, CMU) was almost entirely cut. Insult was added to injury when a graduate student at CMU (Dean Pomerleau) developed a modest neural network (he dubbed ALVNN-autonomous learning vehicle Neural Network) that could be trained under various weather, road and obstacle conditions by driving the vehicle for a few hours in diverse conditions and thereafter would perform flawlessly on the same tasks as the AI programmed tank including varied lighting conditions which proved disastrous for the more brittle AI system. ALVINN morphed over time to other autonomous vehicle systems based on combinations of Neural network technology and various systems integration methodology that recently and ironically won the 1M$ DARPA off-road autonomous vehicle competition (Thrun et al, in press).

By the 1990s the Machine Learning field was bankrupt. Conference attendance declined and many of the leaders in the field began to attend the emerging and soon to be premier Neural Network conference , Neural Information Processing (NIPS), and retooling in Neural Network methods. Through the 1990s innovation between statistical learning theory and neural network architectures and various learning algorithms (esp. Reinforcement Learning; general kernel methods, and Bayesian search methods) that all began to flourish in the "wild west" of neural computation. I noticed this when I was Program Chair of NIPS in 1992, and wrote in the introduction of that years' volume, my sense that Neural Computation was becoming a mosaic or patchwork of statistical learning theory, neural network architectures and algorithms and goals from the now defunct AI/Machine Learning field. I argued that this type of diversity at the time was to be expected for a field that was absorbing so many directions ideas and tensions from other fields. NIPS was turning into a kind of hothouse, a kind of Cambrian period for speciation, in effect an engine for change and diversity. In the 1992 meeting, we carefully struck a chord between the emerging statistical learning science and the cyberpunk neural network slash and burn –"I can make a robot dance" applications. In the first session, which I chaired, I recall having 3 talks in a row, first Michael Kearns (*Estimating Average-Case Learning Curves Using Bayesian, Statistical Physics and VC Dimension Methods* - David Haussler, Michael Kearns, Manfred Opper, and Robert Schapire) who was developing constraints and bounds on learning algorithms more generally, John Moody, who had done a theoretical analysis of why Neural Networks would learn under conditions where they were over parameterized (which of course for the first 9 years was most of the time!) and showed remarkably that they automatically, adaptively modeled the underlying data complexity and developed the notion of "effective number of parameters" (*The Effective Number of Parameters:An Analysis of Generalization and Regularization in Nonlinear Learning Systems* - John E. Moody*)* finally Vladimir Vapnik (*Principles of Risk*

Minimization for Learning Theory - Vladimir Vapnik) who revealed a new concept in risk minimization and a new method that had yet to be tried on any data-something he later called a Support Vector Machine (SVM). You could sense the audience was stirred, that they knew they just heard earth-shattering ideas and directions, but had no idea where it would go, just that field was about to explode in size and directions in what would soon be called the "learning sciences".

Vapnik's paper is of particular interest in the context of the Harman and Kulkarni book. It wasn't till the next year that Vapnik's new algorithm was coded by ATT folks (L. Boutou, I. Guyon) and tested on some difficult (handwritten zip code data) data sets and shown to perform remarkably well. The method was particularly interesting since it appeared not be as ad hoc as other methods that also were engineered to work well with these data sets (Yann Le Cun's "LeNet"), at the same time it appeared to (purposely) lose information about the underlying probability distributions—not a great idea if one wants to understand the feature space. Nonetheless, the lack of underlying explanation of why a classifier works well, never seemed to deter the NIPS crowd and it was found to be remarkably robust in generalization on very hard problems and allowed simple variations leading to a new field now beginning to be called "Large Margin Classifiers". Vapnik had in fact, by theory and development of an accessible, easily generalizable algorithm, created a new subfield in Neural Computation (not by the way in Machine Learning, as often erroneously quoted), that began to draw twice the number of attendees to the NIPS workshop than to the entire conference held the week before (not so much these days).

