Rare Category Detection on $O(dN)$ Time Complexity

Zhenguang Liu[1], Hao Huang[2], Qinming He[1], Kevin Chiew[3], and Lianhang Ma[1]

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
{zhenguangliu,hqm,lhma}@zju.edu.cn
[2] School of Computing, National University of Singapore, Singapore
huanghao@comp.nus.edu.sg
[3] Provident Technology Pte. Ltd., Singapore
kev.chiew@gmail.com

Abstract. Rare category detection (RCD) aims at finding out at least one data example of each rare category in an unlabeled data set with the help of a labeling oracle to prove the existence of such a rare category. Various approaches have been proposed for RCD with quadratic or even cubic time complexity. In this paper, by using histogram density estimation and wavelet analysis, we propose FRED algorithm and its prior-free version iFRED algorithm for RCD, both of which achieve linear time complexity w.r.t. either the data set size N or the data dimension d. Theoretical analysis guarantees its effectiveness, and comprehensive experiments on both synthetic and real data sets verify the effectiveness and efficiency of our algorithms.

Keywords: Rare category detection, wavelet analysis, linear time complexity.

1 Introduction

Emerging from anomaly detection, rare category detection (in short as RCD henceforce) [8,9,15] is proposed to figure out which rare categories exist in an unlabeled data set with the help of a labeling oracle. Different from imbalanced clustering or classification [3], RCD verifies the existence of a rare category by finding out at least one data example of this category. This work has a wealth of potential applications such as network intrusion detection [13], financial security [2], and scientific experiments [15].

Generally, RCD is carried out by two phases [11], i.e., (1) analyzing characteristics of data examples in a data set and picking out candidate examples with rare category characteristics such as compactness [4, 5, 7, 11, 12, 17] and isolation [11, 17], followed by (2) querying the category labels of these candidate examples to a labeling oracle (e.g., a human expert). The first phase involves processing a big amount of data, facing an *efficiency challenge* which aims to achieve low time complexity; while the second phase involves limited labeling budget, leading to a *query challenge* which aims to find out at least one data example for each rare category with as less queries as possible.

Most of the existing approaches (e.g., see [4, 12, 15, 17]) focus on query challenge without addressing too much of efficiency challenge with a time complexity not less than $O(dN^{2-\frac{1}{d}})$. To address both query and efficiency challenges simultaneously, we propose FRED (**F**ast **R**are cat**E**gory **D**etection) algorithm and its prior-free version

V.S. Tseng et al. (Eds.): PAKDD 2014, Part II, LNAI 8444, pp. 498–509, 2014.

Table 1. Algorithms' Time Complexity and Used Prior Information

Algorithm	Complexity	Category	Prior Info Used
Interleave	$O(dN^2)$	Prior Dependent	m
NNDM	$O(dN^{2-\frac{1}{d}})$	Prior Dependent	m, p_1, p_2, ..., p_r
SEDER	$O(d^2N^2)$	Prior Free	–
GRADE	$O(dN^3)$	Prior Dependent	m, p_1, p_2, ..., p_r
GRADE-LI	$O(dN^3)$	Prior Dependent	p_{max}
RADAR	$O(dN^2)$	Prior Dependent	m, p_1, p_2, ..., p_r
CLOVER	$O(dN^{2-\frac{1}{d}})$	Prior Free	–
HMS	$\Omega(dN^2)$	Prior Free	–

iFRED algorithm on $O(dN)$ time complexity which is linear w.r.t. either d or N. This is done by utilizing Histogram Density Estimation (HDE) to estimate local data density and identifying candidate data examples of rare categories through the abrupt density changes via wavelet analysis. On the other hand, the existing RCD approaches [11, 15, 17] are often based on the assumptions of isolation and compactness of rare category examples; in contrast, our algorithms do not require rare categories being isolated from majority categories, and relax the compactness assumption to that every rare category may only be compact on partial dimensions.

2 Related Work

The existing paradigms for RCD can be classified into three groups, namely (1) the mixture model-based [15], (2) the data distribution change-based [4, 5, 7, 8, 11, 12] and (3) the hierarchical clustering-based [17]. A brief review on some representatives of these approaches in terms of time complexity and required prior information about a given data set is shown in Table 1. Note that throughout the paper m stands for the number of all categories, r the number of rare categories, and p_i (where $1 \leqslant i \leqslant r$) the proportion of data examples of rare category R_i out of all data examples in a data set.

