International Journal of Innovative Computing, Information and Control
Volume **6**, Number **3(B)**, March **2010**

ICIC International ©2010 ISSN 1349-4198
pp. 1461–1470

DISTRIBUTED LATENT DIRICHLET ALLOCATION FOR DISTRIBUTED CLUSTER ENSEMBLE

Hongjun Wang, Jianhuai Qi, Weifan Zheng and Mingwen Wang

Information Research Institute
Southwest Jiaotong University of China
Chengdu 610054, P. R. China
wanghongjun@home.swjtu.edu.cn

Received December 2008; revised April 2009

Abstract. *Cluster ensemble is becoming an important research pot and many researchers study in this field. But there is no author to state the problem of distributed cluster ensemble. In this paper the authors initiatively state the problem and introduce the model of distributed latent Dirichlet allocation (D-LDA) for distributed cluster ensemble which is the most important contribution of this paper. First, the latent variables in D-LDA and some terminologies are defined for distributed cluster ensemble. Second, Markov chain Monte Carlo(MCMC) approximation inference for D-LDA is stated in detail. Third, some datasets from UCI are chosen for experiments. Compared with cluster-based similarity partitioning algorithm (CSPA), hyper-graph partitioning algorithm(HGPA) and meta-clustering algorithm(MCLA), the results show D-LDA does work better, furthermore the outputs of D-LDA, as a soft cluster model, can not only cluster the data points but also show the structure of data points.*

Keywords: Distributed cluster ensemble, Distributed latent Dirichlet allocation, Privacy preservation

1. Introduction.

Cluster ensemble is becoming an important research pot which is helpful to handle the problems of privacy preservation, distributed computing and knowledge reuse. And cluster ensemble is a hard problem, the major difficulty of which lies in finding a consensus partition from the outputs of various base clusterings algorithms. There is no explicit correspondence between the labels delivered by different base clusterings. Furthermore, the complexity arises when different base clusterings deliver different numbers of clusters, often resulting in an intractable label correspondence problem. The combination of multiple clustering can also be viewed as finding a median partition with respect to the given partitions which is proven to be NP-complete [1, 2]. Many researchers apply themselves to this field and their goal is to find a robust and stable algorithm for it or focus on finding better partitions of cluster ensemble. Xiaoli's [3] ensemble selection method is based on quality and diversity and his goal is to select a subset of solutions to form a smaller but better performing cluster ensemble than using all available solutions. Cluster ensemble is to combine multi-partitions into an optimal one without access to the original dataset. It can achieve more than a single clustering in several aspects which are also the motivations of cluster ensemble.

Robustness and stability: Robustness and stability is one of the requirements for clustering algorithm. Cluster ensemble combines the results of single clusterings into an optimal one. So it will run better than a single clustering in theory. The basic motivation of cluster ensemble is that cluster ensemble commonly runs better average performance than the best of single clustering and is with lower sensitivity to noise [1, 4]. If cluster