Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

MUVIR: Multi-View Rare Category Detection

Dawei Zhou, Jingrui He, K. Seluk Candan, Hasan Davulcu
Arizona State University
Tempe, Arizona
{dzhou23,jingrui.he,candan,hdavulcu}@asu.edu

Abstract

Rare category detection refers to the problem of identifying the initial examples from under-represented minority classes in an imbalanced data set. This problem becomes more challenging in many real applications where the data comes from multiple views, and some views may be irrelevant for distinguishing between majority and minority classes, such as synthetic ID detection and insider threat detection. Existing techniques for rare category detection are not best suited for such applications, as they mainly focus on data with a single view.

To address the problem of multi-view rare category detection, in this paper, we propose a novel framework named *MUVIR*. It builds upon existing techniques for rare category detection with each single view, and exploits the relationship among multiple views to estimate the overall probability of each example belonging to the minority class. In particular, we study multiple special cases of the framework with respect to their working conditions, and analyze the performance of *MUVIR* in the presence of irrelevant views. For problems where the exact priors of the minority classes are unknown, we generalize the *MUVIR* algorithm to work with only an upper bound on the priors. Experimental results on both synthetic and real data sets demonstrate the effectiveness of the proposed framework, especially in the presence of irrelevant views.

1 Introduction

In contrast to the large amount of data being generated and used everyday in a variety of areas, it is usually the case that only a small percentage of the data might be of interest to us, which form the minority class. However, without initial labeled examples, the minority class might be very difficult to detect with random sampling due to the imbalance nature of the data, and the limited budget for requesting labels from a labeling oracle. Rare category detection has been proposed to address this problem, so that we are able to identify the very first examples from the minority class, by issuing a small number of label requests to the labeling oracle.

In many real-world applications, the data consists of multiple views, or features from multiple information sources. For example, in synthetic ID detection, we aim to distinguish between the true identities and the fake ones generated for the purpose of committing fraud. Each identity is associated with information from various aspects, such as demographic information, online social behaviors, banking behaviors. Another example is insider threat detection, where the goal is to detect malicious insiders in a large organization, by collecting various types of information regarding each employee's daily behaviors. To detect the rare categories in these applications, simply concatenating all the features from multiple views may lead to sub-optimal performance in terms of increased number of label requests, as it ignores the relationship among the multiple views. Furthermore, among the multiple information sources, some may generate features irrelevant to the identification of the rare examples, thus deteriorates the performance of rare category detection.

To address this problem, in this paper, we propose a novel framework named *MUVIR* for detecting the initial examples from the minority classes in the presence of multi-view data. The key idea is to integrate view-specific posterior probabilities of the example coming from the minority class given features from each view, in order to obtain the estimate of the overall posterior probability given features from all the views. In particular, the view-specific posterior probabilities can be inferred from the scores computed using a variety of existing techniques [He and Carbonell, 2007; He *et al.*, 2008]. Furthermore, *MUVIR* can be generalized to handle problems where the exact priors of the minority classes are unknown. To the best of our knowledge, this paper is the first principled effort on rare category detection in the presence of multiple views. Compared with existing techniques, the main advantages of *MUVIR* can be summarized as follows.

1. Effectively leveraging the relationship among multiple views to improve the performance of rare category detection;
2. Robustness to irrelevant views;
3. Flexibility in terms of the base algorithm used for generating view-specific posterior probabilities.

The rest of this paper is organized as follows. After a brief review of the related work in Section 2, we introduce the proposed framework for multi-view rare category detection in Section 3. In Section 4, we test our model on both synthetic data sets and real data sets. Finally, we conclude this paper in Section 5.