Mixture model-based algorithms assume that data examples are generated by a mixture data model and need to iteratively update the model, the computation cost is usually substantial. For example, Interleave algorithm [15] takes $O(dN^2)$ time to update the covariance for each mixture Gaussian.

Data distribution change-based algorithms select data examples with maximal data distribution changes as candidate examples of rare categories. According to the measurements for the data distribution changes, these algorithms can be classified into two sub-groups, namely (1) local density-based, such as SEDER [5], GRADE [7], GRADE-LI [7], and NNDM [4]; and (2) nearest neighborhood-based, such as RADAR [12] and CLOVER [11]. Their time complexities are nearly quadratic or even cubic.

Hierarchical clustering-based algorithms investigate rare category characteristics of clusters on various levels. HMS [17] as a representative uses Mean Shift with increasing bandwidths to create a cluster hierarchy, and adopts *Compactness* and *Isolation* criteria to measure rare category characteristics. Its overall time complexity is $\Omega(dN^2)$.

Besides time complexity, the prior information needed on a given data set leads to the existing algorithms falling into two classes, namely prior-dependent and prior-free.

3 Problem Statement and Assumptions

Adhering to the problem definition by He *et al*. [4, 5, 7] and Huang *et al*. [11, 12], we formally define the problem of rare category detection as follows.

Given: (1) An unlabeled data set $S = \{x_1, x_2, \ldots, x_N\}$ containing m categories; (2) a labeling oracle which is able to give category label for any data example.

Find: At least one data example for each category.

For the data distribution of majority categories, we have the following assumption which is commonly used in the existing work [4, 5, 7, 11, 12, 15] explicitly or implicitly.

Assumption 1. *Data distribution of each majority category is locally smooth on each dimension.*

Gaussian, Poisson, t-, uniform distribution and many other distributions well satisfy this assumption. Thus this assumption can be satisfied by most applications.

For data distribution of rare categories, the exiting work [4–7, 10–12, 15] assumes that each rare category forms a compact cluster in the whole feature space, i.e., data examples from a rare category are similar to each other on every dimension. We relax this assumption to that every rare category may only be compact on partial dimensions.

Assumption 2. *Each rare category forms a compact cluster on partial dimensions or on the whole feature space.*

This assumption is more realistic because in many applications, data examples from a rare category are different from those from a majority category on partial dimensions. For example, panda subspecies are different from giant panda only in fur color and tooth size. According to this assumption, data examples of each rare category should show cohesiveness and form a compact cluster on at least partial dimensions.

Let D_{R_i} be the dimensions such that on each dimension $j \in D_{R_i}$ rare category R_i forms a compact cluster. According to the assumptions, we have following observations.

Observation 1. *In the areas without clusters of rare category examples, data distribution is smooth on each dimension.*

According to Assumption 1, data distribution of each majority category is smooth on each dimension. Due to the additivity of continuous functions, even in the overlapped areas of different majority categories, data distribution is smooth on each dimension. Thus for simplicity, we can assume that there is one majority category R_0 in S.

Observation 2. *Any abrupt change of local data density on each dimension $j \in D_{R_i}$ indicates the presence of rare category R_i.*

According to Assumption 2, data examples of R_i form a compact cluster on dimension $j \in D_{R_i}$, thus the local data density of R_i on dimension j is significant. This significant data density, combining with overlaps of data examples from majority category R_0, brings an abrupt change in local data distribution, which is distinct from the smooth distribution of R_0. Therefore, abrupt changes of local data density on dimension $j \in D_{R_i}$ indicate the presence of rare category R_i.

4 RCD Algorithm via Local Density Change

Based on the observations, we present FRED algorithm for RCD by exploring these abrupt local density changes via three steps. (1) On each dimension of a data set, FRED tabulates data examples into bins of appropriate bandwidth, and estimates the local density of each bin by Histogram Density Estimation (HDE) [16]. (2) By conducting wavelet analysis on estimated density function, FRED locates abrupt changes of local data density and quantitatively evaluates the change rates via our proposed DCR criterion. (3) After summing up each data examples' weighted DCR scores on all dimensions, FRED keeps selecting data examples with maximal DCR scores for labeling until at least one data example is discovered for each category.

4.1 Histogram Density Estimation

To find abrupt changes of local data density, a crucial step is to estimate local data density. We adopt HDE [16] for this goal due to its accuracy and time efficiency.

HDE firstly tabulates the feature space of a single dimension within interval $[s1, s2]$ into w non-overlapped bins $B_1, B_2, \ldots, B_w$, which have the same bandwidth h, and uses the number of data examples in each bin to estimate the local data density.

