

# 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.

## 2.2 Imbalanced Classification

In imbalanced classification, the goal is to construct an accurate classifier that optimizes a discriminative criterion, such as balanced accuracy, G-mean, etc [Cha09]. Existing methods can be roughly categorized into three groups [Cha09], i.e., sampling-based methods [KM97][CBHK02], adapting learning algorithms by modifying objective functions or changing decision thresholds [WC03] [HYKL04], and ensemble-based methods [SKW06][CLHB03]. To be specific, in sampling-based methods, some methods under-sample the majority classes. For example, the one-sided sampling strategy proposed in [KM97] employs Tomek links [Tom76] followed by closest nearest neighbor [HAR68] to discard the majority class examples that lie in the borderline region, are noisy or redundant. In contrast, some sampling-based methods over-sample the hard examples. For example, the DataBoost-IM method proposed in [GV04] generates synthetic examples according to the hard examples identified during the boosting algorithm; the SMOTEBoost algorithm proposed in [CLHB03] applies the SMOTE algorithm [CBHK02] to create new examples from the minority class in each boosting round. Furthermore, some methods combine over-sampling the minority class and under-sampling the majority class. For example, the SMOTE algorithm combined with under-sampling [CBHK02] was proven to outperform only under-sampling the majority class and varying the loss ratios; in [TZCK09], different rebalance heuristics were incorporated into SVM modeling to tackle the problem of class imbalance, including over-sampling, under-sampling, etc.

Imbalanced classification and rare category characterization bear similarity but also have some differences. On one hand, both tasks need labeled examples from all the classes as input. On the other hand, imbalanced classification and rare category characterization have different goals as well as different output. To be specific, in rare category characterization, we only focus on the minority classes, and aim to identify all (or nearly all) the rare examples from the unlabeled data set with high precision and recall; whereas in imbalanced classification, the goal is to construct a classifier that optimizes a discriminative criterion for both the majority and minority classes. Furthermore, in rare category characterization, we are able to obtain a compact representation for the minority classes, which can be provided to domain experts for better understanding of the learning results; whereas in imbalanced classification, such representations are not provided.

### 2.3 Anomaly Detection (Outlier Detection)

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior [CBK09]. Anomalies are often referred to as outliers. According to [CBK09], the majority of anomaly detection techniques can be categorized into classification-based, nearest neighbor-based, clustering-based, information theoretic, spectral, and statistical techniques. For example, in [BWJ01], the authors propose a method based on a technique called pseudo-Bayes estimators to enhance an anomaly detection system's ability to detect new attacks while reducing the false alarm rate as much as possible. In [RRS00], the authors propose a novel formulation for distance-based outliers that is based on the distance of a point from its  $k^{\text{th}}$  nearest neighbor. Then they rank each point on the basis of its distance to its  $k^{\text{th}}$  nearest neighbor and declare the top  $n$  points in this ranking to be outliers. In [YSZ02], the authors propose the *FindOut* algorithm, which is an extension of the *WaveCluster* algorithm [SCZ98] in which the detected clusters are removed from the data and the residual instances are declared as anomalies. In [HXD05], the authors formally define the problem of outlier detection in categorical data as an optimization problem from a global viewpoint based on entropy minimization, and present a local-search heuristic-based algorithm for efficiently finding feasible solutions. In [DGBK07], the authors describe distributed algorithms for doing Principal Component Analysis (PCA) using random projection and sampling based techniques. Using the approximate principal components, they develop a distributed outlier detection algorithm based on the fact that the last principal component enables identification of data points which deviate sharply from the 'correlation structure' of the data. And in [AY01], the authors discuss a new technique for outlier detection which is especially suited to very high dimensional data sets. The method works by finding lower dimensional projections which are locally sparse, and cannot be discovered easily by brute force techniques because of the number of combinations of possibilities.

In general, anomaly detection finds *individual* and *isolated* examples that differ from a given class in an unsupervised fashion. Typically, there is no way to characterize the anomalies since they are often different from each other. There exist a few works dealing with the case where the anomalies are clustered [PKGF03]. However, they still assume that the anomalies are separable from the normal data points. On the other hand, in rare category detection, each rare category consists of a group of points, which form a compact cluster in the feature space and are self-similar. Furthermore, we are dealing with challenging cases where the support regions of the majority

and minority classes overlap with each other.

## 2.4 Clustering

According to [Fun01], clustering refers to the grouping together of similar data items into clusters. Existing clustering algorithms can be categorized into the following two main classes [Fun01]: parametric clustering and non-parametric clustering. In general, parametric methods attempt to minimize a cost function or an optimality criterion which associates a cost to each example-cluster assignment. It can be further classified into two groups: generative models and reconstructive models. In generative models, the basic idea is that the input examples are observations from a set of unknown distributions. For example, in Gaussian mixture models [RR95], the data are viewed as coming from a mixture of Gaussian probability distributions, each representing a different cluster; in C-means fuzzy clustering [Dun73], the membership of a point is shared among various clusters. On the other hand, reconstructive methods generally attempt to minimize a cost function. For example, K-means clustering forms clusters in numeric domains, partitioning examples into disjoint clusters [DHS00]; in Deterministic Annealing EM Algorithm (DAEM) [HB97], the maximization of the likelihood function is embedded in the minimization of the thermodynamic free energy, depending on the temperature which controls the annealing process. For nonparametric methods, two good representative examples are the agglomerative and divisive algorithms, also called hierarchical algorithms [Joh67], that produce dendrograms.