2 Related Work

Multi-view Learning

Multi-view learning targets problems where the features naturally come from multiple information sources, or multiple views. It has been studied extensively in the literature. Co-training [Blum and Mitchell, 1998] is one of the earliest efforts in this area, where the authors proved that maximizing the mutual consistency of two independent views could be used to learn the pattern based on a few labeled and many unlabeled examples. Since then, multi-view learning has been studied in multiple aspects during these years. A portion of the researchers focus on the study of independent assumption for co-training, which is essential in the real world application. [Abney, 2002] refined the analysis of co-training and gave a theoretical justification that their algorithm could work on a more relax independence scenario rather than co-training. [Balcan *et al.*, 2004] proposed an independence expansion and proved that it could guarantee the success of co-training. Another line of work has been devoted to the construction of multiple views and how to combine multiple views. In [Ho, 1998], they apply random sampling algorithm called RSM, which perform bootstrapping in the feature space to separate the views. [Chen *et al.*, 2011] transform the feature decomposition task into an optimization problem, which could automatically divide the feature space into two exclusive subsets. While, in the aspect of how to combine multiple views and learn models, we can separate it into the problems of supervised learning, semi-supervised learning and unsupervised learning. In the category of supervised and semi-supervised learning, [Muslea *et al.*, 2003; 2006] designed a robust semi-supervised algorithm which combined co-learning with active learning. CoMR [Sindhwani and Rosenberg, 2008] proposed a multi-view learning algorithm based on a reproducing kernel Hilbert space with a data-dependent co-regularization norm. In [Yu *et al.*, 2011], author proposed a co-training Bayesian graph model, which is more reliable in handling the case of missing views. SMVC [Günnemann *et al.*, 2014] proposed a Bayesian framework for modeling multiple clusterings of data by multiple mixture distributions. In the category of unsupervised learning, [Long *et al.*, 2008] introduced a general model for unsupervised multiple view learning and demonstrate it in various types of unsupervised learning on various types of multiple view data. The authors of [Song *et al.*, 2013] developed a kernel machine for learning in multi-view latent variable models, which also allows mixture components to be non-parametric and to learn data in an unsupervised fashion.

Different from existing work on multi-view learning, in this paper, we start *de-novo*, i.e., we do not have any labeled examples to start with, but we are able to query the oracle for the labels of selected examples until at least one example has been detected from each minority class.

Rare Category Detection

Rare category analysis has also been studied for years. Up to now, many methods have been approached to address this problem. In this paper, we mainly review the following two existing works on rare category detection. The first one is [He and Carbonell, 2007], in which algorithm NNDM is proposed standing on two assumptions: (i) *data sets have little knowledge about labels* (ii) *there is no separability or near-separability between majority and minority classes*. Both assumptions exactly meet the setting of the problem we want to figure out. The probability distribution function (pdf) of the majority class tends to be locally smooth, while the pdf of minority class tends to be a more compact cluster. In general, the algorithm measures the changes of local density around a certain point. NNDM gives a score to each example, and the score is the maximum difference of local density between one item and all of its neighboring points. By querying the examples with the largest score, it is able to hit the region of minority class with the largest probability.

Another work about rare category detection is [He *et al.*, 2008], the authors provided an upgraded algorithm GRADE based on NNDM. In this algorithm, they took the consideration of the manifold structure in minority class. For example, two examples from the same minority class on the manifold may be far away in Euclidean distance. In this case, they generate a global similarity matrix embedded all of the examples from the original feature space. The items of minority class are made to form a more compact cluster for each minority class. Based on global similarity matrix, they measure the changes of local density for each example. The changes of local density, to some extent, has been enlarged, and made the minority classes easier to be discovered. Furthermore, they provided an approximating algorithm to manage rare category detection with less information about priors of minority classes. In this paper, our proposed framework *MUVIR* is generic in the sense that it can leverage multiple existing RCD methods, such as GRADE, NNDM and etc., to analyze the problem in the multi-view version. To the best of our knowledge, this is the first effort on rare category detection with multiple views.

3 The Proposed Framework

In this section, we introduce the proposed framework *MUVIR* for multi-view rare category detection. Notice that similar as existing techniques designed to address this problem for single-view data, we target the more challenging setting where the support regions of the majority and minority classes overlap with each other, which makes *MUVIR* widely applicable to a variety of real problems.