Let υ_k be the number of data examples in the kth bin B_k and $\hat{f}(k)$ be the estimated local data density at bin k. Then we have

$$\hat{f}(k) = \frac{\upsilon_k}{N * h}, \quad k = 1, 2, ..., w \tag{1}$$

The structure of the histogram is completely determined by two parameters, bandwidth h and bin origin t_0. Well established theories (e.g., [16], [18]) show that bandwidth h has dominant effect and bin origin t_0 is negligible for sufficiently large sample sizes. A very small bandwidth results in a jagged histogram with each distinct observation lying in a separate bin (under-smoothed histogram); and a very large bandwidth results in a histogram with a single bin (over-smoothed histogram) [18]. We propose a criterion on h selection for detecting rare category R_i as

$$|avg(\upsilon_k) - C_i| \leqslant \varepsilon, \quad k = 1, 2, ..., w \tag{2}$$

where C_i is the number of data examples of R_i and ε a relaxation factor. This criterion guarantees that the average bin count is approximate to C_i, which makes the abrupt density change caused by R_i more significant to be detected.

4.2 Wavelet Analysis

After estimating local density, we perform wavelet analysis on the estimated density function to find abrupt density changes, which is the key to detecting rare categories.

First, we provide a brief review of main concepts on wavelet analysis. We define a mother wavelet as a square integrable function $\psi(x)$ that satisfies (1) $\psi(x)$ has a compact support, i.e., $\psi(x)$ has values in a small range and zeros otherwise; (2) $\psi(x)$ is

normalized, i.e., $\int_{-\infty}^{+\infty} \psi(x)\psi^*(x)dx = 1$, where * denotes complex conjugate; and (3) $\psi(x)$ is zero mean, i.e., $\int_{-\infty}^{+\infty} \psi(x)dx = 0$.

A wavelet family can be obtained by translating and scaling the mother wavelet. Mathematically, they are $\psi_{a,b}(x) = \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right)$ for $a, b \in \mathbb{R}$ and $a > 0$ where a is the scale, which is inversely proportional to the frequency, and b represents the translation, which indicates the point of location where we concern [14].

Given these, we define the wavelet analysis of a quadratic integrable function $f(x)$ with real-valued wavelet ψ as

$$WT_f(a, b) = \int_{-\infty}^{+\infty} f(x)\psi_{a,b}(x)dx. \tag{3}$$

Note that (1) wavelet analysis maps a 1-D signal to a 2-D domain of scale (frequency) variable a and location variable b, which allow for location-frequency analysis. (2) For a fixed scale a_0 and translation b_0, $\psi_{a_0,b_0}(x)$ is the wavelet chosen, and $WT_f(a_0, b_0)$ is called the wavelet coefficient, which represents the resemblance index of $f(x)$ on neighborhood of b_0 to $\psi_{a_0,b_0}(x)$, where large coefficients correspond to strong resemblance [14]. Once an appropriate wavelet is chosen, $WT_f(a_0, b_0)$ reflects the amplitude of density change at the point of location b_0. As mentioned above, identifying local density changes is the key to detecting rare categories, thus this amplitude can help us fast locate the location of rare categories.

4.3 Data Distribution Change Rate

To quantify local density change rate for bins, we propose a new criterion defined as

Definition 1 (Data distribution change rate (DCR)). *Given bin density function $\hat{f}$, wavelet basis ψ, scale a, the central point b_0 of bin B, DCR of bin B is defined as*

$$DCR(B) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} \hat{f}(x)\psi(\frac{x - b_0}{a})dx \tag{4}$$

DCR of each bin is calculated by wavelet analysis on $\hat{f}$. In practice, either Mexican hat or Reverse biorthogonal 2.2 (in short as Rbio2.2) wavelet can be chosen as wavelet basis ψ because they are similar in shape as cusps of density function brought by rare categories and have a compact support. Scale a in Eq. (4) is usually set to a positive value smaller than 1, which is the result of balancing the bandwidth of local region and computing cost.

Given DCR definition for bins, DCR of each data example on dimension j can be calculated by four steps. (1) Calculate the optimal bandwidth h by Eq. (2). (2) Divide the feature space of dimension j into bins and calculate bin density function by Eq. (1). (3) Compute DCR score of each bin by Definition 1, negative DCR scores are set to 0 because negative scores indicate drop of local data density and are of no interests to us here. (4) Perform K-means clustering on each bin with $K = v$ as a parameter. Let $x_1, x_2, \ldots, x_K$ be the central data examples of K clusters in bin B, then DCR scores of $x_1, x_2, \ldots, x_K$ are set to the DCR score of B, DCR scores of other data examples in bin B are set to zero.

