



Rare category exploration



Hao Huang^{a,b}, Kevin Chiew^c, Yunjun Gao^{a,*}, Qinming He^a, Qing Li^d

^a College of Computer Science, Zhejiang University, Hangzhou, China

^b School of Computing, National University of Singapore, Computing 1, 13 Computing Drive, Singapore 117417, Singapore

^c Provident Technology Pte. Ltd., 7030 Ang Mo Kio Ave 5, #03-25 Northstar, Singapore 569880, Singapore

^d Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, China

ARTICLE INFO

Keywords:

Rare category exploration
Local community
kNN graph
Histogram density estimation

ABSTRACT

Rare category discovery aims at identifying unlabeled data examples of rare categories in a given data set. The existing approaches to rare category discovery often need a certain number of labeled data examples as the training set, which are usually difficult and expensive to acquire in practice. To save the cost however, if these methods only use a small training set, their accuracy may not be satisfactory for real applications. In this paper, for the first time, we propose the concept of *rare category exploration*, aiming to discover all data examples of a rare category from a seed (which is a labeled data example of this rare category) instead of from a training set. To this end, we present an approach known as the FRANK algorithm which transforms rare category exploration to local community detection from a seed in a kNN (*k*-nearest neighbors) graph with an automatically selected *k* value. Extensive experimental results on real data sets verify the effectiveness and efficiency of our FRANK algorithm.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Rare categories (a.k.a. rare classes) are pervasive in real life. In medical diagnoses, although most of the diseases are well learned by current medicine, doctors still encounter rare diseases which may be new to the world. In the area of finances (Bay, Kumaraswamy, Anderle, Kumar, & Steier, 2006), the majority of transactions are legitimate, but there are still several fraudulent ones. In many applications, a data example of a rare category which cannot be classified into any of the known majority categories is usually discovered by an accident or via a detection approach such as rare category detection (He & Carbonell, 2007; He & Carbonell, 2009; He, Liu, & Lawrence, 2008; Huang, He, He, & Ma, 2011; Huang, He, Chiew, Qian, & Ma, 2013; Pelleg & Moore, 2004; Vatturi & Wong, 2009) which aims at finding out at least one data example from each rare category in an unlabeled data set. When we happen to discover a data example of a rare category, an intuitive idea is how to find out the remaining data examples of the rare category based on this discovered one as a seed. In this paper, we refer to this issue as *rare category exploration*.

Rare category exploration has various practical applications, especially when we are only interested in data examples from

the same rare category of a given data example. Besides the above areas of medical discovery and financial security, rare category exploration can also discover a series of similar feature changes on the surface of the earth through mass remote sensing images based on a specified one (Porter, Hush, Harvey, & Theiler, 2010), distinguish near-duplicate copies of a labeled spam image from a large number of non-spam images (Wang, Josephson, Lv, Charikar, & Li, 2007), and identify all other malicious activities with the same pattern of a detected network intrusion in a huge-volume traffic data set (Wu, Xiong, Wu, & Chen, 2007).

Among the existing research, although the imbalanced classification (e.g., Wu et al., 2007) and semi-supervised learning (e.g., Zhou, Weston, Gretton, Bousquet, & Schölkopf, 2003) can also be used to identify data examples of rare categories from those of majority categories, these two tasks are slightly different from rare category exploration from the following two aspects. (1) First, rare category exploration focuses on accurately discovering all other data examples of a rare category from *only one* seed which is a single labeled data example of this rare category, while imbalanced classification and semi-supervised learning often need *some* labeled data examples to achieve better classification performance in practice. (2) Second, rare category exploration only focuses on data examples with rare category characteristics such as compactness (He & Carbonell, 2007; He & Carbonell, 2009; He et al., 2008; He, Tong, & Carbonell, 2010; Huang et al., 2013; Huang et al., 2011; Vatturi & Wong, 2009) and isolation (Hospedales, Gong, & Xiang, 2013; Huang et al., 2013; Pelleg & Moore, 2004; Vatturi & Wong, 2009). In contrast, rather than specifically focusing on rare

* Corresponding author. Address: College of Computer Science, Zhejiang University, 38 Zheda Road, Hangzhou 310027, China. Tel.: +86 571 87651613; fax: +86 571 87951250.

E-mail addresses: huanghao@comp.nus.edu.sg (H. Huang), kev.chiew@gmail.com (K. Chiew), gaoyj@zju.edu.cn (Y. Gao), hqm@zju.edu.cn (Q. He), itqli@cityu.edu.hk (Q. Li).

category discovery, most of the approaches to imbalanced classification and semi-supervised learning mainly focus on constructing a classifier that optimizes a discriminative criterion for both majority and rare categories; therefore, they usually do not take full advantage of the characteristics of rare categories (He et al., 2010).

In this paper, we propose an approach known as FRANK (**F**ast **R**are **c**ategory **e**xploration using a **K**-nearest neighbors graph) algorithm to rare category exploration. By our approach, we firstly construct a k NN graph from a given data set with an automatically selected k value, and transform rare category exploration to local community detection from a seed in the k NN graph. During the exploration process, the local community keeps absorbing external data examples that are similar to its internal data examples until there is no more improvement of the local community quality evaluated by our proposed Compactness-Isolation (CI) metric. Finally, data examples in the local community are output as the candidate data examples of the rare category.

