Expert Systems With Applications 114 (2018) 503–515

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Rare category exploration with noisy labels

Haiqin Weng[a], Kevin Chiew[b], Zhenguang Liu[c], Qinming He[a,*], Roger Zimmermann[d]

[a] College of Computer Science and Technology, Zhejiang University, 310027, China
[b] CrimsonLogic Pte. Ltd., The Crimson, 31 Science Park Road, 117611, Singapore
[c] Zhejiang Gongshang University, Hangzhou, China
[d] National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 26 March 2018
Revised 23 July 2018
Accepted 24 July 2018
Available online 25 July 2018

Keywords:
Rare category exploration
Similarity matrix
Noisy labels
Label propagation

ABSTRACT

Starting from a few labelled data examples as the seeds, rare category exploration (RCE) aims to find out the target rare category hidden in the given dataset. However, the performance of conventional RCE approaches is very sensitive to noisy labels while the presence of noises in manually generated labels is almost inevitable. To address this deficiency of traditional RCE approaches, this paper investigates the RCE process in the presence of noisy labels, which to the best of our knowledge has not yet been intensively studied by previous research. Based on the assumption that only one labelled data example of the rare category is correctly labelled while the other few data examples may be wrongly labelled, we first propose a label propagation based algorithm SLP to extract the coarse shape of a rare category. Then, we refine the result by proposing a mixture-information based propagation model, RLP. Extensive experiments have been conducted on six real-world datasets, which show that our method outperforms the state-of-the-art RCE approaches. We also show that even with 20% noisy labels, our method is able to achieve a satisfactory accuracy.

1. Introduction

Rare categories were first proposed by Pelleg and Moore in Pelleg and Moore (2004), where a rare category was defined as tiny clusters of similar anomalies. Such rare categories are pervasive in real life. For instance, in cyber security, a rare category can be seen as a cluster of a specific kind of network attacks (Gu, Perdisci, Zhang, & Lee, 2008). Rare category examples (e.g., network attacks) are usually of higher significance than that of data examples from the major category (e.g., normal network accesses). Hence, rare category exploration (RCE) is proposed to conduct a comprehensive analysis on interesting rare categories.

1.1. Rare category exploration (RCE)

RCE is formally formulated as *finding the remaining data examples of a rare category from a few seeds of labelled data examples belonging to this rare category.* That is, extracting data examples of the rare category with the help of a few labels. The RCE technique can be widely used in many real-world applications. For example, in the area of financial transactions, a rare category may correspond to a small fraction of malicious transactions (Bay, Kumaraswamy, Anderle, Kumar, & Steier, 2006). Finding out all the malicious transactions can help us analyze the specific security leaks in a financial system.

Driven by the wide applications of RCE, researchers have explored various possibilities to solve the RCE task. A thread of the existing RCE algorithms relies on a few labelled data examples for model building and training (He, Tong, & Carbonell, 2010; Liu, Huang, He, Chiew, & Gao, 2015). He *el al.* proposed an optimization based solution to train a rare category classifier (He et al., 2010). Liu et al. utilized wavelet transformation for rare category exploration (Liu et al., 2015). Another thread of RCE algorithms concentrates on extracting a rare category only from one single correctly labelled data example from this rare category (Huang, Chiew, Gao, He, & Li, 2014).

1.2. Challenges in practical RCE scenarios

Though various solutions have been proposed for the RCE task, the following challenges still remain in many real-world applications. First, in practice, it is quite common that there may be some erroneous labels (referred to as noisy labels henceforth). These noisy labels are generated from many ways, including the false tagging by crowd sourcing (Kazai, Kamps, & Milic-Frayling, 2013; Krishna et al., 2016), the misdiagnoses by domain

* Corresponding author.
E-mail addresses: hq_weng@zju.edu.cn (H. Weng), kevinchiew@crimsonlogic.com (K. Chiew), lzg@zjgsu.edu.cn (Z. Liu), hqm@zju.edu.cn (Q. He), rogerz@comp.nus.edu.sg (R. Zimmermann).

https://doi.org/10.1016/j.eswa.2018.07.050
0957-4174/

experts (Rebbapragada, Brodley, Sulla-Menashe, & Friedl, 2012; Song, Wang, Zhang, Sun, & Yang, 2015), and even the adversarial countermeasures by attackers (Nelson et al., 2010; Wang, Wang, Zheng, & Zhao, 2014). However, most of the existing methods assume the labels are 100% correct and reliable. It seems that the performance of these algorithms will unfortunately degrade substantially in practical RCE scenarios. Second, rare categories are usually complex-shaped and even overlapped with major categories in some special cases. Therefore, such approaches that extract a rare category only from one correct label may wrongly identify some of the data examples located at the boundary of the rare category, due to the lack of sufficient label information.

