Chapter 2
Survey and Overview

Rare category analysis is related to many research areas, including active learning, where the goal is to improve the classification performance with the fewest label requests to the labeling oracle; imbalanced classification, where the goal is to construct a classifier for imbalanced data sets which is able to identify the under represented classes; anomaly detection (outlier detection), which refers to the problem of finding patterns in the data that do not conform to expected behavior; clustering, which refers to the problem of grouping similar data items into clusters; co-clustering, which generally involves grouping the data from various dimensions; and unsupervised feature selection, where the goal is to select features for the sake of grouping the data without any supervision. In this chapter, we review related work in the above areas, highlighting their differences with rare category analysis. Compared with these research areas, rare category analysis is relatively new. In this chapter, we also briefly introduce some existing work on rare category detection, which is the first task in supervised rare category analysis; whereas the other tasks have not been addressed before.

2.1 Active Learning

The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns [Set10]. In active learning, we assume that the class labels are obtained from a labeling oracle with some cost, and under a fixed budget, we hope to maximally improve the performance of the learning algorithm. According to [Set10], there are three main settings in active learning: membership query synthesis, stream-based selective sampling, and
pool-based sampling.

Many early active learning algorithms belong to membership query synthesis, such as [Ang87] [Ang01] [CGJ96]. One major problem with membership query synthesis is that the synthesized queries often have no practical meanings, and thus no appropriate labels. On the other hand, with stream-based active learning and pool-based sampling, the queries always correspond to real examples. Therefore, their label information can be readily provided by the oracle.

In stream-based selective sampling, given an unlabeled example, the learner must decide whether to query its class label or to discard it. For example, in [CAL92], Cohn et al compute a region of uncertainty, and query examples within it; in [DE95], Dagan et al proposed committee-based sampling, which evaluates the informativeness of an example by measuring the degree of disagreement between several model variants and only queries the more informative ones.

On the other hand, in pool-based sampling, queries are selected from a pool of unlabeled examples. Its major difference from stream-based selective sampling is the large amount of unlabeled data available at query time, which reveals additional information about the underlying distribution. For example, Tong et al [TKK01] proposed an active learning algorithm that minimizes the size of the version space; McCallum and Nigam [MN98] modified the Query-by-Committee method of active learning to use unlabeled data for density estimation, and combined this with EM to find the class labels of the unlabeled examples.

It should be mentioned that in traditional active learning, initially we have labeled examples from all the classes in order to build the very first classifier, which can be improved by actively selecting the training data. On the other hand, in rare category detection, initially we do not have any labeled examples, and the goal is to discover at least one example from each minority class with the fewest label requests. Combining rare category detection and traditional active learning, it has been noticed in [BBL06] and [Das05] that if the learning algorithm starts denovo, finding the initial labeled examples from each class (i.e., rare category detection) becomes the bottleneck for reducing the sampling complexity. Furthermore, in supervised rare category analysis, following rare category detection, the second task is rare category characterization, which works in a semi-supervised fashion. In this task, in order to get a more accurate representation of the minority classes, we can make use of active learning to select the most informative examples to be added to the labeled set.