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 $\theta$ if $s(D)$ provides all the information needed for estimating the parameter $\theta$ from data $D$. We are interested in minimal sufficient statistics.

A statistic $s(D,h)$ 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 $s(D,h)$ as inputs and outputs $h_{i+1}$.

Sufficient Statistics

Learning from Distributed, Semantically Heterogeneous Data

User Ontology

Mappings from $O_1$ ... $O_K$ to $O_U$

Schema level:

Temperature: $D_1$ ≡ $T_{1}$, WindSpeed: $D_1$ ≡ $WS_{1}$, 
Outlook: $D_1$ ≡ $O_{1}$

User Ontology

Temperature: $D_{1}$, WindSpeed: $D_{1}$, 
Outlook: $D_{1}$

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 $A_1, A_2, ..., A_n$ be the attributes of a data source and $\tau_1, \tau_2, ..., \tau_n$ their types.

We say that $D=(D,S,O)$ is an ontology-extended data source if $D$ is a data set, $O$ is an ontology describing the content of the data $D$, $S=\{A_1, A_2, ..., A_n\}$ is the data source schema and the following condition is satisfied: $D \subseteq \tau_1 \times \cdots \times \tau_n$

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

<table>
<thead>
<tr>
<th>Schema level:</th>
<th>Ontology level:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature: $D_1$ ≡ $T_{1}$</td>
<td>Rainy: $D_1$ ≡ Rain: $D_1$</td>
</tr>
<tr>
<td>WindSpeed: $D_1$ ≡ $WS_{1}$</td>
<td>Sunny: $D_1$ ≡ NoPrec: $D_1$</td>
</tr>
<tr>
<td>Outlook: $D_1$ ≡ $O_{1}$</td>
<td>Sunny &amp; Cloudy: $D_1$ ≡ NoPrec: $D_1$</td>
</tr>
<tr>
<td></td>
<td>Rainy: $D_1$ ≡ LightRain: $D_1$</td>
</tr>
<tr>
<td></td>
<td>Snow: $D_1$ ≡ Snow: $D_1$</td>
</tr>
</tbody>
</table>

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 define ontologies and mappings 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.