IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 9, 2019, accepted January 26, 2019. Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2019.2898644

A Novel Parallel Biclustering Approach and Its Application to Identify and Segment Highly Profitable Telecom Customers

QIN LIN[1], HUAILING ZHANG[1], XIZHAO WANG[2], (Fellow, IEEE), YUN XUE[3], HONGXIN LIU[1], AND CHANGWEI GONG[1]

[1]School of Information Engineering, Guangdong Medical University, Dongguan, China
[2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
AQ:1 [3]School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou, China

Corresponding author: Xizhao Wang (xizhaowang@ieee.org)

AQ:2 This work was supported in part by the National Natural Science Foundation of China under Grant 81871433 and Grant 71371063, and in part by the Basic Research Project of Knowledge Innovation Program in Shenzhen under Grant JCYJ20150324140036825.

AQ:3 **ABSTRACT** Identifying and segmenting various kinds of highly profitable customers is a critical issue for telecom enterprises. However, the continual increase in the dimension and the volume of data makes traditional approaches inefficient and even unfeasible. To overcome these problems, a novel statistically motivated parallel large sum submatrix biclustering algorithm based on Spark MapReduce (SP-PLSS) is proposed in this paper. Different from traditional approaches, the SP-PLSS is driven by a newly proposed bicluster model, and clusters both customer samples and consumer attributes simultaneously so that it could finely identify and segment the highly profitable customers who share similarly upscale purchasing behavior on a small fraction of attributes. Furthermore, with the implementation of the MapReduce framework on a Spark platform, the SP-PLSS significantly improves the efficiency and scalability of handling the large dataset. The extensive experiments on a real-world telecom consumption data and synthetic large datasets show that, in comparison with other competing algorithms, the SP-PLSS could provide operators with a comparatively advanced, scalable, and feasible solution in identifying and segmenting highly profitable telecom customers with superior clustering results.

INDEX TERMS Biclustering, clustering effectiveness evaluation spark, MapReduce, market segmentation parallel computing, cloud computing.

I. INTRODUCTION

With the advancement of communication technologies, telecom industry has witnessed the coming of the big data era in recent years. According to the statistics of the International Telecommunication Union, there are more than 7 billion mobile cellular subscriptions by the end of 2015, corresponding to a penetration rate of 97% up from 738 million in 2000 [1]. Due to the limitation of resources and manpower, there is an urgent need for telecom companies to identify and segment various homogeneous subgroups of highly profitable customers who contribute most of the enterprises' revenue [2]. The most obvious benefit is that it can allow operators to deploy resources more effective according to different subgroups' characteristics and then further offer personalized and differentiated services to maintain the good relationship with them. Nevertheless, nowadays conventional approaches of identifying and segmenting highly profitable customers are faced with two main challenges in the context of telecom big data. The first challenge comes from the critical limitation of traditional independent row-column clustering (IRCC) methods and existing biclustering algorithms, especially when dealing with highdimensional consumer records. So far, k-means algorithm [3]–[6] self-organizing maps (SOM) [7]–[12], fuzzy cmeans (FCM) algorithm [13]–[16] and other IRCC methods have been widely considered in market segmentation. Among them, kmeans algorithm is the most commonly used technique. For example, Liu *et al.* [6] put forward a systematically integrated big data mining approach based on k-means to find out high value customers. Another related method SOM that can project high-dimensional input space onto a low-dimensional topology has been recently applied to market segmentation. For instance Yao *et al.* [10] proposed a SOM-Ward clustering algorithm to segment the

The associate editor coordinating the review of this manuscript and approving it for publication was Wanqing Wu.