

## Preface

In many real world problems, rare categories (minority classes) play an essential role despite their extreme scarcity. For example, in financial fraud detection, the vast majority of financial transactions are legitimate, and only a small number may be fraudulent; in Medicare fraud detection, the percentage of bogus claims is small, but the total loss is significant; in network intrusion detection, malicious network activities are hidden among huge volumes of routine network traffic; in astronomy, only 0.001% of the objects in sky survey images are truly beyond the scope of current science and may lead to new discoveries; in spam image detection, near-duplicate spam images are difficult to discover from the large number of non-spam images; in rare disease diagnosis, rare diseases affect less than 1 out of 2000 people, but the consequences can be very severe. Therefore, the discovery, characterization and prediction of rare categories or rare examples may protect us from fraudulent or malicious behaviors, provide aid for scientific discoveries, and even save lives.

This book focuses on the analysis of rare categories, where the majority classes have a smooth distribution, and the minority classes exhibit a compactness property. Furthermore, we focus on the challenging case where the support regions of the majority and minority classes overlap each other. To the best of our knowledge, this book is the first end-to-end investigation of rare categories.

In this book, we focus on both supervised and unsupervised rare category analysis depending on the availability of the label information. In the supervised settings, our first task is rare category detection, which is to discover at least one example from each minority class with the help of a labeling oracle. For this task, simply applying random sampling proves very expensive in terms of the number of label requests to the oracle, so we need to design more effective methods. Then, given labeled examples from all the classes, our second task is rare category characterization. The goal here is to find a compact representation for the minority classes in order to identify all the rare examples with high precision and recall. On the other hand, in the unsupervised settings, we do not have access to a labeling

oracle. Here we propose to co-select candidate examples from the minority classes and the relevant features, which benefits both tasks (rare category selection and feature selection). For each of the above tasks, we have designed effective algorithms with theoretical guarantees as well as good empirical results.

Jingrui He  
08/31/2011



<http://www.springer.com/978-3-642-22813-1>

Analysis of Rare Categories

He, J.

2012, VIII, 136 p.,

ISBN: 978-3-642-22813-1