

## **High Performance Knowledge Discovery and Data Mining Systems Using Workstation Clusters**

William H. Hsu, Kansas State University / NCSA

Michael Welge, Thomas Redman, Loretta Auvil and David Tcheng, NCSA

Poster URLs:

<http://lightsaber.ncsa.uiuc.edu/NCSA/D2KIntro.exe>

<http://lightsaber.ncsa.uiuc.edu/NCSA/HP-KDD.ppt>

- Abstract

We present D2K, a Java-based visual programming system for information visualization and data mining. This poster focuses on the problem of scaling up high-level optimization algorithms (wrappers) for performance enhancement in machine learning. In particular, we report results from a small Beowulf cluster, and similar network-of-workstation systems, that incorporate the MLC++ machine learning library into a wrapper for performance enhancement in D2K. The wrapper performs relevance determination (finding attributes that are relevant to a prediction objective) using a parallel, distributed, genetic algorithm running on this cluster. It achieves a linear speedup due to a high degree of task parallelism as well as improved model efficiency in the number of relevance attributes discovered, compared to the search algorithm A\* using an existing heuristic.