3.1 Notation

Suppose that we are given a set of unlabeled examples $\mathcal{S} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, which come from m distinct classes, i.e. $y_i \in \{1, \cdots, m\}$. Without loss of generality, assume that $y_i = 1$ corresponds to the majority class with prior p^1, and the remaining classes are minority classes with prior p^c. Furthermore, each example $\boldsymbol{x}_i$ is described by features from V views, i.e., $\boldsymbol{x}_i = [(\boldsymbol{x}_i^1)^T, \dots, (\boldsymbol{x}_i^V)^T]^T$, where $\boldsymbol{x}_i^v \in \mathbb{R}^{d_v}$, and d_v is the dimensionality of the v^{th} view. In our proposed model, we repeatedly select examples to be labeled by an oracle, and the goal is to discover at leaset one example from each minority class by requesting as few labels as possible.

3.2 Multi-View Fusion

In this section, for the sake of exposition, we focus on the binary case, i.e., $m = 2$, and the minority class corresponds to $y_i = 2$, although the analysis can be generalized to multiple minority classes. As reviewed in Section 2, existing techniques for rare category detection with single-view data essentially compute the score for each example according to the change in the local density, and select the examples with the largest scores to be labeled by the oracle. Under mild conditions [He *et al.*, 2008; He and Carbonell, 2007], these scores reflect $P(\boldsymbol{x}, y = 2)$, thus are in proportion to the conditional probability $P(y = 2|\boldsymbol{x})$.

For data with multi-view features, running these algorithms [He *et al.*, 2008; He and Carbonell, 2007] on each view will generate scores in proportion to $P(y = 2|\boldsymbol{x}^v)$, $v = 1, \ldots, V$. Next, we establish the relationship between these probabilities and the overall probability $P(y = 2|\boldsymbol{x})$.

Theorem 1. *If the features from multiple views have weak dependence given the class label $y_i = 2$ [Abney, 2002], i.e., $P(\boldsymbol{x}|y = 2) \geq \alpha \prod_{v=1}^{V} P(\boldsymbol{x}^v|y = 2)$, $\alpha > 0$, then*

$$P(y = 2|\boldsymbol{x}) \geq C(\prod_{v=1}^{V} P(y = 2|\boldsymbol{x}^v)) \times \left(\frac{\prod_{v=1}^{V} P(\boldsymbol{x}^v)}{P(\boldsymbol{x})}\right) \tag{1}$$

where $C = \frac{\alpha}{(p^2)^{V-1}}$ is a constant.

Proof.

$$\begin{aligned} P\left(y = 2|\boldsymbol{x}\right) &= \frac{P(y = 2)P(\boldsymbol{x}|y = 2)}{P(\boldsymbol{x})} \\ &\geq \frac{P(y = 2)\alpha \prod_{v=1}^{V} P(\boldsymbol{x}^v|y = 2)}{P(\boldsymbol{x})} \\ &= \alpha \frac{P(y = 2) \prod_{v=1}^{V} \frac{P(y=2|\boldsymbol{x}^v)P(\boldsymbol{x}^v)}{P(y=2)}}{P(\boldsymbol{x})} \\ &= \alpha \frac{\prod_{v=1}^{V} P(y = 2|\boldsymbol{x}^v)P(\boldsymbol{x}^v)}{P(\boldsymbol{x})(P(y = 2))^{V-1}} \\ &= \frac{\alpha}{(p^2)^{V-1}} \prod_{v=1}^{V} P(y = 2|\boldsymbol{x}^v) \frac{\prod_{v=1}^{V} P(\boldsymbol{x}^v)}{P(\boldsymbol{x})} \end{aligned} \tag{2}$$

□

As a special case of Theorem 1, when the features from multiple view are conditionally independent given the class label, i.e., $\alpha = 1$, we have the following corollary.