In unsupervised rare category analysis, one important problem we want to address is rare category selection, i.e., selecting a set of examples which are likely to come from the minority classes. General-purpose clustering algorithms do not fit here because the proportions of different classes are very skewed and the support regions of the majority and minority classes overlap with each other. In this case, general-purpose clustering algorithms tend to overlook the minority classes and generate clusters within the majority classes. Therefore, we need to develop new methods for rare category selection which leverage the property of the minority classes.

## 2.5 Co-clustering

The idea of using compression for co-clustering can be traced back to the information-theoretic co-clustering algorithm [DMM03], where the normal-

ized non-negative contingency table is treated as a joint probability distribution between two discrete random variables that take values over the rows and columns. Then co-clustering is defined as a pair of mappings from rows to row clusters and from columns to column clusters. According to information theory, the optimal co-clustering is the one that minimizes the difference between the mutual information of the original random variables and the clustered random variables. The algorithm for minimizing the above criterion intertwines both row and column clustering at all stages. Row clustering is done by assessing closeness of each row distribution, in relative entropy, to certain ‘row cluster prototypes’. Column clustering is done similarly, and this process is iterated until it converges to a local minimum. It can be theoretically proven that the proposed algorithm never increases the criterion, and gradually improves the quality of co-clustering.

Although the information-theoretic co-clustering algorithm can only be applied to bipartite graphs, the idea behind this algorithm can be generalized to more than two types of heterogeneous objects. For example, in [GLZ<sup>+</sup>07], the authors proposed the CBGC algorithm. It aims to do collective clustering for star-shaped inter-relationships among different types of objects. First, it transforms the star-shaped structure into a set of bipartite graphs; then it formulates a constrained optimization problem, where the objective function is a weighted sum of the Rayleigh quotients on different bipartite graphs, and the constraints are that clustering results for the same type of objects should be the same. Follow-up work includes high order co-clustering [GGP07]. Another example is the spectral relational clustering algorithm proposed in [LZWY06]. Unlike the previous algorithm, this algorithm is not restricted to star-shaped structures. It is based on a general model, collective factorization on related matrices. This model simultaneously clusters multi-type interrelated objects based on both the relation and the feature information. It exploits the interactions between the hidden structures of different types of objects through the related factorizations which share matrix factors, i.e., cluster indicator matrices. The resulting spectral relational clustering algorithm iteratively updates the cluster indicator matrices using the leading eigenvectors of a specially designed matrix until convergence. More recently, the collective matrix factorization proposed by Singh and Gordon [SG08a] [SG08b] can also be used for clustering k-partite graphs.

Other related work includes (1) GraphScope [SFPY07], which uses a similar information-theoretic criterion as cross-association for time-evolving graphs to segment time into homogeneous intervals; and (2) multi-way distributional clustering (MDC) [BEYM05] which was demonstrated to outper-

form the previous information-theoretic clustering algorithms at the time the algorithm was proposed.

At first glance, one may apply co-clustering algorithms to simultaneously address the problem of rare category selection and feature selection in unsupervised rare category analysis. The problem here is similar to the one mentioned in Section 2.4. That is, due to the extreme skewness of the class proportions and the overlapping support regions, general-purpose co-clustering algorithms may not be able to correctly identify the few rare examples or the features relevant to the rare categories; whereas we propose an algorithm for co-selecting the rare examples and the relevant features, which addresses this problem by making use of the clustering property of the minority classes.

## 2.6 Unsupervised Feature Selection

Generally speaking, existing feature selection methods in the unsupervised setting can be categorized as wrapper models and filter models. The wrapper models evaluate feature subsets based on the clustering results, such as the FSSEM algorithm [DB00], the mixture-based approach which extends to the unsupervised context the mutual-information based criterion [LJF02], and the ELSA algorithm [KSM00]. The filter models are independent of the clustering algorithm, such as the feature selection algorithm based on maximum information compression index [MMP02], the feature selection method using distance-based entropy [DCSL02], and the feature selection method based on Laplacian score [HCN05].

In unsupervised rare category analysis, one of the problems we want to address is feature selection, i.e., selecting a set of features relevant to the minority classes. In our settings, since the class proportions are very skewed, the general-purpose wrapper and filter methods would fail by selecting the features primarily relevant to the majority classes. Therefore, we need new feature selection methods that are tailored for rare category analysis.

## 2.7 Rare Category Detection

In rare category detection, the goal is to find at least one example from each minority class with the help of a labeling oracle, minimizing the number of label requests. Researchers have developed several methods for rare category detection. For example, in [PM04], the authors assumed a mixture model to fit the data, and experimented with different hint selection methods, of which

Interleaving performs the best; in [FM06], the authors studied functions with multiple output values, and used active sampling to identify an example for each of the possible output values; in [DH08], the authors presented an active learning scheme that exploits cluster structure in the data, which was proven to be effective in rare category detection; and in [VW09], the authors proposed a new approach to rare category detection based on hierarchical mean shift, where a hierarchy is created by repeatedly applying mean shift with an increasing bandwidth on the data. Different from most existing work on rare category detection, which assumes that the majority and minority classes are separable / near-separable from each other in the feature space, in Chapter 3 of this book, we target the more challenging cases where the support regions of different classes are not separable. Furthermore, besides empirical evaluations of the proposed algorithms, we also prove their effectiveness theoretically; whereas most existing algorithms do not have such guarantees.



Analysis of Rare Categories

He, J.

2012, VIII, 136 p.,

ISBN: 978-3-642-22813-1