Knowledge Discovery from Disparate Earth Data Sources

Doina Caragea and Vasant Honavar
Iowa State University

Abstract

Advances in data collection and data storage technologies have made it possible to acquire massive Earth science data sets. In principle, these data sets could be transformed into great scientific discoveries. However, due to the heterogeneous nature and to the scale of the available Earth science data, traditional analysis methods are challenged and much of these data remain largely unexplored. We have developed a general strategy for transforming machine learning algorithms for learning from a single data source into algorithms for learning from disparate, semantically heterogeneous data sources. We believe that our strategy could be adapted and used for data exploration and knowledge discovery from Earth science data.

1 Introduction

Large amounts of Earth data have been acquired and stored worldwide through remote sensing instruments, Earths satellites, terrestrial observations, and ecosystem monitoring technologies, among others [3]. Although largely unexplored, potentially, these data contain information that can be used to answer fundamental scientific questions (such as, how to understand and protect life on Earth, what is the connection between Sun and Earth, etc.), as well as practical questions (such as, how to predict and avoid hazards, how to improve the agricultural crop, how to predict the effects of the global change, etc.). Machine learning techniques [20, 13], in addition to traditional statistical techniques [9], offer some of the most cost-effective approaches to analyzing, exploring and extracting knowledge (features, correlations, and other complex relationships and hypotheses that describe potentially interesting regularities) from such data sources [16]. However, the applicability of current knowledge acquisition techniques to Earth science problems is challenged by the nature and the scale of the data available [3]. More precisely:

(a) Data repositories are large in size, dynamic, and physically distributed. Consequently, it is neither desirable nor feasible to gather all of the data in a centralized location. Hence, there is a need for efficient algorithms for analyzing and exploring multiple distributed data sources without transmitting large amounts of data.

(b) Autonomously developed and operated data sources often differ in their structure and organization (e.g., relational databases, flat files, etc.) and the operations that can be performed on the data sources (e.g., types of queries - relational queries, statistical queries, keyword matches). Hence, there is a need for theoretically well-founded strategies for efficiently obtaining the information needed for analysis within the operational constraints imposed by the data sources.

(c) Autonomously developed data sources are semantically heterogeneous. The ontological commitments associated with a data source (and hence its implied semantics) are typically determined by the data source designers, based on their understanding of the intended use of the data. Effective use of multiple sources of data in a given context requires reconciliation of semantic differences. Hence, there is a need for methods that can dynamically and efficiently extract and integrate information needed for knowledge acquisition, from a user's perspective (note that there has been significant efforts aimed at construction of ontologies, e.g., SWEET at sweet.jpl.nasa.gov).

There exist many Earth science data-rich problem domains, for which the available data is inherently distributed, autonomous and semantically heterogeneous, for example:

- **Global change**: Understanding and predicting global climate change and its effects require building predictive models that capture the relationships between measurable indicators (e.g., biosphere, hydrosphere, cryosphere, ocean, and other potentially related data) [24]. Relevant data are gathered independently by US, European, and Japanese geostationary satellite systems, and are available from individual data repositories (e.g., those included in NASA's Global Change Master Directory).

- **Terrestrial ecology**: Some examples of terrestrial ecology data include agriculture data (related to land use, crops, etc.), biosphere data (vegetation, etc.), hydrosphere data (related to water quality, ground water, etc.), ocean data (related to temperature, circulation, etc.), solid earth data (volcanos, rocks, minerals, etc.). Some of these data can be accessed through the web sites for the NASA's Global Change Master Directory, the Global Change Information System, and the NASA Earth Observing System Web Site. Usually, several types of data need to be used when addressing problems, such as: agriculture efficiency [27], terrestrial carbon sinks and effects [25], disaster prevention and management [5], etc.

Against these background, we have developed a general strategy for transforming a large class of traditional machine learning algorithms into algorithms for learning from distributed, autonomous, semantically heterogeneous data sources [8, 7]. We have illustrated our strategy using examples from biology [6]. We believe that the same strategy could be successfully adopted and used to explore and extract knowledge from Earth science data.

2 Strategy for Learning from Disparate Data Sources

Given a data set D, a hypothesis class H, and a performance criterion P, an algorithm L for learning (from centralized data D) outputs a hypothesis $h \in H$ that optimizes a performance criterion P [20]. In pattern classification applications, h is a classifier (e.g., a decision tree, a support vector machine classifier, etc.). The data set D typically consists of a set of training examples. Each training example is an ordered tuple of attribute values, where one of the attributes corresponds to a class label and the remaining attributes represent inputs to the classifier. The goal of learning is to produce a hypothesis that optimizes the performance criterion (e.g., minimizing classification error on the training data and the complexity of the hypothesis).

