Machine Learning in Infrastructure Security

V Rao Vemuri

Center for Digital Security
Department of Applied Science
Department of Computer Science
University of California, Davis, USA
rvemuri@ucdavis.edu

Extended Abstract: Democratic societies throughout the world, it appears, are facing a new type of threat dubbed "asymmetric threat." In this new threat environment the world governments are faced with a number of low-intensity conflicts characterized by less discriminate attacks on civilian populations, infrastructures and the like. For the purposes of this paper, an asymmetric war is defined as an environment in which a government is faced with a faceless and stateless enemy. Even states can hide behind the façade of a stateless terrorist group and conduct attacks. Contemporary cross-border terrorism is an example.

The advent of Internet is a mixed blessing in this context. On one hand the Internet gives the terrorist organizations an opportunity to open a new front in the shape of cyber warfare. On the other hand the internet-based technologies can be used to fight back terrorism. In defense establishments, an often-suggested solution to this terrorism problem is to spin strands of Information Technology (IT) that would weave data gathered from diverse sensors into a vast electronic dragnet. Some defense technologists point out that the types of computerized data sifting and pattern matching that might flag suspicious activities are not much different from programs already in use by private companies.

In the civilian world, digital technologies like e-mail, online shopping and travel booking, automatic teller machines used by banks, cell phone networks and credit-card payment terminals are already gathering information about people and their transactions. Now it is possible to link for the first time such different electronic sources as video feeds from airport surveillance cameras, credit card transactions, airline reservations and telephone calling records. The data could be filtered through software that would constantly look for suspicious patterns of behavior.

What is needed in this new threat environment is a human-computer team that would dramatically improve the capability of human analysts to make inferences in complex domains - both data-rich and data-poor. Such a team would vastly increase the ability to identify key facts hidden in immense quantities of irrelevant information, to assemble large numbers of disparate facts in order to reach valid conclusions, and to produce new patterns that assist future analyses. To realize this vision, improvements are necessary in the state of the art in knowledge discovery, data mining, and machine learning (KDD-ML) such that a human-computer team can:

Learn using prior knowledge - Make effective use of a wide variety of knowledge sources, including common-sense knowledge bases, domain-specific knowledge bases, and direct interaction with human experts.

- Learn actively - Request new data and analyses that optimally improve learning and inference.
- Learn incrementally and cumulatively - Incrementally improve existing knowledge, and make use of that knowledge in subsequent learning and inference.
- This talk addresses these issues by focusing attention on the discovery of new knowledge - that is knowledge that we do not already possess.

Although this discovery does not necessarily mean a first person experience, the problem suggests that there is a person (or an agent) *searching* for knowledge. In data-rich environments, this activity typically is not just a search for a needle in a haystack – rather, it is like assembling a needle from pieces of several needles strewn in a haystack. In data-poor environments, this is like deducing the shape of an extinct, prehistoric animal (an inverse problem) from a bone fragment.

"How will you look for it, Socrates, when you do not know at all what it is? How will you aim to search for something you do not know at all? If you should meet with it, how will you know that this is the thing that you did not know?

"Do you know what a debater's argument you are bringing up, that a man cannot search either for what he knows, or for what he does not know? He cannot search for what he knows – since he knows it, there is no need to search – nor for what he does not know, for he does not know what to look for" [3].

This learning paradox, called Meno's dilemma, says that "we cannot learn what we do not already know because we are unable not only to search for it, but also to recognize it should we stumble on to it."

"In going beyond what is already known, one cannot but go blindly. If one can go wisely, this indicates already achieved wisdom of some general sort" [1]. Reasoning, therefore, which is a wise method of obtaining seemingly new knowledge, is nothing but an exploration of what is already implicitly known. The discovery of truly new knowledge needs investigation along unconventional lines. That is what this proposal is all about.

Instead of putting the burden on the subject alone in the search for new knowledge, it stands to reason to bring in the environment, in which the knowledge resides, as a participant. That is, the subject (an agent) and the knowledge embedded in its environment are being treated here as two sides of the same coin. This perspective harks back at the age-old definition of a "system" as "a plant and its environment" and Tichonov's regularization procedure [4, 5], which is a mathematical statement of this desire to bring the information embedded in the environment in the form of auxiliary conditions and constraints. This perspective gives some hope for success in finding mathematical theories on which to build the proposed knowledge discovery edifice.

Using this unified framework, the brain (or the mind) can bypass Meno's dilemma in the same way biological evolution avoids the necessity of a Creator; that is, utilize a *variation mechanism* to make new discoveries. The origin of new knowledge in the information space is analogous to the origin of new species in the Darwinian space.

The central issue here is *discovery* of knowledge. That is, how is it possible for a human to acquire new knowledge? First of all, it should be noted that there does not exist any method or algorithm for the discovery of features of a totally unknown environment. Once the agent knows what it is looking for, it is easy enough to design a search algorithm to bring it from its previous state of ignorance to its new state of knowledge. But such a method would defeat the purpose because it would implicitly utilize *a posteriori* knowledge in the design of a knowledge *discovery* system.

Evolution does not *search* for new adaptations. Evolution may *find* new adaptations, but does so *without* searching. During evolution, the evolving structure may go through numerous variations. These variations are produced without a goal in mind. In a way, evolution produces solutions to non-existing problems. Those representing adaptations to non-existent environments will soon perish. Stated differently, one can visualize a *selection* process operating on the environment to produce a structure and *variation* as a blind (or random) mechanism that modifies the structure.

In phylogenetic adaptation, the structure (here, the species) has two components: phenotype and genotype. Similarly, in ontogenetic adaptation, the structure (here, the model) also has two components: neural activity pattern and synaptic configuration. Since ontogenetic adaptation has been viewed widely as a *learning* process, much research continues to get devoted to finding suitable structures that constitute a good learner. This is perhaps one reason for the popularity and success of connectionism.

Biological evolution, however, suffers from what Lorenz calls "the generational dead time," which refers to the time required to "generate and test" a new variation [2]. A potential source of variation that can operate at a faster pace would be the spontaneous neural background activity. (It is useful to assume here that knowledge has a neural correlate – i.e., an act of knowing corresponds to some neural activity.)

The talk begins with an outline of the challenges in a broad context and concludes with some specific results obtained in the context of computer network security.

References

1 Campbell, D. T. (1974) "Evolutionary Epistemology," in P. A. Schlipp (ed.) *The Philosophy of Karl Popper,* pp 413-463, Open Court, LaSalle. Reprinted in (eds.) G. Radnitzky and W. W. Bartley III (1987), *Evolutionary Epistemology, Rationality and the Sociology of Knowledge*, Open Court, LaSalle.

2 Lorenz, K. (1977*) Behind the Mirror: A Search for a Natural History of Human Knowledge,* Methuenand Co. Ltd., London

3 Plato, (9181) "Meno," in *Five Dialogues,* (Translation by G, M, A, Grube), Hackett, Indianapolis, USA

4 Tenorio, L. (2001) "Statistical Regularization of Inverse Problems," *SIAM Review,* Vol. 43, No. 2, pp 347-366.

5 Tichonov, A. N. (1963) "Regularization of Incorrectly Posed Problems," *Soviet Math. Dokl.* Vol. 4, pp 1624-1627.