Corollary 1. *If the features from multiple views are conditionally independent given the class label, then Inequality 1 becomes equality, and $C = \frac{1}{(p^2)^{V-1}}$.*

Proof. Notice that when the features from multiple views are conditionally independent given the class label, we have

$$P(\boldsymbol{x}|y = 2) = \prod_{v=1}^{V} P(\boldsymbol{x}^v|y = 2)$$

The rest of the proof follows by changing the inequality in Equation 2 to equality. □

Based on the above analysis, in *MUVIR*, we propose to assign the score for each example as follows.

$$s(\boldsymbol{x}) = \prod_{v=1}^{V} s^v(\boldsymbol{x}^v) \left(\frac{\prod_{v=1}^{V} P(\boldsymbol{x}^v)}{P(\boldsymbol{x})}\right)^d \tag{3}$$

where $s^v(\boldsymbol{x}^v)$ denotes the score obtained based on the v^{th} view using existing techniques such as NNDM [He and Carbonell, 2007] or GRADE [He *et al.*, 2008]; and $d \geq 0$ is a parameter that controls the impact of the term related to the marginal probability of the features. In particular, we would like to discuss two special cases of Equation 3.

Case 1. If the features from multiple views are conditionally independent given the class label, and they are marginally independent, i.e., $P(\boldsymbol{x}) = \prod_{v=1}^{V} P(\boldsymbol{x}^v)$, then Corollary 1 indicates that $d = 0$;

Case 2. If the features from multiple views are conditionally independent given the class label, then Corollary 1 indicates that $d = 1$.

In Section 4, we study the impact of the parameter d on the performance of *MUVIR*, and show that in general, $d \in (0, 1.5]$ will lead to reasonable performance.

Notice that the proposed score in Equation 3 is robust to irrelevant views in the data, i.e., the views where the examples from the majority and minority classes cannot be effectively distinguished. This is mainly due to the first part $\prod_{v=1}^{V} s^v(\boldsymbol{x}^v)$ on the right hand side of Equation 3. For example, assume that view 1 is irrelevant such that the distribution of the majority class ($P(\boldsymbol{x}|y = 1)$) is the same as the minority class ($P(\boldsymbol{x}|y = 2)$). In this case, the view-specific score $s^1(\boldsymbol{x}^1)$, which reflects the conditional probability $P(y = 2|\boldsymbol{x})$, would be the same for all the examples. Therefore, when integrated with the scores from the other relevant views, view 1 will not impact the *relative* score of all the examples, thus it will not degrade the performance of the proposed framework.

3.3 *MUVIR* Algorithm

The proposed *MUVIR* algorithm is described in Algorithm 1. It takes as input the multi-view data set, the priors of all the classes ($p^1, p^2, \ldots, p^m$), as well as some parameters, and outputs the set of selected examples together with their labels.

MUVIR works as follows. In Step 2, we compute the view-specific score for each example, which can be done using any existing techniques for rare category detection. In Step 3, we estimate the view-specific density using kernel density estimation; whereas in Step 5, we estimate the overall density by pooling the features from all the views together. Finally, Steps 6 to 16 aim to select candidates according to $P(y = c|\boldsymbol{x})$. To be specific, in Step 7, we skip class c if examples from this class have already been identified in the previous iterations. Step 10 implements the feedback loop by excluding any examples close to the labeled ones from being selected in future iterations. Notice that the threshold ϵ depends on the algorithm used to obtain the view-specific scores. For example, it is set to the smallest k-nearest neighbor distance in NNDM [He and Carbonell, 2007], and the largest k-nearest neighbor global similarity in GRADE [He *et al.*, 2008]. Step 11 updates the view-specific score for each example with enlarged neighborhood for computing the change in local density [He and Carbonell, 2007; He *et al.*, 2008]. In Step 13, we compute the overall score based on Equation 3, and select the example with the maximum overall score to be labeled by the oracle in Step 14. In Step 15, if the labeled example is from the target class in this iteration, we proceed to the next class; otherwise, we mark the class of this examples as labeled.

