

Student Research Symposium 2003

XiaoJin (Jerry) Zhu

Semi-Supervised Learning

Many natural language applications, like speech recognition, information retrieval, machine translation etc., employ classification with statistical machine learning methods. To perform classification well one needs large amount of labeled data, which is often hard to obtain. On the other hand unlabeled data may be relatively easy to collect, but traditionally it was ignored for classification. It is of great interest to find ways to use both labeled and unlabeled data.

I propose an approach based on a Gaussian random field model to learn from both labeled and unlabeled data. Labeled and unlabeled data are represented as vertices in a weighted graph, with edge weights encoding the similarity between instances. The learning problem is then formulated in terms of a Gaussian random field on this graph, where the mean of the field is characterized in terms of harmonic functions, and is efficiently obtained using matrix methods or belief propagation. The resulting learning algorithms have intimate connections with random walks, electric networks, and spectral graph theory. Promising experimental results are presented for synthetic data, digit classification, and text classification tasks.