A Density-Biased Sampling Technique to Improve Cluster Representativeness

Ana Paula Appel
Adriano A. Paterlini
Elaine P. M. de Sousa
Caetano Traina Jr.

University of São Paulo at São Carlos – Brazil - USP
Computer Science Department - ICMC

Poster Highlight

- Data mining algorithms $\longrightarrow$ are computationally very expensive.
- **Sampling** $\longrightarrow$ Reduces complexity and Summaries are obtained faster;
- Clusters have different sizes $\longrightarrow$ problems with uniform sampling;
- *How to sample a multi-dimensional dataset without missing the clusters, even if they are* **unbalanced** *and without* **any previous knowledge** *about the clusters?*
- **BBS - Biased Box Sampling** is a new technique that selects points in such a way to preserve the representativeness of each cluster.
- It is **robust to withstand high-dimensionality** drawbacks; **insensitive to noise** and **linear** on both number of points N and on the number E of attributes.