Algorithm 1 *MUVIR* Algorithm

Input: Unlabeled data set $\mathcal{S}$ with features from V views, $p^1, \ldots, p^m$, d, ϵ.
Output: The set I of selected examples and the set L of their labels.
1: **for** v=1 : V **do**
2: Compute the view-specific score $s^v(\boldsymbol{x}_i^v)$ for all the examples using existing techniques for rare category detection, such as GRADE [He *et al.*, 2008];
3: Estimate $P(\boldsymbol{x}_i^v)$ using kernel density estimation;
4: **end for**
5: Estimate $P(\boldsymbol{x}_i)$ using kernel density estimation on all the features combined;
6: **for** c=2 : m **do**
7: If class c has been discovered, continue;
8: **for** $t = 2 : n$ **do**
9: **for** $v = 1 : V$ **do**
10: For each $\boldsymbol{x}_i$ that has been labeled by the oracle, $\forall i, j = 1, \ldots, n, i \neq j,$, if $\|\boldsymbol{x}_i^v, \boldsymbol{x}_j^v\|_2 \leq \epsilon$, then $s^v(\boldsymbol{x}_j^v) = -\infty$;
11: Update the view-specific score $s^v(\boldsymbol{x}_i^v)$ using existing techniques such as GRADE [He *et al.*, 2008];
12: **end for**
13: Compute the overall score for each example $s(\boldsymbol{x}_i)$ based on Equation 3;
14: Query the label of the example with the maximum $s(\boldsymbol{x}_i)$
15: If the label of $\boldsymbol{x}_i$ is from class c, break; otherwise, mark the class of $\boldsymbol{x}_i$ as labeled.
16: **end for**
17: **end for**

3.4 *MUVIR* with Less Information (*MUVIR*-LI)

In many real applications, it may be difficult to obtain the priors of all the minority classes. Therefore, In this subsection, we introduce *MUVIR*-LI, a modified version of Algorithm 1, which replaces the requirement for the exact priors with an upper bound p for all minority classes. Compared with *MUVIR*, *MUVIR*-LI is more suitable in real world applications.

MUVIR-LI is described in Algorithm 2. It works as follows. Step 2 calculates the specific score s^v for each example. The only difference from *MUVIR* is that here we use upper bound p to calculate s^v, which is a less accurate measurement of changing local density than in *MUVIR*. The same as *MUVIR*, we estimate the view specific density and the overall density by applying kernel density estimation in Step 3 and Step 5. The while loop from Step 6 to Step 16 is the query processing. We calculate the overall score for each example and select the examples with the largest overall score to be labeled by oracle. We end the loop until all the classes has been discovered.

Algorithm 2 *MUVIR*-LI Algorithm

Input:
Unlabeled data set $\mathcal{S}$ with features from V views, p, d, ϵ.
Output:
The set I of selected examples and the set L of their labels.
1: **for** $v = 1 : V$ **do**
2: Compute the view-specific score $s^v(\boldsymbol{x}_i^v)$ for all the examples using existing techniques for rare category detection, such as GRADE-LI [He *et al.*, 2008];
3: Estimate $P(\boldsymbol{x}_i^v)$ using kernel density estimation;
4: **end for**;
5: Estimate $P(\boldsymbol{x}_i)$ using kernel density estimation;
6: **while** not all the classes have been discovered **do**
7: **for** $t = 2 : n$ **do**
8: **for** $v = 1 : V$ **do**
9: For each $\boldsymbol{x}_i$ that has been labeled by the oracle, $\forall i, j = 1, \ldots, n, i \neq j,$, if $\|\boldsymbol{x}_i^v, \boldsymbol{x}_j^v\|_2 \leq \epsilon$, then $s^v(\boldsymbol{x}_j^v) = -\infty$;
10: Update the view-specific score $s^v(\boldsymbol{x}_i^v)$ using existing techniques such as GRADE-LI [He *et al.*, 2008];
11: **end for**;
12: Compute the overall score for each example $s(\boldsymbol{x}_i)$ based on Equation 3;
13: Query the label of the example with the maximum $s(\boldsymbol{x}_i)$
14: Mark the class that $\boldsymbol{x}$ belongs to as discovered.
15: **end for**;
16: **end while**