Our contributions can be summarized as follows. (1) To the best of our knowledge, it is the first time that we explicitly identify and solve the problem of *rare category exploration*. (2) We propose an efficient algorithm known as FRANK for rare category exploration which outperforms the existing approaches in terms of finding out the remaining data examples of the objective rare category from a seed. (3) We provide an automatic selection of the k value to construct a suitable k NN graph for FRANK, which can help further improve the performance of our FRANK algorithm. Note that since the automatic k selection is a distance-based approach which is also under a curse of dimensionality, our FRANK algorithm is not intended for high-dimensional data sets. Experimental results demonstrate that FRANK can handle data sets with up to 20 dimensions in satisfactory performance.

The remaining sections are organized as follows. We review the related work in Section 2 and present our problem statement in Section 3. In Section 4, we first introduce our assumption about rare categories, based on which we then transform our rare category exploration to local community detection in a k NN graph constructed from a given data set and propose a local community metric. In Section 5, we introduce the k selection for the k NN graph construction, and present our FRANK algorithm in Section 6, following which we report the experimental results in Section 7 before concluding the paper in Section 8.

2. Related work

The related work to rare category exploration can be classified into four categories, namely (1) imbalanced classification, (2) semi-supervised learning, (3) local community detection, and (4) rare category detection.

2.1. Imbalanced classification

Imbalanced classification refers to constructing a classifier for a data set with imbalanced category membership. There are four types of methods proposed for imbalanced classification, namely (1) the compactness-based methods (He et al., 2010), (2) the sampling-based methods (Hospedales et al., 2013; Wu et al., 2007), (3) the ensemble-based methods (Fernández-Baldera & Baumela, 2014), and (4) the methods adapting learning algorithms by moving decision thresholds or modifying classifiers' objective functions (Huang, Yang, King, & Lyu, 2004). Among these methods, the most related work to ours is the compactness-based method RACH (He et al., 2010), which constructs an optimization framework for finding a hyper-ball that encloses data examples of the objective rare

category with a minimum radius. RACH has some good properties such as a solid mathematical foundation and paying more attention to characteristics of rare categories than traditional imbalanced classification.

Compared with our FRANK algorithm, when imbalanced classification methods are used for rare category detection, they have the following flaw, i.e., imbalanced classification usually requires more computation since they conduct a holistic analysis on all data examples of a give data set. By contrast, FRANK carries out a local analysis on a minority part of the data set, and thus is more efficient.

2.2. Semi-supervised learning

Semi-supervised learning reveals the inherent data patterns from both labeled and unlabeled data examples. There are various paradigms for semi-supervised learning, such as co-training (Zhang, Wen, Wang, & Jiang, 2014), self-training (Liu, Jun, & Ghosh, 2009; Van Vaerenbergh, Santamaría, & Barbano, 2013), semi-supervised manifold learning (Zhou et al., 2003), and learning from positive and unlabeled data examples (Li & Liu, 2003; Liu, Dai, Li, Lee, & Yu, 2003). Among these semi-supervised approaches, the most related work to ours is (1) the work conducted by Zhou et al. (2003), which spreads the relevance from seeds to unlabeled data examples via a connected graph constructed with the given data set, and (2) the work on learning from positive and unlabeled data examples (Li & Liu, 2003; Liu et al., 2003), which aims at identifying the undiscovered data examples within the same category of the positive data examples.

Nevertheless, different from our proposed FRANK algorithm, these semi-supervised approaches are not specifically designed for rare category exploration. An extreme lack of labeled data examples often raises a risk of performance degradation for them. The experiments in this paper show that FRANK significantly outperforms them when there is only one labeled data example available as the seed for a rare category.

2.3. Local community detection

Local community detection aims at finding out local communities from some starting vertices in a graph structure. Each local community expands from a starting vertex by absorbing external vertices into it until either the expansion rate falls below some pre-defined threshold or the local community quality stops improving (Huang, Sun, Liu, Song, & Weninger, 2011b; Ma et al., 2013). Our approach to rare category exploration is inspired by local community detection since rare categories share much in common with local communities, i.e., their members are very similar to each other and differ from those outside.

In order to take full advantage of local community detection, our proposed FRANK algorithm adopts an automatic k selection to transform a given data set to a suitable k NN graph, in which the sub-graph corresponding to the objective rare category can form a community which is more compact than others and thus easier to be identified.

2.4. Rare category detection

Rare category detection discovers at least one data example for each rare category in an unlabeled data set by selecting candidate data examples for labeling. Methods proposed so far for rare category detection can be classified into four categories, namely (1) the mixture model-based (Pelleg & Moore, 2004), (2) the nearest neighborhood-based (Huang et al., 2011; Huang et al., 2013), (3) the density differential-based (He & Carbonell, 2007; He & Carbonell,

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات