Expert Systems With Applications 63 (2016) 173–186

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Rare category exploration via wavelet analysis: Theory and applications

Zhenguang Liu[a], Kevin Chiew[b], Luming Zhang[c,d], Beibei Zhang[a], Qinming He[e,*], Roger Zimmermann[a]

[a] School of Computing, National University of Singapore, 117417, Singapore
[b] Handal Indah Sdn Bhd, 728789, Singapore
[c] Department of Computer and Information, Hefei University of Technology, Anhui, 230009, China
[d] National University of Singapore Research Institute, Suzhou, 215123, China
[e] College of Computer Science, Zhejiang University, Hangzhou, 310027, China

ARTICLE INFO

Article history:
Received 2 May 2016
Revised 15 June 2016
Accepted 16 June 2016
Available online 29 June 2016

Keywords:
Rare category exploration
Wavelet transform
Linear time complexity
Bandwidth selection

ABSTRACT

Rare category exploration (in short as RCE) aims to discover all the remaining data examples of a rare category from a known data example of the rare category. A few approaches have been proposed to address this problem. Most of them, however, are on quadratic or even cubic time complexities w.r.t. data set size *n*. More importantly, the F-scores (harmonic mean of precision and recall) of the existing approaches are not satisfactory. Compared with the existing solutions to RCE, this paper proposes a novel approach with a linear time complexity and achieves a higher F-score of mining results. The key steps of our approach are to reduce search space by performing wavelet analysis on the data density function, and then refine the coarse mining result in the reduced search space via fine-grained metrics. A solid theoretical analysis is conducted to prove the feasibility of our solution, and extensive experiments on real data sets further verify its effectiveness and efficiency.

1. Introduction

Given a data set for mining, besides the major categories in which all data examples are normal, there may be a few rare categories each of which contains some anomalous data examples with the same pattern. These anomalies may be of special interest to researchers (Alhammady & Ramamohanarao, 2004; O'Reilly, Gluhak, & Imran, 2015). In financial activities, for example, among thousands of daily transactions, a few of them may be anomalies as compared with normal transactions (He, Tong, & Carbonell, 2010). These anomalies may intrigue us because they could be fraudulent transactions which can harm the stakeholders and should be identified early to prevent further fraud from happening (He, 2010).

Rare category exploration (in short as RCE) aims to identify all the remaining data examples of a rare category from a given data example of the rare category. That is, upon detecting an interesting data example, a natural idea is to identify all the other data examples similar to the detected one (Huang, Chiew, Gao, He, & Li, 2014a). RCE is different from rare category detection problem. Rare category detection aims to find at least one data example for each rare category to prove the existence of this category in the unlabeled data set, while RCE focuses on mining all the data examples of an interesting rare category from the unlabeled data set.

The aim of RCE has enabled it to have a wide variety of applications. For example, in medical diagnoses, after detecting a patient with a rare disease, RCE can identify other patients who have similar symptoms from a vast collection of records; in financial security, after detecting a fraudulent transaction, RCE allows us to identify other fraudulent transactions of the same type, which helps analyze the security leaks of the system and prevent new fraudulent transactions (He, 2010; Liu, Huang, He, Chiew, & Gao, 2015). Other reported applications include detection of oil spills in satellite radar images, detection of network intrusions (Sun, Kamel, Wong, & Wang, 2007), etc.

Due to the wide application prospects, a number of RCE solutions have been proposed in the literatures. He *et al.* converted RCE to a convex optimization problem and tried to represent the rare category with a hyperball (He et al., 2010). Huang et al. formulated RCE as a local community detection problem and derived a

* Corresponding author.Fax: +86 571 86971956.
E-mail addresses: lzg@nus.edu.sg, zhenguangliu@zju.edu.cn (Z. Liu), kchiew@handalindah.com.my (K. Chiew), beibei.zhang@connect.polyu.hk (B. Zhang), hqm@zju.edu.cn (Q. He), rogerz@comp.nus.edu.sg (R. Zimmermann).

http://dx.doi.org/10.1016/j.eswa.2016.06.033
0957-4174/

solution based on the compactness and isolation assumptions of rare categories (Huang et al., 2014a).

However, the time complexities of most existing RCE algorithms are quadratic or even cubic. The high computational expenses may hinder their applications to big data sets. More importantly, the F-scores of the existing RCE algorithms are not satisfactory. F-score is the harmonic mean of precision and recall, which is commonly used (Huang et al., 2014a; Liu, Chiew, He, Huang, & Huang, 2014) to measure the quality of RCE algorithms. Higher F-score means higher quality of mining the target rare category. Unsatisfactory F-scores of the existing RCE algorithms may lead to misunderstanding of the intrinsic characteristics of the target rare category.

This paper proposes a novel approach termed RCEWA (Rare Category Exploration via Wavelet Analysis) which achieves a linear time complexity w.r.t. the size of the data set and a higher F-score as compared with existing algorithms. In detail, we (1) perform feature space partition to build the data density function, (2) apply wavelet analysis (which is continuous wavelet transform in our settings) on the data density function to locate the target rare category, and (3) refine the coarse shape via k-means operation and connected sub-cluster search to obtain the final individual data examples of the rare category.

