Discriminative Similarity for Data Clustering

Yingzhen Yang[1], Ping Li[2]
[1] Arizona State Uniersity
[2] Cognitive Computing Lab, Baidu Research

Introduction

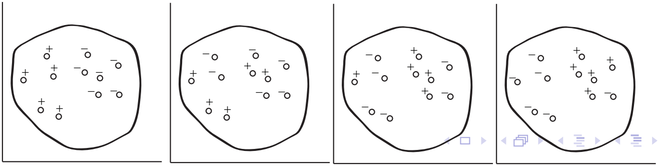
- Similarity-based clustering highly depends on the similarity measure.
- Examples of similarity measures:
  - Kernel similarity: similarity induced by kernel function
  - Sparse similarity: similarity induced by sparse representation, especially effective for data lying on subspaces or manifolds
  - Other similarity measures, such as Random Forest-based similarity

Introduction

- However, most existing similarity measures do not account for cluster labels for the purpose of separating different clusters.
    - It is worthwhile to emphasize that the similarity is used for separating different clusters.

- How to consider cluster labels in the similarity measure?

- Our solution: train a similarity-based classifier based on a candidate cluster labeling.
    - A candidate cluster labeling is defined as a hypothetical labeling.



Introduction

- Each hypothetical labeling corresponds to a candidate data partition, and clustering methods search for the hypothetical labeling which is optimal in some sense.
    - Each hypothetical labeling is associated with a similarity-based classifier trained on such hypothetical labeling. The optimal hypothetical labeling corresponds to the similarity-based classifier with minimum generalization error

- In this way, clustering is formulated as a multi-class classification problem.

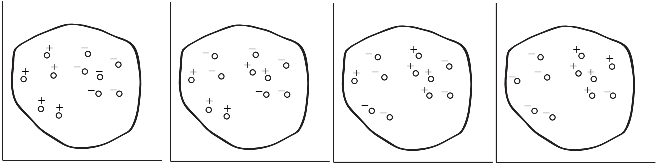


Figure 1: Illustration of binary hypothetical labelings

Our Contribution

- We present a new similarity measure for data clustering, and the similarity measure is induced by the generalization error of a similarity-based classifier trained on a specific hypothetical cluster labeling.

- Our theoretical contribution: deriving the generalization bound for general similarity-based classifier (Theorem 4.2 in the paper)
    - This bound is the first principled result about generalization error bound for general similarity-based classifier with strong connection to the established generalization error bound for Support Vector Machines (SVMs) or Kernel Machines.
    - Please refer to more details in Section B.1 of the paper.

Our Contribution

- We use kernel as the similarity function in the similarity-based classifier.
- By minimizing the error bound for the similarity-based classifier, we propose Clustering by Discriminative Similarity via unsupervised Kernel classification (CDSK).
- Please refer to Section 5 of the paper for the details of CDSK and its experimental results.

Thank you!