Traditional machine learning algorithms assume centralized access to data. Our general strategy for designing algorithms for learning from distributed data follows from the observation that most of the learning algorithms use only certain statistics computed from data

(e.g., mean value of an attribute, counts of instances that have specified values for some subset of attributes, the most frequent value of an attribute, etc.), in the process of generating a hypothesis. This observation yields a natural decomposition of a learning algorithm into two components (see Figure 1 (Left)): (a) an information gathering component that formulates and sends a statistical query to a data source; and (b) a hypothesis generation component that uses the resulting statistic to modify a partially constructed hypothesis (and further invokes the information gathering component as needed). In this model, the only

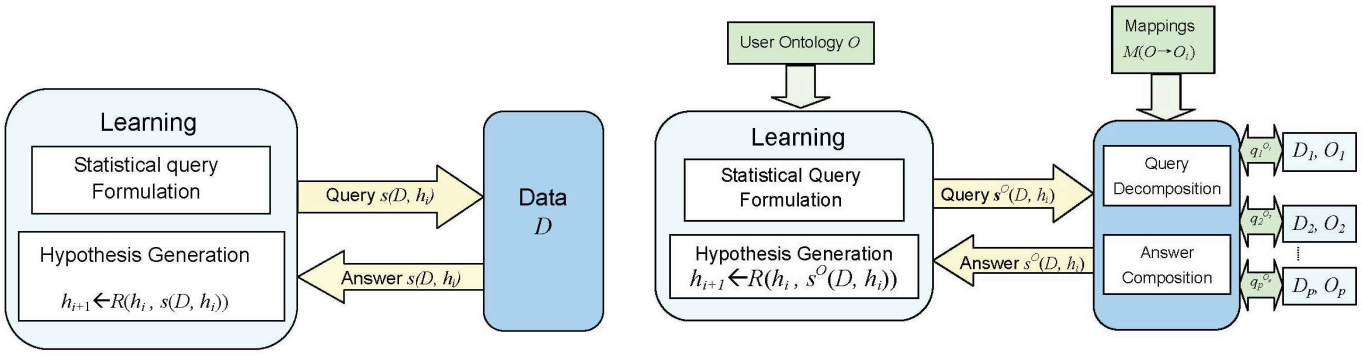

Figure 1: (Left) Learning = Statistical Query Answering + Hypothesis Generation (Right) General Framework for learning classifiers from semantically heterogeneous distributed data.

interaction of the learner with the data repository is through queries for the relevant statistics. Thus, in principle, the information gathering from distributed data $D_1, ..., D_p$ would entail decomposing each statistical query q posed by the learner into sub queries $q_1, ..., q_n$ that can be answered by the individual data sources $D_1, ..., D_p$, respectively, and combining the answers to the sub queries into an answer for the original query q. However, in general, the distributed data sources $D_1, ..., D_p$ are heterogeneous. To deal with the syntactic heterogeneity, we assume the existence of data source wrappers that can be used to answer statistical queries from various types of data sources [19, 21, 15].

To deal with the semantic heterogeneity, we explicitly associate an ontology with each data source. The ontology associated with a data source includes hierarchies corresponding to attribute value taxonomies (AVTs), for example a taxonomy of cyclone types. A user may want to explore the distributed data sources from his or her own perspective, corresponding to an ontology O_U. Mappings between AVTs in the user ontology and AVTs in the data source ontologies are specified through the means of semantic correspondences between similar attributes, for example, $WindSpeed(kmh) \equiv Speed(mph)$ or $Hurricane \subseteq TropicalStorm$. These mappings allow a user to view the distributed data as a single table structured according to his or her ontology. Therefore, they can be used to answer user queries that are expressed in terms of O_U. This allows us to extend our strategy for learning from data into a strategy for learning from distributed, semantically heterogeneous data (see Figure (Right)).

Note that, because the values of some data attributes can be organized in hierarchies (e.g., cyclone types) and data can be specified at different levels of abstraction in these hierarchies, the data of interest to a user may be partially specified. Zhang et al. [31, 32] developed ontology-aware algorithms for learning from partially specified data. Our approach

to learning from distributed, semantically heterogeneous data, can be seen as an extension of the approach to learning from partially specified data in the presence of AVTs [31, 32].

