

# AirIDM: A Collection of Data Source Independent Learning Algorithms Through the Means of Sufficient Statistics and Data Source Wrappers

Doina Caragea, Adrian Silvescu, and Vasant Honavar  
Artificial Intelligence Research Laboratory  
Department of Computer Science  
Iowa State University, Ames, IA 50011  
{dcaragea|silvescu|honavar}@cs.iastate.edu

AirIDM is a collection of machine learning algorithms, which are data source independent through the means of sufficient statistics and data source wrappers. They work with general data sources where data can be stored in any format as long as wrappers for accessing and getting sufficient statistics from those data sources are provided. Some of the algorithms in AirIDM are adapted from Weka implementations by separating the information extraction and hypothesis generation components.

Figure 1 shows the general architecture of AirIDM. As can be seen a learning algorithm is regarded as a `TRAINER` that generates a `HYPOTHESIS` from `SUFFICIENT STATISTICS`. Each `DATA SOURCE` is wrapped by a `DATA SOURCE WRAPPER`. The `TRAINER` registers sufficient statistics with the `DATA SOURCE WRAPPER` which populates them by accessing the corresponding `DATA SOURCE`. Once the `SUFFICIENT STATISTICS` are populated, they are used by the `TRAINER` to get `parameters` that are needed to build a current `HYPOTHESIS`. This process may repeat a few time (e.g., for decision tree algorithm). When a hypothesis is built, a `USER` can get `hypothesis` from the `TRAINER` and use it to classify new unseen data.

AirIDM is an open source software issued under the GNU General Public License. In the current release, we provide wrappers for data that can be seen as a single table (INDUS wrapper, Weka wrapper), as a collection of tables (multi relational data wrapper) or as a sequence (sequence wrapper). We have implemented sufficient statistics of type joint counts, which are the sufficient statistics needed by a large class of algorithms (e.g., Naive Bayes Bayes Networks Relational Learning Decision Trees with a variety of splitting criteria). The algorithms currently implemented are Naive Bayes and Decision Tree Algorithm.

Because of the modular design of AirIDM and the clear separation of concerns between hypothesis generation and information extraction, AirIDM can be easily linked

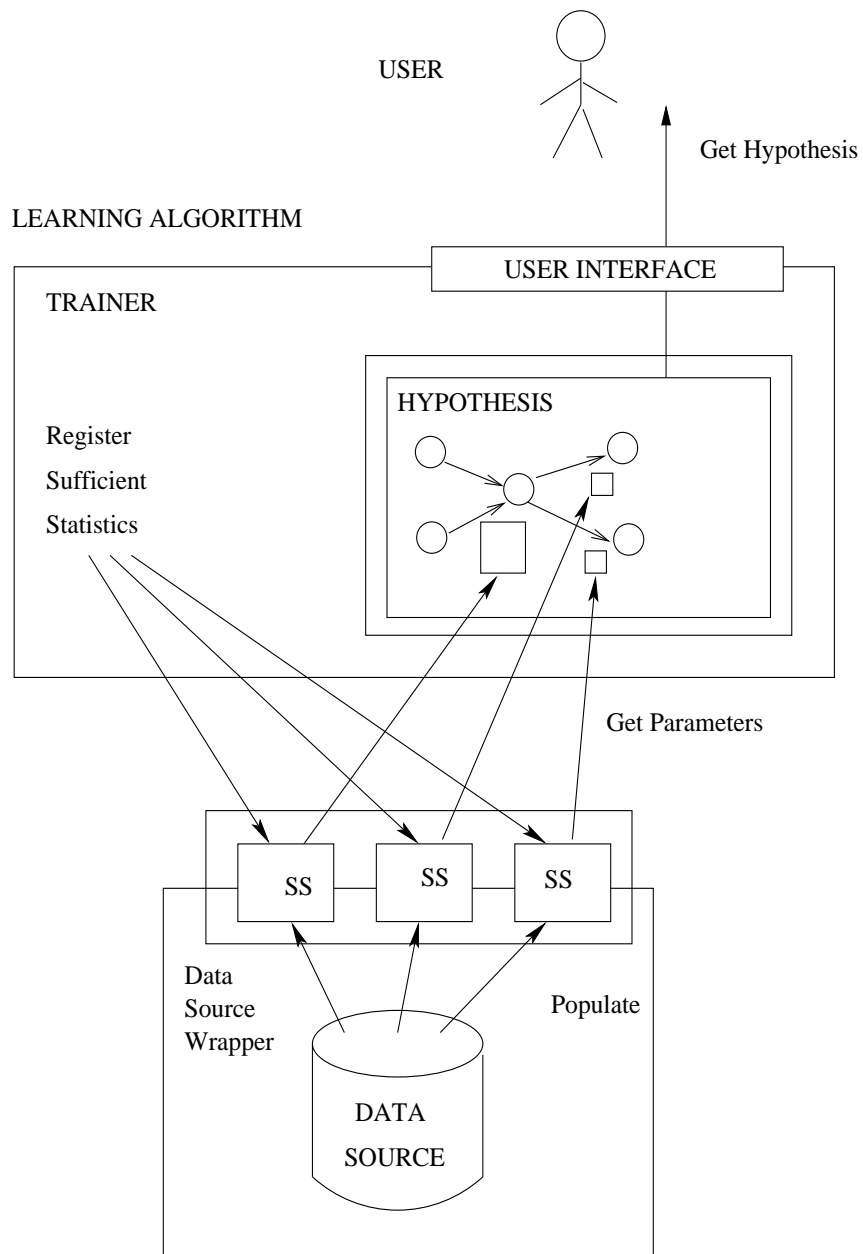


Figure 1: AirIDM: Learning algorithms are independent of data sources through the means of sufficient statistics and wrappers.

to INDUS to obtain a system for learning from heterogeneous distributed autonomous data sources. Thus, we have written an INDUS wrapper that provides sufficient statistics to the trainer and linked it with the AirlDM implementations of Naive Bayes and Decision Tree algorithms. Using that, we have implemented algorithms for learning Naive Bayes and Decision Tree classifiers from horizontally and vertically distributed data sources by having the query answering engine register the sufficient statistics that it gets from the trainer with the corresponding wrappers and composing the statistics populated by the wrappers into the sufficient statistics needed by the trainer. Thus, in the case of horizontally distributed data, each count statistic is registered with the wrapper of each distributed data source and the answers are added up to get the overall count. In the case of vertically fragmented data, the query answering engine identifies the wrapper that can be used to populate each sufficient statistic count and the answer is sent back to the trainer.

Therefore, we can achieve learning from distributed data in a way which is transparent to the learning algorithm, meaning that from the algorithm point of view it makes no difference if the data comes from a single or multiple data sources or if these data sources are represented as relational tables or flat file or any other format. Furthermore, if the distributed data sources are heterogeneous, the query answering engine can perform mappings from data sources ontologies to user ontology and the algorithms remain unchanged.