Algorithm 1. Fast Rare Category Detection Algorithm (FRED)

Input: $S = \{x_k | 1 \leqslant k \leqslant N\}$, proportions of rare categories $p_1, p_2, \ldots, p_r$, dimension of data examples d, number of categories m

Output: The set Q of queried data examples and the set L of their labels

1 Initialize $Q = \varnothing, L = \varnothing$;
2 **for** $i = 1 : m$ **do**
3 **if** *category i is a majority category* **then**
4 $C_i = N * \max(p_l), 1 \leqslant l \leqslant r$;
5 **else**
6 $C_i = N * p_i$;
7 Set DCR of $\forall x_k \in S$ (denoted as $DCR(x_k)$) to 0;
8 **for** $j = 1 : d$ **do**
9 Calculate DCR of $\forall x_k \in S$ on dimension j (denoted as $DCR_j(x_k)$) by running the four steps introduced in Sec. 4.3;
10 calculate DCR of $\forall x_k \in S$, namely $DCR(x_k) = \sum_{j=1}^{d} W_j DCR_j(x_k)$;
11 Set the DCR of $\forall q \in Q$ to $-\infty$;
12 **while** $\max_{x_k \in S}(DCR(x_k)) > 0$ **do**
13 Query $s = \arg\max_{x_k \in S}(DCR(x_k))$ for its category label ℓ;
14 $Q = Q \cup s, L = L \cup \ell$;
15 **if** *s belong to an undiscovered category* **then**
16 break;
17 Set the DCR of s to $-\infty$;

4.4 FRED Algorithm

Algorithm 1 presents FRED algorithm which works as follows. Given proportions of rare categories, data dimension d, and the number of categories m, we first initialize hints set Q and their label set L to empty (line 1). Then for each rare category, (1) we compute the count of data examples C_i (lines 3–6), which will be used in the h selection step of DCR score calculation. (2) Then we calculate DCR score of $\forall x_k \in S$ on each dimension (lines 8–9). (3) For $\forall x_k \in S$, we sum up its weighed DCR score on each dimension as its final DCR score (line 10). It is recommended that $W_1, W_2, \ldots$, and W_k have the same value, whereas users with domain knowledge can modify them. (4) Next, we keep proposing the data example with maximal DCR score to the labeling oracle until a new category is found (lines 12–17). Note that DCR scores of selected data examples are set to $-\infty$ (lines 11 & 17) to prevent them from being chosen twice.

The time complexity of FRED consists of two parts, (1) DCR score computation and (2) sampling. (1) In DCR computation on each dimension, the most time consuming step is K-means clustering, which takes $O(N)$ time complexity. Note that on each dimension the time complexity of HDE is $O(N)$ and the time complexity of wavelet analysis is $O(w)$, where w is the number of bins and $w < N$. So the overall time complexity of DCR score computation is $O(dN)$. (2) Since one data example will never be selected twice according to Algorithm 1, the time complexity of sampling is $O(N)$. Thus the time complexity of FRED is $O(dN)$ which is linear w.r.t. either d or N.

Algorithm 2. Prior-free Rare Category Detection Algorithm (iFRED)

Input: $S = \{x_k | 1 \leqslant k \leqslant N\}$, sample size u, parameter β, ϵ
Output: The set Q of queried instances and the set L of their labels

1 Initialize $Q = \varnothing, L = \varnothing$;
2 $scale = 1$;
3 **for** $j = 1 : d$ **do**
4 Find bandwidth h_j on dimension j by Cross Validation;
5 **while** *labeling budget is not exhausted* ***and*** $scale > \epsilon$ **do**
6 Set DCR of $\forall x_k \in S$ to 0;
7 **for** $j = 1 : d$ **do**
8 Calculate DCR of $\forall x_k \in S$ on dimension j (denoted as $DCR_j(x_k)$) with bandwidth h_j;
9 calculate DCR of $\forall x_k \in S$, namely $DCR(x_k) = \sum_{j=1}^{d} W_j DCR_j(x_k)$;
10 Set the DCR of $\forall q \in Q$ to $-\infty$;
11 **while** $\max_{x_k \in S}(DCR(x_k)) > 0$ *and labeling budget is not exhausted* **do**
12 Query u data examples (denoted as set U) that have the maximum DCR scores for their category labels L_U;
13 $Q = Q \cup U, L = L \cup L_U$;
14 **if** U *all belong to discovered categories* **then**
15 $h_j = h_j * \beta, 1 \leqslant j \leqslant d$;
16 $scale = scale * \beta$;
17 break;
18 Set the DCR of each data example in U to $-\infty$;

5 iFRED Algorithm

We propose iFRED algorithm as a prior-free version of FRED for scenarios where no prior knowledge about the given data set is available.

The difference between iFRED and FRED is twofold. (1) For bandwidth h selection, iFRED algorithm cannot follow the criterion introduced in Eq. (2) because the number of data examples C_i in each rare category is not available. Instead, it uses Cross Validation [16] to find the original bandwidth h. Furthermore, if current h is not efficient in finding rare categories, iFRED reduces h by setting $h = h * \beta$, $0 < \beta < 1$. (2) In sampling phase, iFRED does not choose one data example each time for labeling, instead, each time it picks up u ($u \in \mathbb{N}^*$) data examples to measure the efficiency of current h in detecting rare categories. If at least one of the u data examples belongs to a new category, then current h is efficient; otherwise it sets $h = h * \beta$, $0 < \beta < 1$.

Algorithm 2 presents iFRED algorithm which works as follows. (1) The initialization phase (lines 1–4) initializes hints set Q and their label set L to empty, $scale$ is initialized to 1 and bandwidth on each dimension is initialized by Cross Validation. (2) The computation phase (lines 6–9) calculates DCR of each data example. (3) The sampling phase (lines 10–18) chooses each time u data examples of maximum DCR for labeling. If at least one of the u selected data examples belongs to a new category, we continue the sampling loop; otherwise we break out from the sampling loop, update

h_j ($1 \leqslant j \leqslant d$) by setting $h_j = h_j * \beta$ and then continue the loop of computation and sampling phase until the labeling budget is exhausted or $scale$ is too small (line 5, where ϵ is the threshold).

Time complexity of iFRED consists of two parts, (1) DCR score computation and (2) sampling. (1) In DCR computation, since the time complexity of Cross Validation on each dimension is $O(N)$ and the other three steps of DCR computation on each dimension takes $O(N)$ time complexity as analyzed in Sec. 4.4, the time complexity of DCR score computation is $O(dN)$. (2) Since one data example will never be selected twice according to Algorithm 2, the time complexity of sampling is $O(N)$. Thus the overall time complexity of iFRED is $O(dN)$ which is linear w.r.t. either d or N.

6 Effectiveness Analysis

In this section, we prove that if Assumptions 1 & 2 are fulfilled, our algorithms will sample repeatedly in the region where rare category examples occur with high probability. Without loss of generality, assume that we are searching for rare category $R_i, 1 \leqslant i \leqslant r$. Let B_{R_i} be the bins where data examples of R_i cluster together, D_{R_i} the dimensions that on each dimension $j \in D_{R_i}$ rare category R_i forms a compact cluster.

Claim 1. *According to the bandwidth selection criterion of FRED and iFRED, a cusp of bin density function will appear in B_{R_i} on each dimension j where $j \in D_{R_i}$.*

Proof. Since $\hat{f}(k) = \frac{v_k}{N*h}$ (see Eq. (1)), a sharp cusp of bin density function is equivalent to a cusp of bin count function. We prove from the following three points that a cusp of bin count function will appear in B_{R_i} on dimension $j \in D_{R_i}$.

(1) On dimension $j \in D_{R_i}$, the compact rare category examples of R_i will cluster together in the same bin or Q adjacent bins where Q is a small integer.

(2) Let γ_1 be the data distribution of majority categories at B_{R_i} and γ_2 be the data distribution of R_i at B_{R_i}. Then the bin count of B_{R_i} should be $\gamma_1 + \gamma_2$; whereas nearby bins without rare category examples have bin count of $\gamma_1 \pm \xi$. By Assumption 2, γ_2 is significant. Note that ξ is very small because data distribution of majority categories changes slowly according to Assumption 1, thus $\gamma_2 \gg \xi$.

(3) Let C_i be the number of data examples of R_i. For FRED, according to the h selection criterion (see Eq. (2)), $avg(v_k) \approx C_i$, thus γ_1 will not be too large than γ_2; for iFRED, the bandwidth h will keep reducing, resulting in smaller and smaller bins where γ_1 will not be too large than γ_2.

Therefore, bins with rare categories will have significantly higher bin counts than nearby bins without rare categories. Claim 1 is proven. ■

Claim 2. *According to DCR criterion, bins with cusp will get significantly high DCR scores while bins without cusp will get low DCR scores approximate to* 0.

Proof. Here we use Mexican hat wavelet (denoted by $\hat{\psi}(x)$) as an example wavelet to prove Claim 2 (wavelet Rbio2.2 can also be used in the same way). The shape of $\hat{\psi}_{a,b}(x) = \frac{1}{\sqrt{a}}\hat{\psi}\left(\frac{x-b}{a}\right)$ is shown in Fig. 1(a). Since the support interval of $\hat{\psi}(x)$ is $[-5, 5]$, thus the support interval of $\hat{\psi}_{a,b}(x)$ is $[-5a + b, 5a + b]$, outside which the

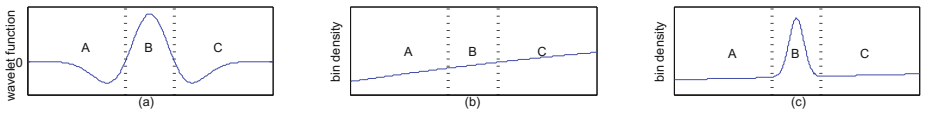

Fig. 1. Wavelet Analysis

value of $\hat{\psi}_{a,b}(x)$ is zero. Thus wavelet analysis of $f(x)$ by $\hat{\psi}_{a,b}(x)$ is $WT_f(a,b) = \int_{-5a+b}^{5a+b} f(x)\hat{\psi}_{a,b}(x)dx$.

Since FRED and iFRED use a positive fixed small scale a to detect local density changes, the integral interval $[-5a+b, 5a+b]$ is very narrow, which means that the data distribution change of majority categories is trivial. Fig. 1(b) shows the bin density function on the interval without cusps, and Fig. 1(c) shows the bin density function on the interval with one cusp. As shown in Fig. 1, wavelet analysis of $f(x)$ by $\hat{\psi}_{a,b}(x)$ is

$$WT_f(a,b) = \int_{A+B+C} f(x)\hat{\psi}_{a,b}(x)dx \tag{5}$$

For local areas without cusp, the bin density change on this interval is trivial. Thus

$$WT_f(a,b) \approx 0 \tag{6}$$

For local areas with cusps, Eq. (5) has a significant value because $\int_B f(x)\hat{\psi}_{a,b}(x)dx$ dominates the integration and has a significant value as shown in Fig. 1. Combining this conclusion with Eq. (6), we know that bins with cusps of bin density function will get a significantly high coefficients and bins without cusps will get coefficients approximate to zero. Therefore, Claim 2 is proven. ■

Claim 3. *In FRED and iFRED, representative data examples of rare category R_i where $1 \leqslant i \leqslant r$ will get significantly high DCR scores, whereas data examples with locally smooth data density will get low DCR scores approximate to* 0 .

Proof. From Claims 1 & 2, we know that B_{R_i} will get significantly high DCR score on each dimension j where $j \in D_{R_i}$. The significantly high dimensional DCR score of B_{R_i} will pass to representative data examples of R_i in the K-means clustering steps of FRED and iFRED. Since the DCR score of each data example is the sum of its weighted DCR scores on each dimension, representative data examples of R_i will have significantly high DCR scores; whereas according to Claim 2, bins with locally smooth data density will get low DCR scores approximate to 0, these low DCR scores will pass to representative data examples of these bins. Claim 3 is proven. ■

According to Assumption 1, the pdf of majority data examples is locally smooth. Combining this conclusion with Claim 3, we know that representative data examples of rare categories have significantly higher probabilities to be selected for labeling.

7 Experimental Evaluation

In this section, we conduct experiments to verify the efficiency and effectiveness of FRED and iFRED algorithms from two aspects, namely (1) time efficiency and scalability on data size N and dimension d, and (2) number of queries required for rare

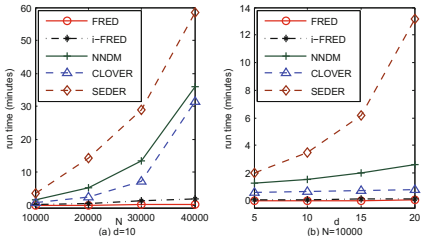

Fig. 2. Time Scalability Over N and d

Table 2. Properties of Real Data Sets

Data set	N	d	m	Largest class (%)	Smallest class (%)
Iris	106	4	3	47.17	5.66
Vertebral	310	6	3	48.39	19.35
Wine Quality	1589	11	5	42.86	1.13
Pen Digits	7143	16	10	10.92	5.15
Letter	19500	16	26	4.17	1.20
Shuttle	43494	9	6	78.42	0.03

category discovery. All algorithms are implemented with MATLAB 7.11 and running on a server computer with Intel Core 4 2.4GHz CPU and 20GB RAM.

7.1 Scalability

In this experiment, we compare our methods with NNDM, SEDER, and CLOVER on a synthetic data set where the pdf of majority categories is Gaussian and the pdf of rare categories is uniform within a small region. The synthetic data set satisfies that (1) the data size N ranges from 10000 to 40000 , (2) rare category R_1 forms a compact cluster in the densest area of the data set and has 395 data examples, (3) rare category R_2 forms a compact cluster in the moderate dense area and has 100 data examples, and (4) rare category R_3 forms a compact cluster in the low dense area and has 157 data examples.

(1) We set data dimension to 10 and vary the data size from 10000 to 40000 with incremental 10000. Fig. 2(a) shows the comparison results which agrees well with the time complexities shown in Table 1; i.e., the SEDER curve raises steeply due to its $O(d^2N^2)$ time complexity, followed by the curves of NNDM and CLOVER with $O(dN^{2-\frac{1}{d}})$ complexity, and lastly the curves of FRED and iFRED with linear complexity w.r.t. N.

(2) We set the size of the data set to 10000 and vary the data dimension from 5 to 20 with incremental 5. Fig. 2(b) shows that (1) iFRED, NNDM and CLOVER consumes much less time than SEDER; (2) time consumption by FRED, iFRED, NNDM and CLOVER grows linearly with data dimension.

(3) Our algorithms are much more efficient than other tested algorithms, e.g., on the data set with data size 40000, the runtime of CLOVER in seconds is 1884, NNDM 2153, SEDER 3499, whereas FRED only needs 5 seconds and iFRED 114 seconds. Experiments on real data sets such as Shuttle and Letter [1] have similar observations.

7.2 Efficiency

In RCD, the efficiency of an algorithm is evaluated by the number of queries needed to discover all categories in a data set. Usually these queries involve expensive human experts' work, thus our goal is to discover each category with minimal number of queries.

This experiment compares our algorithms with other algorithms on six real data sets from the UCI data repository [1], as detailed in Table 2. Specifically, categories with too

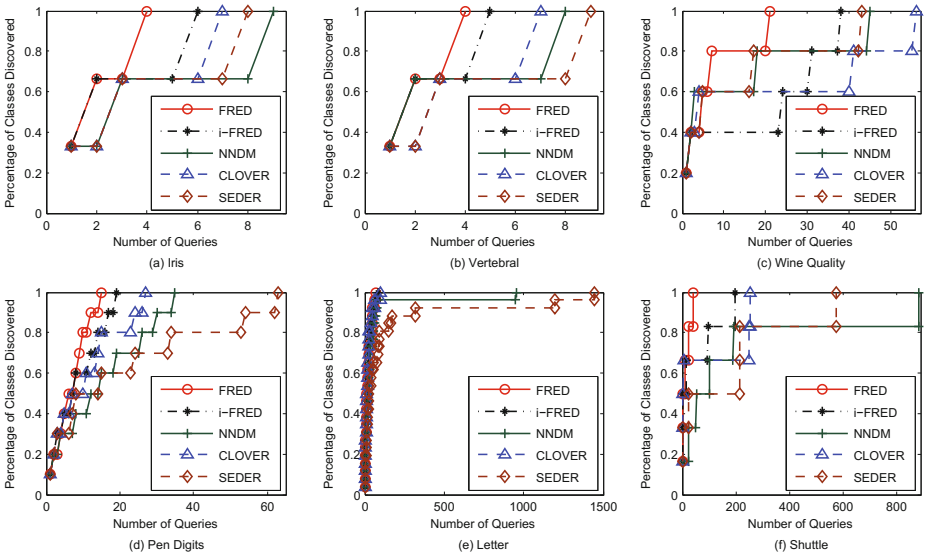

Fig. 3. Performance Comparisons on Real Data Sets

few ($\leqslant 10$) data examples are removed in order to satisfy compactness characteristic of rare categories, and sets Iris, Pen Digits and Letter are sub-sampled to create skewed data sets. Parameter v is set to either 1 or 2 for both FRED and iFRED.