4 Experimental Results

In this section, we will present the results of our algorithm on both synthetic data sets and real data sets in multiple special scenarios, such as data sets with different number of irrelevant features, data sets with multiple classes and data sets with very rare categories, such as class proportion of 0.02%.

4.1 Synthetic Data Sets

Binary Class Data Sets

For binary classes, we perform experiment on 3600 synthetic data sets, and each scenario has independent 100 data sets.

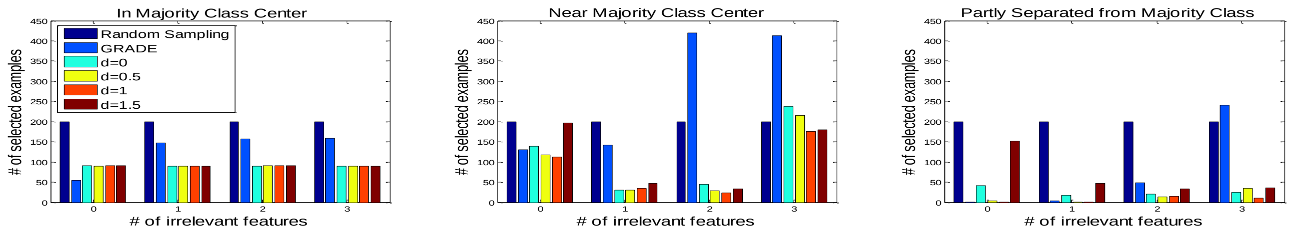

Figure 1: Prior of minority class is 0.5%

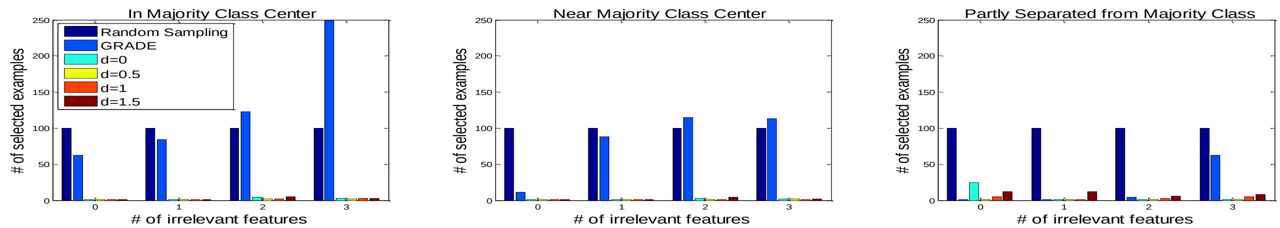

Figure 2: Prior of minority class is 1%

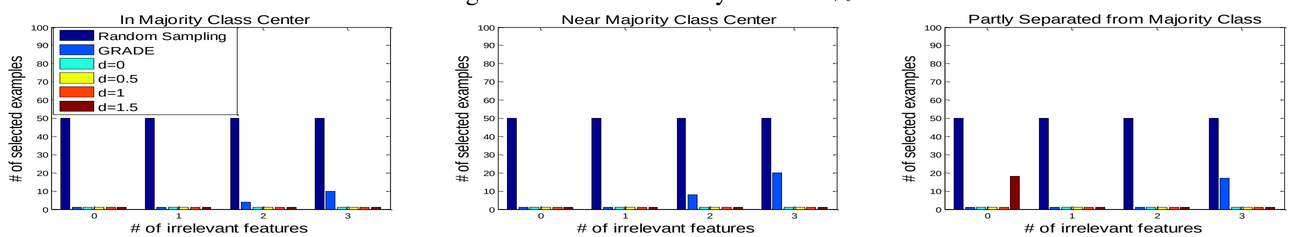

Figure 3: Prior of minority class is 2%

We consider the following three special conditions: (i) different number of irrelevant features, i.e. from 0 to 3 irrelevant features; (ii) different priors for minority class, i.e. 0.5%, 1%, 2%; (iii) different levels of correlation between majority class and minority class, ie. minority class stays in the center of majority class, minority class stays around the center of majority class, minority class stays at the boundary of majority class. Besides, as the distribution of majority class tends to be more scattered and the distribution of minority class is more compact, we set each data set with 5000 examples and $\sigma_{majority} : \sigma_{minority} = 40 : 1$.

In the experiment, we compare *MUVIR* with GRADE [He *et al.*, 2008] and random sampling. Fig. 1 shows the results when the prior of minority class is 0.5%. Using random sampling, we need to label 200 examples on average to identify the minority class. In most cases, other approaches outperform random sampling. However, the learning model generated by GRADE algorithm performs worse with the increasing of irrelevant features. In contrast, *MUVIR* is more efficient and stable rather GRADE. The experiment with minority proportions of 1% and 2% are represented in Fig. 2 and Fig. 3. In these two experiment, *MUVIR* outperforms GRADE and random sampling in each condition with any setting of d. Comparing these three figures, we have the following observations for binary class data sets: (i) *MUVIR* is more reliable especially when dealing with data sets containing irrelevant features. (ii) In the case of data sets with no irrelevant features, the performance of *MUVIR* with different values of d are roughly the same. (iii) In the case of data sets with irrelevant features, *MUVIR* with $d = 1$ outperforms other methods.