1.3. Our methodology

To address the above challenges, we investigate an RCE scenario in the presence of noisy labels based on the following two basic assumptions, namely (1) the label for the seed of the objective rare category is correct, and (2) noisy labels exist across the other few labelled data examples from both rare and major categories. To this end, we first propose a label propagation based algorithm known as SLP (seed label propagation) to identify the coarse shape of the objective rare category from a seed. Equipped with a compactness based similarity matrix, SLP propagates the seed label along the high density areas of the objective rare category. Then, we develop a mixture-information based propagation model termed as RLP to extract the latent and useful information in the other few labels while explicitly addressing the noisy label issue. Specifically, RLP first refines the noisy labels and the coarse shape; then, it propagates the refined labels over the whole dataset to find those data examples not included in the coarse shape. Without any rigid assumption on the structure of a rare category, our algorithms can find complex-shaped rare categories in the presence of noisy labels.

In brief, the key contributions of this work are:

1. To our knowledge we are the first to explicitly define and address the rare category exploration problem in the presence of noisy labels.
2. We propose a compactness based similarity matrix among data examples for capturing the characteristics of rare categories. To handle a dataset with complex shaped rare categories, we propose the kernelized version of this compactness based similarity matrix.
3. Extensive experiments have been conducted on six real world datasets. Experimental results show that our proposed methods can achieve a relatively satisfactory performance even when 20% noisy labels are presented. As another contribution, we have proven the effectiveness of the proposed approaches via theoretical analysis.

The remaining sections are organized as follows. We review the related work in Section 2, and give the problem formulation and assumptions in Section 3. In Section 4, we first introduce the construction of the compactness based similarity matrix and its kernelized version, and then present the label propagation based algorithm SLP and the label-noise robust algorithm RLP. Finally, we report the experimental results and findings in Section 5 and conclude the paper in Section 6.

2. Related work

The related work can be classified into five categories, namely (1) imbalanced classification, (2) semi-supervised learning, (3) rare category detection, (4) rare category exploration, and (5) noisy labels learning.

2.1. Imbalanced classification

Imbalanced classification refers to the problem of constructing a classifier in the presence of under-represented datasets (Ertekin, Huang, & Giles, 2007; He et al., 2010; Hospedales, Gong, & Xiang, 2013; Liu & Zhou, 2006). Imbalanced classification can handle class distribution skews. There are four types of algorithms proposed for imbalanced classification, namely (1) nearest-neighbor based algorithms (He et al., 2010), (2) sampling based algorithms (Hospedales et al., 2013), (3) interactive learning based algorithms (Ertekin et al., 2007), and (4) cost-sensitive based algorithms (Liu & Zhou, 2006). Notably, the state-of-the-art RCE algorithm, RACH (He et al., 2010) falls into the category of cost-sensitive based imbalanced classification algorithm.

2.2. Semi-supervised learning

Semi-supervised learning uses both labelled and unlabelled data to perform an otherwise supervised learning or unsupervised learning task (Zhou & Belkin, 2014). Label propagation (LP) is a kind of semi-supervised learning (Chapelle, Weston, & Schölkopf, 2002; Li, Lu, Lin, Shen, & Price, 2015; Wang & Zhang, 2006; Zhou, Bousquet, Lal, Weston, & Schölkopf, 2003). The key to label propagation algorithms lies in two assumptions (Chapelle et al., 2002; Zhou et al., 2003), namely (1) nearby data examples are likely to have the same label belonging to the same major or rare category, and (2) data examples in the same cluster are likely to have the same label. Our method to construct a compactness based similarity matrix is motivated by the label propagation algorithm called LNP (Wang & Zhang, 2006), which propagates the labels of data examples to the whole dataset using their linear neighborhoods with sufficient smoothness.

Remarks. For the scenario of RCE in the presence of noisy labels, our algorithms can recover true label information from the given noisy labels, and focuses on finding the remaining examples of the objective rare category. Compared with our algorithms, when imbalanced classification methods are used for this scenario, they have several flaws, e.g., imbalanced classification methods usually require more computation since they conduct a holistic analysis on all data examples of a give data set. Semi-supervised learning may fail to work due to the lack of sufficient and correct label information. For example, LP may propagate the wrong label information to neighbors if there exists a large percentage of noises in the seeds, and fail to extract a rare category.

2.3. Rare category detection

By requesting only a small number of labels from a labelling oracle, Rare Category Detection (RCD) aims to discover a few data examples of a rare category to provide support for the existence of this rare category. Since RCD can find at least one data example of a rare category, it is often followed by an RCE process to discover the remaining data examples of the rare category. Pelleg et al. proposed the first RCD algorithm which assumed a mixture model to fit data examples (Pelleg & Moore, 2004). He et al. developed a k-nearest neighbors (kNN for short) based method to detect a rare category via unsupervised local-density-differential sampling strategy (He & Carbonell, 2007; He, Liu, & Lawrence, 2008). Besides, there exists some prior-free algorithms which addressed the case that any prior knowledge was not available (He & Carbonell, 2009; Huang, He, Chiew, Qian, & Ma, 2013; Liu, Chiew, He, Huang, & Huang, 2014a). For high dimensional datasets, Weng et al. proposed a tree-based algorithm to detect data examples of a rare category which have high scores of relative density (Weng, Liu, Chiew, & He, 2015). Most recently, Zhou et al. proposed two incremental