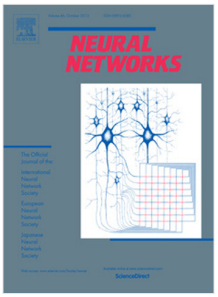
Accepted Manuscript



Effects of additional data on Bayesian clustering

Keisuke Yamazaki

PII: S0893-6080(17)30151-X
DOI: http://dx.doi.org/10.1016/j.neunet.2017.06.015
Reference: NN 3780

To appear in: *Neural Networks*

Received date : 14 July 2016
Revised date : 26 June 2017
Accepted date : 30 June 2017

Please cite this article as: Yamazaki, K., Effects of additional data on Bayesian clustering. *Neural Networks* (2017), http://dx.doi.org/10.1016/j.neunet.2017.06.015

Effects of Additional Data on Bayesian Clustering

Keisuke Yamazaki
k.yamazaki@aist.go.jp
Artificial Intelligence Research Center,
National Institute of Advanced Industrial Science and Technology
2-3-26 Aomi Koto-ku, Tokyo, Japan

Abstract

Hierarchical probabilistic models, such as mixture models, are used for cluster analysis. These models have two types of variables: observable and latent. In cluster analysis, the latent variable is estimated, and it is expected that additional information will improve the accuracy of the estimation of the latent variable. Many proposed learning methods are able to use additional data; these include semi-supervised learning and transfer learning. However, from a statistical point of view, a complex probabilistic model that encompasses both the initial and additional data might be less accurate due to having a higher-dimensional parameter. The present paper presents a theoretical analysis of the accuracy of such a model and clarifies which factor has the greatest effect on its accuracy, the advantages of obtaining additional data, and the disadvantages of increasing the complexity.
Keywords: unsupervised learning, semi-supervised learning, hierarchical parametric models, latent variable estimation

1 Introduction

Hierarchical probabilistic models, such as mixture models, are often used for data analysis. These models have two types of variables: observable and latent. Observable variables represent the data that can be observed,