3 Summary and Discussion

We have outlined a strategy for learning classifiers from distributed, semantically heterogeneous data sources. Our strategy couples machine learning techniques with information integration techniques, making the process of knowledge acquisition from such sources transparent to the end user, as long as the implicit ontologies associated with the data are made explicit and mappings between a user ontology and data source ontologies are specified by domain experts (in principle, they could also be semi-automatically learned from data and validated by experts [11, 23]). More broadly, our research in the domain of knowledge acquisition from scientific data has led to the development of:

(a) A general theoretical framework for learning predictive models (e.g., classifiers) from large, physically distributed data sources [8].

(b) A theoretically sound approach to formulation and execution of statistical queries across semantically heterogeneous data sources [7].

(c) Statistically sound approaches to learning classifiers from *partially specified data* resulting from data described at different levels of abstraction [31].

(d) Tools to support collaborative development of modular ontologies [2].

(e) INDUS, a modular, extensible, open-source software toolkit[1] for data-driven knowledge acquisition from large, distributed, autonomous, semantically heterogeneous data sources [6].

Related work includes several approaches to distributed learning [22, 18, 26, 12] and information integration [17, 10, 14], in general, and in the Earth science domain [4, 28], in particular. However, to the best of our knowledge, none of these approaches combine data mining and information integration techniques into a system that can be easily used by end users to explore and extract knowledge from large, distributed, autonomous, semantically heterogeneous data sources.

Our algorithms and tools have been successfully applied to data-driven knowledge acquisition tasks that arise in bioinformatics [1, 6, 29, 30]. We believe that the same approach could be applied to data exploration and knowledge discovery problems in the Earth science domain, as a lot of tasks in this domain are similar in nature to those in bioinformatics and can be decomposed into information gathering and information processing components. For example, Zhang et al. [33] showed that the task of visualizing forest cover type data can be decomposed into these two components. The authors used the decomposition to design a strategy for visualizing data from large, distributed data repositories.

Research in the area of knowledge discovery from Earth science data is still in its infancy, posing many challenges due to the large amounts of data involved and the semantic heterogeneity nature of these data. Our contributions to the general problem of knowledge acquisition from distributed, semantically heterogeneous data sources represent important steps towards solutions to problems that arise in Earth science environments.

[1] http://www.cild.iastate.edu/software/indus.html

References

[1] C. Andorf, A. Silvescu, D. Dobbs, and V. Honavar. Learning classifiers for assigning protein sequences to gene ontology functional families. In *Fifth International Conference on Knowledge Based Computer Systems (KBCS 2004)*, India, 2004.

[2] J. Bao, D. Caragea, and V. Honavar. Towards collaborative environments for ontology construction and sharing. In *The 2006 International Symposium on Collaborative Technologies and Systems (CTS 2006)*, 2006. submitted.

[3] J. Behnk, E. Dobinson, S. Graves, T. Hinke, D. Nichols, P. Stolorz, and P. Newsome. Nasa workshop on issues in the application of data mining to scientific data, 1999. Final Report.

[4] T. L. Benyo and W. H. Jones. Exploring diverse data sets and developing new theories and ideas with project integration architecture. Technical Report NASA-TM-2005-213415, NASA, Glenn Research Center, Cleveland, Ohio, 2005.

[5] G. Brakenridge, E. Anderson, S. Nghiem, S. Caquard, and T. Shabaneh. Flood warnings, flood disaster assessments, and flood hazard reduction: the roles of orbital remote sensing. In *Symposium on Remote Sensing of the Environment*, Honolulu, HI, 2003.

[6] D. Caragea, J. Pathak, J. Bao, A. Silvescu, C. Andorf., D. Dobbs, and V. Honavar. Information integration and knowledge acquisition from semantically heterogeneous biological data sources. In *Proceedings of the 2nd International Workshop on Data Integration in Life Sciences (DILS 2005)*, volume 3615, pages 175–190, San Diego, CA, 2005. Berlin: Springer-Verlag.

[7] D. Caragea, J. Pathak, and V. Honavar. Learning classifiers from semantically heterogeneous data. In *Proceedings of the International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*, volume 3291, pages 963–980, Agia Napa, Cyprus, 2004. Springer-Verlag.

[8] D. Caragea, A. Silvescu, and V. Honavar. A framework for learning from distributed data using sufficient statistics and its application to learning decision trees. *International Journal of Hybrid Intelligent Systems*, 1(2):80–89, 2004.

[9] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury Press, Belmont, CA, 2001.

[10] S. Davidson, J. Crabtree, B. Brunk, J. Schug, V. Tannen, G. Overton, and C. Stoeckert. K2/Kleisli and GUS: experiments in integrated access to genomic data sources. *IBM Journal*, 40(2), 2001.

