DATA FUSION: A FIRST STEP IN DECISION INFORMATICS

By

Jiaqi Hu

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: Decision Sciences and Engineering Systems

The original of the complete thesis is on file
In the Rensselaer Polytechnic Institute Library

Examining Committee:

James M. Tien, Thesis Co-Advisor

Wai Kin (Victor) Chan, Thesis Co-Advisor

Daniel Berg, Member

Yingrui Yang, Member

John E. Mitchell, Member

Rensselaer Polytechnic Institute
Troy, New York

August 2008
(For Graduation December 2008)

ABSTRACT

This research proposes to develop data fusion tools that allow for the simultaneous utilization of qualitative and quantitative data in clustering analysis. From the data fusion perspective, the integration of multiple data sources can happen at different levels including data level, feature level, similarity level and decision level. We categorize methods that have been employed in the simultaneous utilization of qualitative and quantitative data, based on the fusion levels. We highlight two critical research areas where multiple data sources are to be fused at the feature and similarity level, respectively, which implies less information loss and potential flexible integration scheme. For the feature level fusion, we extend a probabilistic model for the mixed type data modeling to model the dependency between the qualitative and quantitative data, and embed the feature identification in the model estimation procedure. We also propose a model initialization strategy to reduce the influence of the initial configuration on the model estimation. We formulate the model estimation in an optimization framework where the penalized log likelihood is maximized. For the similarity level fusion, we propose a sub-sampling based method to search for the weight configuration in the weight sum rule. We also propose a voting based threshold strategy for noise reduction when the max rule is applied. We show through empirical studies that fusing quantitative and qualitative data can produce better results in clustering analysis than using individual data sources alone.