THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Author Name(s)

[Report]
Report Title

Abstract

Write your abstract here.

1 Introduction

The importance of dataspace systems has already been recognized and emphasized in handling heterogeneous data In many applications of big data, there is usually a large gap between the number of labeled examples needed to obtain high prediction accuracy, and the number of labeled examples that could be realistically obtained. These problems can be found in social networks. In the meanwhile, however, usually a large amount of labeled data sets can be collected from related applications of various types, such as, labeled documents in social tagging systems (e.g., wikipedia, ODP) for web mining and classified images from social sites such as Flickr. This talk will show examples about how free labeled heterogeneous source data provide useful supervision to a related target task.

Three challenging sub-issues need to be solved:

1. The source data may be generated from a different feature space from the target data (e.g., source is text data while target is image data).
2. The source data may be drawn from a distribution different from the target data. For example, the source data is dominated by a Gaussian distribution while the target data is dominated by a multinomial distribution, which violates the i.i.d. assumption.
3. The source and target data may have totally different output spaces.

References

[1] R. Caruana, Multitask learning. Machine Learning 28(1) 41-75, 1997.

[2] S. J. Pan and Q. Yang, A Survey on Transfer Learning. Technical Report HKUST-CS08-08, Hong Kong University of Science and Technology, 2009.

[3] X. Shi, Q. Liu, W. Fan, Q. Yang, and P. S. Yu, Predictive Modeling with Heterogeneous Sources. In Proc. of SDM10, 2010.