Learning from Categorical and Numerical Imbalanced Data

Hao Canh Nguyen (450026)

School of Knowledge Science,
Japan Advanced Institute of Science and Technology

February 10, 2006

Keywords: Imbalanced Data, Rule Learning, Sampling, ROC analysis, Manifold Learning.

Imbalanced data learning has recently begun to receive considerable attention from the research and industrial communities. Imbalanced data is problematic as traditional machine learning methods fail to achieve satisfactory results due to the skewed class distribution. Solutions to the problem generally use traditional machine learners to make a bias decision in favor of the smaller class. To make a bias decision, one need to have a good assumption of some kind of data distribution. The thesis proposes two methods to learn imbalanced data problem, one is a rule learner for categorical data using local data distributions and another is a family of sampling algorithms for numerical data using manifold modeling.

For categorical data, we deal with imbalanced data problem using example weighting to make a bias decision. Higher weights are assigned to small class examples to avoid being overshadowed by the large class ones. In this work, we introduce a scheme to weight examples of small class based solely on local data distributions. The approach is for categorical data, and a rule learning algorithm is constructed taking the weighting scheme into account. The approach proves favorable performance to other rule learning systems. We conclude that local data distributions contain information that would be useful for the imbalanced data problem.

For numerical data, we explicitly model the distribution of small class data to make a bias decision. We utilize the flexibility of manifold modeling for the small class data distribution. Based on recent advances in manifold learning algorithms, we design basic sampling strategies to account for skewed class distribution by generating synthetic small class data. We combine these strategies to create a family of three sampling algorithms. Experimental evaluation shows that the proposed algorithms can learn effectively imbalanced data sets. We conclude that manifold is flexible and useful enough to account for the imbalanced data problem.