Fig. 3 illustrates the comparison results on six data sets. From the figure, we have the following observations. (1) On all six data sets, FRED requires the least queries to discover all categories among all tested algorithms and performs significantly better than other algorithms. For example, on Shuttle set (Fig. 3(f)), FRED needs 42 labeling queries, iFRED 197, CLOVER 254, SEDER 575, and NNDM 884. (2) On all six data sets, iFRED takes up the second place right after FRED, especially on Vertebral, Pen Digits and Letter sets, its query number is almost as less as that of FRED. (3) A steeper curve means that the algorithm can discover new categories with fewer labeling queries. The curves of FRED and iFRED are much steeper than those of others, meaning that they need much less queries to discover a new category.

8 Conclusion

In this paper, we have proposed FRED and iFRED algorithms for RCD which have achieved $O(dN)$ time complexity and required least labeling queries. After using HDE to estimate local data density, they have been able to effectively identify candidate examples of rare categories through the abrupt density changes via wavelet analysis. Theoretical analysis has proven the effectiveness of our algorithms, and comprehensive experiments have further verified the efficiency and effectiveness.

For the next stage of study, a promising direction is to investigate new methods for the estimation of local data density. Another suggestion for future study is to work on

sub-space selection which may bring with breakthroughs on this topic since in many scenarios rare categories are distinct on only a few dimensions.

Acknowledgment. This work is supported by the National Key Technology R&D Program of the Chinese Ministry of Science and Technology under Grant No. 2012BAH94F03.

References

1. Asuncion, A., Newman, D.: UCI Machine Learning Repository (2007)
2. Bay, S., Kumaraswamy, K., Anderle, M., Kumar, R., Steier, D.: Large scale detection of irregularities in accounting data. In: ICDM, Hong Kong, China, pp. 75–86 (2006)
3. Garcia-Pedrajas, N., Garcia-Osorio, C.: Boosting for class-imbalanced datasets using genetically evolved supervised non-linear projections. Progress in AI 2(1), 29–44 (2013)
4. He, J., Carbonell, J.: Nearest-neighbor-based active learning for rare category detection. In: NIPS 2007, Vancouver, British Columbia, Canada, December 3-6, pp. 633–640 (2007)
5. He, J., Carbonell, J.: Prior-free rare category detection. In: SDM 2009, Sparks, Nevada, USA, April 30-May 2, pp. 155–163 (2009)
6. He, J., Carbonell, J.: Coselection of features and instances for unsupervised rare category analysis. Statistical Analysis and Data Mining 3(6), 417–430 (2010)
7. He, J., Liu, Y., Lawrence, R.: Graph-based rare category detection. In: ICDM 2008, Pisa, Italy, December 15-19, pp. 833–838 (2008)
8. He, J., Tong, H., Carbonell, J.: An effective framework for characterizing rare categories. Frontiers of Computer Science 6(2), 154–165 (2012)
9. Hospedales, T.M., Gong, S., Xiang, T.: Finding rare classes: Active learning with generative and discriminative models. TKDE 25(2), 374–386 (2013)
10. Huang, H., Chiew, K., Gao, Y., He, Q., Li, Q.: Rare category exploration. Expert Systems with Applications (2014)
11. Huang, H., He, Q., Chiew, K., Qian, F., Ma, L.: Clover: A faster prior-free approach to rare category detection. Knowledge and Information Systems 35(3), 713–736 (2013)
12. Huang, H., He, Q., He, J., Ma, L.: Radar: Rare category detection via computation of boundary degree. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011, Part II. LNCS (LNAI), vol. 6635, pp. 258–269. Springer, Heidelberg (2011)
13. Khor, K., Ting, C., Phon-Amnuaisuk, S.: A cascaded classifier approach for improving detection rates on rare attack categories in network intrusion detection. Applied Intelligence 36(2), 320–329 (2012)
14. Nenadic, Z., Burdick, J.: Spike detection using the continuous wavelet transform. IEEE Transactions on Biomedical Engineering 52(1), 74–87 (2005)
15. Pelleg, D., Moore, A.: Active learning for anomaly and rare-category detection. In: NIPS 2004, Vancouver, British Columbia, Canada, December 13-18, pp. 1073–1080 (2004)
16. Scott, D.W.: Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley, New York (1992)
17. Vatturi, P., Wong, W.: Category detection using hierarchical mean shift. In: KDD 2009, Paris, France, June 28-July 1, pp. 847–856 (2009)
18. Wand, M.P.: Data-based choice of histogram bin width. The American Statistician 51(1), 59–64 (1997)