Gold Panning from the Mess: Rare Category Exploration, Exposition, Representation, and Interpretation

Dawei Zhou
davidchouzdw@gmail.com
Arizona State University
Tempe, AZ

Jingrui He
jingrui.he@gmail.com
Arizona State University
Tempe, AZ

ABSTRACT

In contrast to the massive volume of data, it is often the rare categories that are of great importance in many high impact domains, ranging from financial fraud detection in online transaction networks to emerging trend detection in social networks, from spam image detection in social media to rare disease diagnosis in the medical decision support system. The unique challenges of rare category analysis include: (1) the highly-skewed class-membership distribution; (2) the non-separability nature of the rare categories from the majority classes; (3) the data and task heterogeneity, e.g., the multi-modal representation of examples, and the analysis of similar rare categories across multiple related tasks. This tutorial aims to provide a concise review of state-of-the-art techniques on complex rare category analysis, where the majority classes have a smooth distribution, while the minority classes exhibit a compactness property in the feature space or subspace. In particular, we start with the context, problem definition and unique challenges of complex rare category analysis; then we present a comprehensive overview of recent advances that are designed for this problem setting, from rare category exploration without any label information to the exposition step that characterizes rare examples with a compact representation, from representing rare patterns in a salient embedding space to interpreting the prediction results and providing relevant clues for the end users' interpretation; at last, we will discuss the potential challenges and shed light on the future directions of complex rare category analysis.

ACM Reference Format:
Dawei Zhou and Jingrui He. 2019. Gold Panning from the Mess: Rare Category Exploration, Exposition, Representation, and Interpretation. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332268

1 TUTORIAL DESCRIPTION

1.1 Tutorial Website

https://sites.google.com/view/kdd19-tutorial-rca/home

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3332268

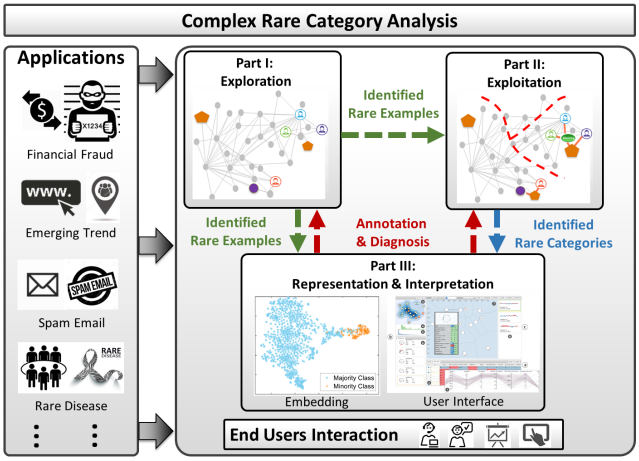

Figure 1: An overview of complex rare category analysis.

1.2 Audience

- **Targeted audience**: The tutorial is designed for anyone with the basic knowledge of data mining, artificial intelligence, or machine learning. The content level of this tutorial is designed for 50% novice, 30% intermediate, 20% expert.
- **Audience prerequisites**: We aim to attract both researchers and practitioners working in high impact application domains such as finance, healthcare, social networks / media, manufacturing, etc.

1.3 Outline

- **Introduction**: In this part, we will start with the definition of rare categories, along with the key challenges in rare category analysis problems. Then, we will provide an overview of rare category analysis (illustrated in Fig. 1), summarizing the underlying problems and major techniques in this problem setting. Finally, we will discuss the impact of rare category analysis in real-world applications.
  - What is a rare category?
  - What are the key challenges in complex rare category analysis?
  - What are some real-world applications for complex rare category analysis?
- **Part I: Rare Category Exploration**: In this part, we will briefly introduce the exploratory step for rare category analysis, which is also referred to as *rare category detection*. Depending on the availability of multi-modal representation of

examples, rare category detection can be categorized into the following two groups:
  - Homogeneous rare category detection
  - Heterogeneous rare category detection
- **Part II: Rare Category Exploitation**: In this part, we will discuss the existing work of rare category exploitation, which aims to capture a compact representation for the rare categories in various types of data including:
  - Attributed data
  - Time series data
  - Graph/network data
- **Part III: Rare Category Representation & Interpretation**: In this part, we will discuss the recent advances of Rare Category Representation and Interpretation, which serve as an investigation step for the end users to visualize the data distribution and inspect the underlying prediction process in the previous steps (i.e., Part I, Part II).
  - Rare Category Representation
  - Rare Category Interpretation
- **Part IV: Applications**: In this part, we will introduce various applications of the analysis of complex rare categories, ranging from financial fraud detection in online transaction networks to emerging trend detection in social networks, from spam image detection in social media to rare disease diagnosis in medical decision support system.
- **Part V: Conclusion and Future Directions**: In this part, we will conclude the existing work and share our thoughts regarding the future directions of complex rare category analysis such as:
  - Rare Category Interpretation
  - Rare Category Robustness
  - Rare Category Generation

2 BIOGRAPHIES

In-person presenters and contributors include Dawei Zhou and Jingrui He. Their biographies and expertises are elaborated as follows.

