Journal of Database Management
Volume 30 • Issue 2 • April-June 2019

Multi-Objective Optimization-Based Networked Multi-Label Active Learning

Lei Li, Hefei University of Technology, Hefei, China

Yuqi Chu, Luoyang Optoelectro Technology Development Center, Luoyang, China

Guanfeng Liu, Macquarie University, NSW, Australia

https://orcid.org/0000-0001-8980-4950

Xindong Wu, Mininglamp Academy of Sciences, Mininglamp Technologies, Beijing, China

ABSTRACT

Along with the fast development of network applications, network research has attracted more and more attention, where one of the most important research directions is networked multi-label classification. Based on it, unknown labels of nodes can be inferred by known labels of nodes in the neighborhood. As both the scale and complexity of networks are increasing, the problems of previously neglected system overhead are turning more and more seriously. In this article, a novel multi-objective optimization-based networked multi-label seed node selection algorithm (named as MOSS) is proposed to improve both the prediction accuracy for unknown labels of nodes from labels of seed nodes during classification and the system overhead for mining the labels of seed nodes with third parties before classification. Compared with other algorithms on several real networked data sets, MOSS algorithm not only greatly reduces the system overhead before classification but also improves the prediction accuracy during classification.

KEYWORDS

Active Learning, Multi-Label Classification, Multi-Objective Optimization, Networked Data

INTRODUCTION

With the fast development of network applications (Batini & Rula, 2015; Cao et al., 2016; Long & Siau, 2007), network research has attracted more attention from both academic researchers and industrial engineers (Bhagat, Cormode, & Muthukrishnan, 2011; Bu et al., 2018; Li, et al., 2016), where one of the important research directions is networked multi-label classification (Wang & Sukthankar, 2013; Wu, Zhao, & Li, 2016; Zhang, et al., 2010). Specifically, unknown labels of nodes can be inferred by the known labels of other nodes in the neighborhood, and these inferred labels can be further used in user classification, community detection, or personalized recommendation (Guo, et al., 2017; Hong, et al., 2015; Li, et al., 2015). Networked data, which is different from traditional data with simple structures (Bhagat, Cormode, & Muthukrishnan, 2011), can reflect the complex relations, such as friendship or colleagues in life, or co-authorship of an article (Guo, et al., 2017; Miller, Perlman, & Brehm, 2007), between nodes in network environments (Fakhraei, et al., 2015; Otte & Rousseau, 2002), which makes it difficult to classify labels in networks (McDowell & Aha, 2016).

DOI: 10.4018/JDM.2019040101

Mathematically, the multi-labeled network can be represented as a graph G=(W, E, C, Y), where W={ w_1 , w_2 ,..., w_n } is a set of nodes, E is a set of edges that connect pairs of nodes, C={ c_1 , c_2 ,..., c_k } represents labels with K classes, and $Y_i =(y_i^1, y_i^2, \cdots, y_i^K)\in \{0,1\}^K$ denotes the multi-labels that are associated with node w_i (if w_i belongs to label c_k , y_i^k = 1, otherwise y_i^k = 0). W is divided into two disjoint parts: S, i.e., nodes whose labels are known which are named as seed nodes, and S = nseed and U, i.e., nodes whose labels need to be classified. The networked multi-label classification problem is to use S to infer the labels for nodes in U. Our seed node selection is to actively learn the set S to satisfy some objectives under certain conditions.

In the literature, there are some existing algorithms to find out how to classify networked multi-labels, including MLKNN (Wang & Sukthankar, 2013), RPC (Zhang & Zhou, 2005), wvRN (Hüllermeier, et al., 2008), SCRN (Macskassy & Provost, 2003), and so on. Therein, the former two algorithms MLKNN and RPC belong to the traditional networked multi-label classification, where firstly the multi-labels are dispatched into single labels, and then the single label classification can be handled with existing networked single label classification approaches, finally the predicted single labels can be combined into the results of networked multi-label classification. The latter two algorithms wvRN and SCRN adopt relational classifications to solve the networked multi-label classification, which comprehensively utilizes the topological structure of the network to predict the unknown labels of nodes by network propagation. In network environments, we tend to adopt the latter approaches for the multi-label classification, as they have properly utilized the topological structure of the network (Liu, Wang, & Orgun, 2010; Liu, Wang, & Orgun, 2013; Liu, et al., 2018). In most of existing approaches, little attention has been paid to the seed node selection (Li, et al., 2017; Wu, Zhao, & Li, 2016). However, different seed nodes lead to different accuracy of networked multi-label classification, not to mention system overhead for obtaining the labels of seed nodes before classification. In particular, as the scale of networks (Yu, et al., 2014; Yu, et al., 2016), especially social networks, are expanding significantly (Fakhraei, et al., 2015; Guo, et al., 2017; Jiang, et al., 2016), and the complexity of data relations is increasing constantly, the overhead consumed by obtaining the information of seed nodes by third parties before classification, for example, the time consumption and system memory occupation, is rising constantly (Hu et al., 2019; Raza et al., 2018). In order to control the system overhead and classify the networked multi-labels, in this paper we propose a novel Multi-objective Optimization-based networked multi-label Seed node Selection algorithm (named as MOSS) to improve both the prediction accuracy of unknown labels of nodes during classification (Yan & Guo, 2016) and the system overhead for mining the labels of seed nodes by third parties before classification. Compared with other networked multi-label classification algorithms, with network structure oriented multi-objective optimization process, MOSS algorithm not only greatly reduces the system overhead before classification but also improves the prediction accuracy during classification.

Technically, this paper introduces a multi-objective optimization algorithm to optimally select seed nodes from network environments with active learning (Chen et al., 2019), endowing seed node selection with both network structure and functional characteristics such as improving the prediction accuracy during classification, and reducing the time consumption and the system memory occupation for mining the labels of seed nodes before classification. More specifically, with the network structure-oriented decision process including crossover, mutation and population evolution from genetic algorithms, the selected nodes which satisfy the targeted requirements are taken as seed nodes, which are useful for the prediction of the unknown labels of nodes.

This paper is organized as follows. Firstly, we describe the basic concepts and related definitions about the multi-objective optimization approach NSGA-II and summarize recent research results about networked multi-label classification. Secondly, it introduces main ideas and steps of networked multi-label classification in our proposed MOSS algorithm, more specifically the model establishment by NSGA-II and the network structure-oriented selection of pareto optimal solution. In addition, the experimental results of the proposed MOSS algorithm are calculated on real networks, showing that

Related Content