IOWA STATE UNIVERSITY

Artificial Intelligence Research Laboratory
Bioinformatics and Computational Biology Program
Computational Intelligence, Learning, and Discovery Program
Department of Computer Science

NASA DM 2006

Knowledge Discovery from Disparate Earth Data Sources

Doina Caragea and Vasant Honavar

Motivation: Collaborative and Interdisciplinary e-Science

Available: large amounts of data in many application domains (e.g., global change and terrestrial ecology).

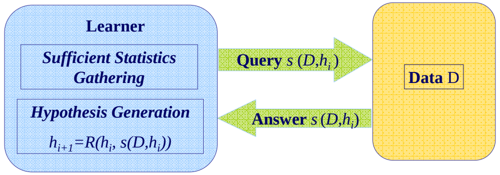
Opportunities: share data and findings between scientists working on related problems.

Challenges: large amounts of data; heterogeneous structure; different ontological commitments; constraints imposed by autonomous data sources.

Needed: knowledge discovery from large, autonomous, distributed and semantically heterogeneous data sources according to a user view.

Traditional Machine Learning Algorithms -- centralized access to data

Learning Classifiers from Data Revisited

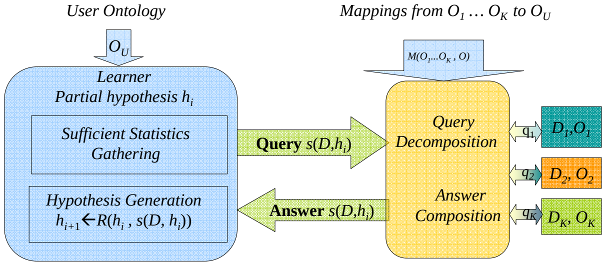


Sufficient Statistics

A statistic $s(D)$ is called a ***sufficient statistic*** for a parameter θ if $s(D)$ provides all the information needed for estimating the parameter θ from data D. We are interested in ***minimal sufficient statistics***.

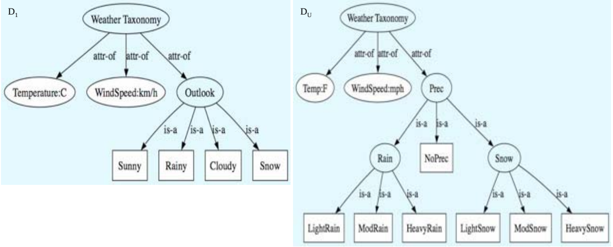
A statistic $s(D,h_i)$ is called a sufficient statistic for the ***refinement*** of a hypothesis h_i into h_{i+1} if there exists a refinement algorithm R that accepts h_i and $s(D,h_i)$ as inputs and outputs h_{i+1}.

Learning from Distributed, Semantically Heterogeneous Data



Ontologies

An ontology is a specification of objects, categories, properties and relationships used to conceptualize a domain of interest. Hierarchies (e.g., *isa* hierarchies) are a common type of ontologies. Hierarchies can be seen as orderings over a set of terms. Types of attributes that describe a data set can be defined as a hierarchical ontology.



Ontology-extended data sources

Let $A_1, A_2, \dots, A_n$ be the attributes of a data source and $\tau_1, \tau_2, \dots, \tau_n$ their types, We say that $\mathbf{D}=(D,S,O)$ is an *ontology-extended data source* if D is a data set, O is an ontology describing the content of the data D, $S=\{A_1{:}\,\tau_1, A_2{:}\,\tau_2, \dots, A_n{:}\,\tau_n\}$ is the data source schema and the following condition is satisfied: $D \subseteq \tau_1 \times \dots \times \tau_n$

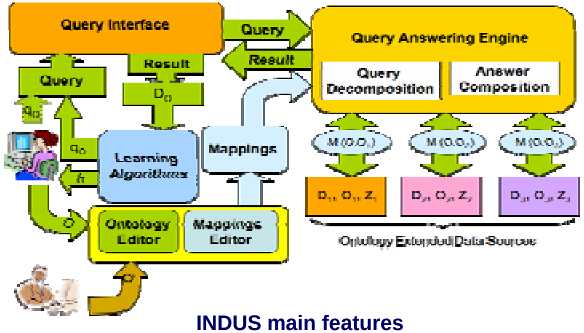
User view

A *user view* with respect to a set of ontology-extended data sources is given by a user schema and ontology and a set of semantic correspondences from data source meta-data to user meta-data.

Semantic correspondences

Schema level:	Ontology level:
Temperature : $D_1 \equiv$ Temp : D_U WindSpeed : $D_1 \equiv$ WindSpeed : D_U Outlook : $D_1 \equiv$ Prec : D_U	Rainy : $D_1 \equiv$ Rain : D_U Sunny : $D_1 \subseteq$ NoPrec : D_U Sunny & Cloudy : $D_1 \equiv$ NoPrec : D_U Rainy : $D_1 \supseteq$ LightRain : D_U Snow : $D_1 \supseteq$ Snow : D_U Etc.

INDUS: An Ontology-Based Approach to Information Integration and Knowledge Discovery from Distributed, Semantically Heterogeneous, Autonomous Data Sources



INDUS main features

- A clear distinction between data and the semantics of the data: makes it easy to define mappings from data source ontologies to user ontologies
- User-specified ontologies: each user can specify his or her ontology and mappings from data source ontologies to the user ontology; there is no single global ontology.
- A user-friendly ontology and mappings editor: this can be easily used to specify ontologies and mappings; however, a predefined set of ontologies and mappings are also available in a repository.
- Knowledge acquisition capabilities: machine learning algorithms can be easily linked to INDUS, making it an appropriate tool for information integration as well as knowledge acquisition tasks.

INDUS prototype: web address

http://www.cild.iastate.edu/software/indus.html

Acknowledgements: This work is supported in part by grants from the National Science Foundation (IIS 0219699), and the National Institutes of Health (GM 066387) to Vasant Honavar.