Dawei Zhou is currently a Ph.D. student at the School of Computing, Informatics and Decision Systems Engineering, Arizona State University. Before that, he received the M.S degree from Department of Electrical and Computer Engineering, University of Rochester in 2014 and B.E degree from College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications in 2012. His current research interests include rare category analysis, active learning, and semi-supervised learning, with applications in financial fraud detection, social network analysis. Dawei Zhou has worked on rare category analysis for 4 years, which results in 10 publications at major conferences (e.g., IJCAI, AAAI, KDD, SDM, ICDM, BigData) and journals (e.g., TKDD, DMKD, Frontier), such as [11–14, 14–21]. For more information, please visit his homepage at https://sites.google.com/view/dawei-zhou/home.

Dr. Jingrui He is an associate professor at the School of Computing, Informatics and Decision Systems Engineering, Arizona State University. She received her Ph.D. degree from Machine Learning Department, Carnegie Mellon University in 2010. Her research interests include rare category analysis, heterogeneous learning, semi-supervised learning, and active learning, with applications in semiconductor manufacturing, social media analysis, traffic prediction, healthcare, etc. Dr. He has worked on the topic of rare category analysis for over 10 years, which results in more than 30 publications at major conferences (e.g., NIPS, IJCAI, AAAI, KDD, ICML, ICDM), journals (e.g., TKDE, TKDD, DMKD) and 1 books, such as [1–3, 3–10].

REFERENCES

[1] J. He. *Analysis of rare categories*. Springer Science & Business Media, 2012.
[2] J. He and J. Carbonell. Prior-free rare category detection. In *SDM*, pages 155–163. SIAM, 2009.
[3] J. He and J. Carbonell. Coselection of features and instances for unsupervised rare category analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2010.
[4] J. He and J. G. Carbonell. Nearest-neighbor-based active learning for rare category detection. In *NIPS*, 2008.
[5] J. He and J. G. Carbonell. Rare class discovery based on active learning. 2008.
[6] J. He, Y. Liu, and R. Lawrence. Graph-based rare category detection. In *ICDM*. IEEE, 2008.
[7] J. He, H. Tong, and J. Carbonell. Rare category characterization. In *ICDM*. IEEE, 2010.
[8] J. He, H. Tong, and J. Carbonell. An effective framework for characterizing rare categories. *Frontiers of Computer Science*, 2012.
[9] H. Lin, S. Gao, D. Gotz, F. Du, J. He, and N. Cao. Rclens: Interactive rare category exploration and identification. *IEEE transactions on visualization and computer graphics*, 2017.
[10] P. Yang, J. He, and J.-Y. Pan. Learning complex rare categories with dual heterogeneity. In *SDM*. SIAM, 2015.
[11] P. Yang, H. Yang, H. Fu, D. Zhou, J. Ye, T. Lappas, and J. He. Jointly modeling label and feature heterogeneity in medical informatics. *ACM Transactions on Knowledge Discovery from Data (TKDD-2015)*, 2015.
[12] S. Zhang, D. Zhou, M. Y. Yildirim, S. Alcorn, J. He, H. Davulcu, and H. Tong. Hidden: hierarchical dense subgraph detection with application to financial fraud detection. In *SDM*. SIAM, 2017.
[13] D. Zhou, J. He, K. S. Candan, and H. Davulcu. Muvir: Multi-view rare category detection. In *IJCAI*, 2015.
[14] D. Zhou, J. He, Y. Cao, and J. Seo. Bi-level rare temporal pattern detection. In *ICDM*. IEEE, 2016.
[15] D. Zhou, J. He, H. Davulcu, and R. Maciejewski. Motif-preserving dynamic local graph cut. In *BigData*. IEEE, 2018.
[16] D. Zhou, J. He, H. Yang, and W. Fan. Sparc: Self-paced network representation for few-shot rare category characterization. In *SIGKDD*. ACM, 2018.
[17] D. Zhou, A. Karthikeyan, K. Wang, N. Cao, and J. He. Discovering rare categories from graph streams. *DMKD*, 2017.
[18] D. Zhou, J. Luo, V. Silenzio, Y. Zhou, J. Hu, G. Currier, and H. Kautz. Tackling mental health by integrating unobtrusive multimodal sensing. In *AAAI*, 2015.
[19] D. Zhou, K. Wang, N. Cao, and J. He. Rare category detection on time-evolving graphs. In *ICDM*. IEEE, 2015.
[20] D. Zhou, S. Zhang, M. Y. Yildirim, S. Alcorn, H. Tong, H. Davulcu, and J. He. A local algorithm for structure-preserving graph cut. In *SIGKDD*. ACM, 2017.
[21] D. Zhou, L. Zheng, J. Xu, and J. He. Misc-gan: A multi-scale generative model forgraphs. *Frontiers in Big Data*, 2019.