
Title: Removing confounding factors via constraint-based clustering: An application to finding homogeneous groups of multiple sclerosis patients

Author: Jingjing Liu Carla E. Brodley Brian C. Healy Tanuja Chitnis

PII: S0933-3657(15)00083-4
DOI: http://dx.doi.org/doi:10.1016/j.artmed.2015.06.004
Reference: ARTMED 1412

To appear in: *ARTMED*

Received date: 5-4-2015
Revised date: 6-6-2015
Accepted date: 26-6-2015

Please cite this article as: Jingjing Liu, Carla E. Brodley, Brian C. Healy, Tanuja Chitnis, Removing confounding factors via constraint-based clustering: An application to finding homogeneous groups of multiple sclerosis patients, *Artificial Intelligence In Medicine* (2015), http://dx.doi.org/10.1016/j.artmed.2015.06.004

# Removing confounding factors via constraint-based clustering: An application to finding homogeneous groups of multiple sclerosis patients

Jingjing Liu

*Department of Computer Science, Tufts University, 161 College Ave, Medford, MA, USA, 02155. jliu02@tufts.edu.*

Corresponding Author: Carla E. Brodley

*College of Computer and Information Science, Northeastern University, 440 Huntington Avenue 202 West Village H Boston, Massachusetts 02115. c.brodley@neu.edu. Phone: (617) 373-5204*

Brian C. Healy

*Biostatistics Center, Massachusetts General Hospital, Boston, MA, USA, 02114. bchealy@partners.org.*

Tanuja Chitnis

*Partners Multiple Sclerosis Center, Brigham and Women's Hospital, Brookline, MA, USA, 02115. tchitnis@rics.bwh.harvard.edu.*

**Abstract**

*Objectives:* Confounding factors in unsupervised data can lead to undesirable clustering results. For example in medical datasets, age is often a confounding factor in tests designed to judge the severity of a patient's disease through measures of mobility, eyesight and hearing. In such cases, removing age from each instance will not remove its effect from the data as other features will be correlated with age. Motivated by the need to find homogeneous groups of multiple sclerosis (MS) patients, we apply our approach to remove physician subjectivity from patient data.

*Methods:* We present a method based on constraint-based clustering to remove the impact of such confounding factors. Given knowledge about which feature (or set of features) is a confounding factor, call it $F$. Our method