SVM and its Kernel or Large Margin or whatever it is called now –classifiers have become de rigueur for solving hard classification problems. Introduction of nonlinear kernels and the "kernel trick", slack variables and Cost parameters as well as ways to visualize ALPHA per voxel, made the method truly useful and frankly what one turns to when other approaches fail. I have recently used it with Brain Imaging Data (Hanson & Halchenko, 2008) and found it both amazing and annoying. We were able to classify brain images from subjects viewing either pictures of Faces or pictures of Houses using full brain (40k voxels) with 92% out-of-sample cross validation for all subjects. Although remarkable, SVM does not really allow one to examine the diagnosticity of the underlying features without some strong assumptions which unfortunately are probably not true. Nonetheless, we forged ahead using a method called Recursive feature elimination which allowed us to titrate down to the necessary(?) number of voxels required to maintain high generalization accuracy. We were able to find 100s of contiguous voxels that are most diagnostic for these two categories, which unfortunately for some theories of cognitive/perceptual modularity, were highly overlapped. But that's another story.

The reason for imparting all of this nitty-gritty detail is to try and bring SVM a bit down to earth as well as the notions of transduction that are framed so beautifully in the book. One point in passing is that it is highly unlikely that SVM has anything to do with human cognition or categorization. As Harman and Kulkarni point out, it has no way to express a “prototype”, again due to the lack of a probability distribution in the implementation of the method (also categorical perception is not really an example of SVM…but this is also a long story—see Harnad, Hanson & Lubin, 1991). Of course, this is the very tactic that allows it to attack such high-dimensional spaces. On the other hand this is not quite the same as solving the so called “curse of dimensionality” which is often confused with the power of SVM in classifying problems, which has more to do with its ability to pick good examples per class. Frankly, SVM, as my friend Yann Le Cun once put it, is nothing more than a dumb Perceptron, that can find interesting exemplar cases but can't remember how many it found. Now this is a bit harsh, and probably missing the larger picture of this remarkable idea. Nonetheless, there is a bit of over-romanticism here concerning SVM and I can't help to think that the Harman and Kulkarni treatment of SVM is not unlike Woody Allen's view of Manhattan: “'Chapter One. He adored New York City. He idolized it all out of proportion.' Uh, no, make that: 'He-he...romanticized it all out of proportion. Now...to him...no matter what the season was, this was still a town that existed in black and white and pulsated to the great tunes of George Gershwin.'” SVM is an interesting statistical learning theory tool ...but SVM ..Gershwin.. I don't think so.

Stephen José Hanson
Rutgers University
jose@psychology.rutgers.edu

References

Harnad, S., Hanson, S.J. & Lubin, J. (1991). ‘Categorical Perception and the Evolution of Supervised Learning in Neural Nets’. IN D.W. Powers & L. Reeker (eds.), *Working Papers of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology*, pp. 65-74.

Hanson, S. & Halchenko, Y.O. (2008). ‘Brain Reading Using Full Brain Support Vector Machines for Object Recognition: There is no “Face” Identification Area’. *Neural Computation*, 20, 2:486-503.

Hanson, S. J. Cowan, J.D., & Giles, C.L. (1993) ‘Advances in Neural Information Processing Systems 5’, *NIPS Conference*, Denver, Colorado, USA, 1992], Morgan Kaufmann 1993.

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., Gale, J., Halpenny, M., Hoffmann, G., Lau, K., Oakley, C., Palatucci, M., Pratt, V., Stang, P., Strohband, S., Dupont, C., Jendrossek, L.-E., Koelen, C., Markey, C., Rummel, C., van Niekerk, J., Jensen, E., Alessandrini, P., Bradski, G., Davies, B., Ettinger, S., Kaehler, A., Nefian, A. and Mahoney, P. ‘Stanley, the robot that won the DARPA Grand Challenge’. *Journal of Field Robotics*, In Press.