Learning in Open-Ended Dynamic Distributed Environments

Doina Caragea
Artificial Intelligence Research Laboratory
Department of Computer Science
Iowa State University, Ames, IA 50011
dcaragea@cs.iastate.edu

In some domains (e.g., molecular biology), data repositories are large in size, dynamic, and physically distributed. Consequently, it is neither desirable nor feasible to gather all the data in a centralized location for analysis. Hence, efficient distributed learning algorithms that can operate across multiple data sources without the need to transmit large amounts of data and cumulative learning algorithms that can cope with data sets that grow at rapid rate are needed.

The problem of *learning from distributed data* can be summarized as follows: data is distributed across multiple sites and the learner's task is to discover useful knowledge from all the available data. For example, such knowledge might be expressed in the form of a decision tree or a set of rules for pattern classification. A distributed learning algorithm L_D is said to be *exact* with respect to the hypothesis inferred by a learning algorithm L, if the hypothesis produced by L_D, using distributed data sets D_1 through D_n is the same as that obtained by L when it is given access to the complete data set D, which can be constructed (in principle) by combining the individual data sets D_1 through D_n.

Our approach to distributed learning is based on a decomposition of the learning task into *information extraction* and *hypothesis generation* components. This involves identifying the information requirements of a learning algorithm and designing efficient means of providing the needed information to the hypothesis generation component, while avoiding the need to transmit large amounts of data. This offers a general strategy for transforming a batch or centralized learning algorithm into an exact distributed algorithm. In this approach to distributed learning, only the information extraction component has to effectively cope with the distributed nature of data in order to guarantee provably exact learning in the distributed learning.

We have used this approach to construct provably exact distributed algorithms for support vector machines and also for decision tree learning from horizontally as well as vertically distributed data by gathering sufficient information used further to generate the hypothesis (Caragea, Silvescu, and Honavar, 2000). Our definition of *sufficient information* for a data set is relative to a specific learning algorithm, e.g. a decision tree that implements a particular search strategy through the space of decision trees; or an algorithm that chooses a maximal margin hyperplane for a binary classification task as in the case of SVM algorithm. Thus, the relative frequencies of instances that satisfy certain constraints on the values of their attributes represent sufficient information for decision tree algorithm, while the weight vector (expressed in terms of support vectors) that defines the maximal margin hyperplane represents sufficient information for SVM algorithm. Note that we are interested in characterizing the minimal information requirements of a learning algorithm, i.e. *minimal sufficient information* that needs to be extracted in order to determine the output of the learning algorithm.

We have formalized the treatment of distributed learning outlined above by introducing a family of learning, information extraction and information composition operators and establishing sufficient conditions for provably exact distributed and cumulative learning in terms of general algebraic properties of the operators (Caragea, Silvescu and Honavar, 2001). This theoretical framework provides a basis for a unified treatment of a diverse body of recent work related to: distributed learning approaches based on combining multiple models learned from disjoint data sets; parallel formulation of learning algorithms; techniques for scaling up distributed learning algorithms; algorithms based on distributed computation of sufficient information etc.

We plan to build up on our preliminary results mentioned above to design, implement, and analyze distributed and cumulative learning algorithms. New algorithms and software for distributed and cumulative learning will not only advance the state of the art in information technology, but also contribute to advances in emerging data-rich disciplines such as biological sciences where data available is huge, distributed and rapidly evolving.

References

Caragea, D., Silvescu, A., and Honavar, V., 2000. Agents that Learn from Distributed Dynamic Data Sources. In: *Proc. of the Workshop on Learning Agents, Agents 2000.*

Caragea, D., Silvescu, A., and Honavar, V., 2001. Invited Chapter. Towards a Theoretical Framework for Analysis and Synthesis of Agents That Learn from Distributed Dynamic Data Sources. In: *Emerging Neural Architectures Based on Neuroscience*. Berlin: Springer-Verlag.