The main contributions of this paper can be summarized as follows. (1) We propose a novel approach for RCE which achieves a linear time complexity. (2) We provide a solid theoretical proof for the effectiveness of using wavelet analysis for RCE. (3) Extensive experiments on various data sets show that our algorithm significantly and consistently outperforms the existing algorithms in terms of F-score.

The remaining sections are organized as follows. We review the related work in Section 2, then give the problem statement in Section 3. Next, we elaborate on our proposed RCEWA algorithm in Section 4 with a solid theoretical analysis. After that, we present the potential applications of RCE and summarize the findings of experimental results in Section 5. At last, we conclude the paper in Section 6.

2. Related work

For a better understanding, we classify the closely related work of RCE into three parts, namely rare category detection, imbalanced classification and existing approaches to the RCE problem. The details are introduced below. We also note that existing works on correlation analysis (e.g., Eches, Dobigeon, and Tourneret (2011), Giotis and Guruswami (2006), Klein, Mathieu, and Zhou (2015), Swamy (2004)) can be extended for the task of RCE by utilizing the correlation between data examples. However, these approaches usually tend to perform unsatisfactorily for RCE due to not taking full advantage of rare category properties.

2.1. Rare category detection

Rare category detection aims to discover at least one data example for each rare category in an unlabeled data set (He & Carbonell, 2007, 2009; Huang, He, He, & Ma, 2011). RCE can be seen as a natural follow-up action of rare category detection. That is, after detecting a data example of a rare category that is interesting, an natural idea is to find out the other data examples in the same rare category (He et al., 2010; Huang et al., 2014a). Existing algorithms for rare category detection typically utilize the compactness and isolation properties to search data examples of rare categories. These algorithms can be classified into three groups, *i.e.,* the model-based (Liu et al., 2014), the neighbor-based (He & Carbonell, 2007; 2009; Huang, He, Chiew, Qian, & Ma, 2013; Huang et al., 2011), and the hierarchical-clustering-based (Vatturi & Wong, 2009; Weng, Liu, Chiew, & He, 2015) approaches.

2.2. Imbalanced classification

Imbalanced classification has been proposed to construct a classifier that optimizes a discriminative criterion for both major categories and rare categories (He et al., 2010; Kim, Kang, & Kim, 2015; López, Triguero, Carmona, García, & Herrera, 2014; Tang, Zhang, Chawla, & Krasser, 2009; Thammasiri, Delen, Meesad, & Kasap, 2014). Existing methods on imbalanced classification can be employed for RCE by returning the data examples which are classified into rare categories (He et al., 2010). Notably, their performance usually cannot meet the expectations because they are not specially designed for the RCE task.

2.3. Existing approaches to RCE problem

Existing approaches to RCE can be classified into three general types, *i.e.,* optimization-based approaches, community-detection-based approaches and density-based approaches.

Optimization-based approaches. Optimization-based approaches convert RCE into a convex optimization problem (He et al., 2010), and try to enclose the rare category data examples with a minimum-radius hyper-ball. Optimization-based approaches, *e.g.,* RACH (He et al., 2010), can handle the scenario in which the target rare category overlaps with a major category. Notably, to build a training set, they require a certain number of labeled data examples which might be difficult and expensive to acquire in practice. Besides, the computational time complexities of these approaches become quadratic on data set size when only one seed data example is known. This is because the filtering process of these approaches (He et al., 2010) will fail in this case.

Community-detection-based approaches. Community-detection-based approaches formulate the RCE problem into a local community detection problem (Huang et al., 2014a). These approaches keep absorbing external data examples until no improvement in the quality of the local community is observed. Community-detection-based approaches, *e.g.,* FRANK (Huang et al., 2014a), require the data examples of a rare category being isolated from others. The time complexities are quadratic w.r.t. data set size since they have to construct the kNN graph of data examples.

Density-based approaches. Density-based approaches (Liu et al., 2015) organize data examples into small clusters and perform density analysis to find candidate data examples of the target rare category. Density-based approaches require different parts of a rare category to have close data density. FREE (Liu et al., 2015) as a typical density-based approach, to the best of our knowledge, is the only method for RCE that achieves a linear time complexity w.r.t. data set size. However, since it requires different parts of a rare category to have close data density, the performance of FREE becomes quite unsatisfactory for the scenarios where the data density of a rare category change smoothly (*e.g.,*follows a Gaussian distribution).

3. Problem statement and preliminary

Based on the existing work on RCE (Huang et al., 2014a; Liu et al., 2015), we formulate RCE as follows.

Given an unlabeled data set and a seed which is a known data example of a rare category **S**, find the remaining data examples of rare category **S** from the given data set.

As aforementioned, most of the existing work makes two assumptions on the basic properties of a rare category explicitly or implicitly (He & Carbonell, 2009; He et al., 2010; Huang et al., 2014a; Huang et al., 2013; Liu et al., 2014; Liu et al., 2015; Sun et al., 2007). The two assumptions are *compactness* and *isolation*, meaning that in terms of data distribution of the given data set,