[11] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Y. Halevy. Learning to match ontologies on the semantic web. *VLDB*, 12(4), 2003.

[12] P. Domingos. Knowledge acquisition from examples via multiple models. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 98–106, Nashville, TN, 1997. Morgan Kaufmann.

[13] R. Duda, E. Hart, and D. Stork. *Pattern Recognition*. Wiley, 2000.

[14] B. Eckman. A practitioner's guide to data management and data integration in bioinformatics. *Bioinformatics*, pages 3–74, 2003.

[15] J.R. Gruser, L. Raschid, M. E. Vidal, and L. Bright. Wrapper generation for web accessible data sources. In *COOPIS '98: Proceedings of the 3rd IFCIS International Conference on Cooperative Information Systems*, pages 14–23, Washington, DC, USA, 1998. IEEE Computer Society.

[16] J. Han, R. B. Altman, V. Kumar, H. Mannila, and D. Pregibon. Emerging scientific applications in data mining. *Communications of ACM*, 45(8):54–58, 2002.

[17] R. Hull. Managing semantic heterogeneity in databases: A theoretical perspective. In *PODS*, pages 51–61, Tucson, Arizona, 1997.

[18] H. Kargupta, B.H. Park, D. Hershberger, and E. Johnson. Collective data mining: A new perspective toward distributed data mining. In H. Kargupta and P. Chan, editors, *Advances in Distributed and Parallel Knowledge Discovery*. MIT Press, 1999.

[19] L. Liu, C. Pu, W. Han, D. Buttler, and W. Tang. Building an extensible wrapper repository system: A metadata approach. In *Proceedings of IEEE Metadata Conference*, 1999.

[20] T.M. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[21] I. Muslea, S. Minton, and C. A. Knoblock. Hierarchical wrapper induction for semistructured information sources. *Autonomous Agents and Multi-Agent Systems*, 4(1/2):93–114, 2001.

[22] B. Park and H. Kargupta. Constructing simpler decision trees from ensemble models using Fourier analysis. In *Proceedings of the 7th Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'2002)*, pages 18–23, Madison, WI, June 2002. ACM SIGMOD.

[23] P.Mitra, N.F. Noy, and A.R. Jaiswal. Ontology mapping discovery with uncertainty. In *Fourth International Conference on the Semantic Web (ISWC-2005)*, Galway, Ireland, 2005.

[24] C. Potter. Predicting climate change effects on vegetation, soil thermal dynamics, and carbon cycling in ecosystems of interior alaska. *Ecological Modelling*, 175:1–24, 2004.

[25] C. Potter, S. Klooster, V. Genovese, and R. Myneni. Satellite data helps predict terrestrial carbon sinks. *Eos - Transactions of the American Geophysical Union*, 84:502–508, 2003.

[26] A. Srivastava, E. Han, V. Kumar, and V. Singh. Parallel formulations of decision-tree classification algorithms. *Data Mining and Knowledge Discovery*, 3(3):237–261, 1999.

[27] K. L. Wagstaff, D. Mazzoni, and S. Sain. Harvist: A system for agricultural and weather studies using advanced statistical models. In *Proceedings of the Earth-Sun Systems Technology Conference*, 2005.

[28] N. Wiegand and N Zhou. Ontology-based geospatial web query system. In P. Agouris and A. Croitoru, editors, *Next Generation Geospatial Information: From Digital Image Analysis to Spatio-Temporal Databases*, ISPRS Book series. Balkema, Taylor & Francis, 2005.

[29] C. Yan, D. Dobbs, and V. Honavar. A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics*, 20:i371–378, 2004.

[30] C. Yan, V. Honavar, and D. Dobbs. Identifying protein-protein interaction sites from surface residues - a support vector machine approach. *Neural Computing Applications*, 13:123–129, 2004.

[31] J. Zhang, D. Caragea, and V. Honavar. Learning ontology-aware classifiers. In *Proceedings of the Eight International Conference on Discovery Science (DS 2005), Springer-Verlag Lecture Notes in Computer Science*, volume 3735, pages 308–321, Singapore, 2005. Berlin: Springer-Verlag.

[32] J. Zhang, D.-K. Kang, A. Silvescu, and V. Honavar. Learning accurate and concise naive bayes classifiers from attribute value taxonomies and data. *Journal of Knowledge and Information Systems*, 2005.

[33] J. Zhang, L. Miller, D. Cook, A. Hardjasamudra, and H. Hofman. Densityplot matrix display for large distributed data. In *Proceedings of the Third International Workshop on Visual Data Mining, Third IEEE International Conference on Data Mining*, pages 59–70, Melbourne, FL, 2003.