Chapter 4

Rare Category Characterization

In Chapter 3, we have introduced various algorithms for rare category detection, which result in a set of labeled examples. Based on this labeled set, a natural follow-up step is rare category characterization, i.e., to characterize the minority classes in order to identify all the rare examples in the data set. For example, in Medicare fraud detection, once we have discovered a bogus claim for durable equipments (e.g., wheelchairs, breathing machines), we may want to find the fraud patterns related to such equipments in order to prevent similar fraudulent claims in the future. To this end, in this chapter, we focus on rare category characterization, the second task in the supervised setting.

Rare category characterization is to characterize the minority classes for the purpose of understanding and correctly classifying those classes. Here, our key observation is that the minority classes often exhibit compactness. That is, each minority class often forms a compact cluster. For example, fraudulent people often make multiple similar transactions to maximize their profits [CPF06]. For rare diseases, the patients with the same type of rare disease often share similar genes or chromosomal abnormalities [EUR05].

In this chapter, we propose RACH by exploring such compactness for rare category characterization. The core of RACH is to represent the minority classes with a hyper-ball. We present the optimization framework as well as an effective algorithm to solve it. Furthermore, we show how RACH can be naturally kernelized. We also analyze the complexity of RACH. Finally, we justify the effectiveness of the proposed RACH by both theoretical analysis and empirical evaluations.
The key content of this chapter can be summarized as follows.

**Problem Formulation.** We formulate the problem of rare category characterization as an optimization problem, which takes into account both labeled and unlabeled examples, and imposes different constraints for different types of data;

**Algorithm Design.** We design an effective algorithm to find the solution of the optimization problem. It repeatedly converts the original problem into a convex optimization problem, and solves it in its dual form by a projected subgradient method, which is well justified theoretically.

The rest of this chapter is organized as follows. In Section 4.1, we propose the optimization framework to provide a compact representation for the minority class with justification, followed by the conversion of this framework to the convex optimization problem as well as its dual form. Then we introduce the RACH algorithm to solve the dual problem with performance guarantees in Section 4.2, and the kernelized RACH algorithm in Section 4.3. Finally, following the experimental results presented in Section 4.4, we give a brief summary of rare category characterization in Section 4.5.

### 4.1 Optimization Framework

In this section, we present our optimization framework, after we introduce additional notation, assumptions and the pre-processing step.

Besides the notation introduced in Chapter 1, we also assume that there is only one majority class and one minority class in the data set, i.e., $m = 2$. (Multiple majority and minority classes can be converted into several binary problems.) Throughout this chapter, we will use calligraphic capital letters to denote sets. Let $x_1, \ldots, x_{n_1} \in \mathbb{R}^d$ denote the labeled examples from the majority class, which correspond to $y_i = 1, i = 1, \ldots, n_1$; let $x_{n_1+1}, \ldots, x_{n_1+n_2} \in \mathbb{R}^d$ denote the labeled examples from the minority class, which correspond to $y_j = 2, j = n_1+1, \ldots, n_1+n_2$; let $x_{n_1+n_2+1}', \ldots, x_{n'}' \in \mathbb{R}^d$ denote all the unlabeled examples. Here, $n_1$, $n_2$, and $n'$ denote the number of labeled examples from the majority class, the number of labeled examples from the minority class, and the total number of examples, both labeled and unlabeled. $d$ is the dimensionality of the input space. Our goal is to identify with high precision and recall a list of unlabeled examples which are believed to come from the minority class.

For rare category characterization, we make the following assumptions: the rare examples from the same minority class are very close to each other,