Multi-classes Data Sets with Imprecise Prior

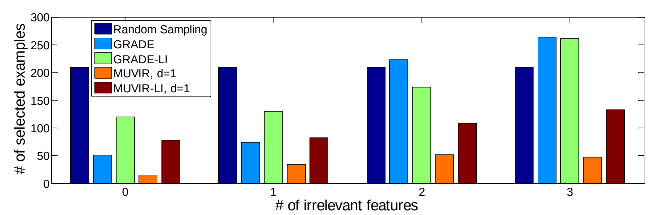

Figure 4: Multi-class data sets

For multi-class data sets, we compare the performances among different approaches. In particular, GRADE-LI [He *et al.*, 2008] and *MUVIR*-LI are only provided with an upper bound p on the proportion of all the minority classes. The multi-class data sets consisting of 9000 examples correspond to majority class, and the other 1000 examples correspond to 4 minority classes. The proportions of minority classes are 4%, 3%, 2%, 1%. Similar to previous experiments, we will discuss the scenario data sets contain different number of irrelevant features. Each value we represented in the figure is the median value of results from 100 same scenario data sets. From Fig. 4, we can have the following conclusions: (i) *MUVIR* outperforms all other algorithms in multi-class data sets; (ii) GRADE only performs good when data sets have 1 or 0 irrelevant feature; (iii) *MUVIR*-LI is more reliable than GRADE-LI in all scenarios. The reason that our models have better performance is that both *MUVIR* and *MUVIR*-LI are capable to exploit the relationship among multiple views and extract useful information to make predictions.

Parameter Analysis

From previous experiments, we found different parameter settings may result in different outcomes. In this experiment, we will focus on analyzing the impact from degree d and upper bound prior p. To measure the impact of these parameters, we generate 400 data sets with minority class proportion 1%. The number of irrelevant features varies from 0 to 3, and each case has 100 data sets. In Fig. 5, the X axis represents different values of degree d, and Y axis represents the number of selected examples on average. From Fig. 5, we can see that *MUVIR* performs better when $d \in (0, 1.5]$. In the following experiments, we will focus on studying the performance of our algorithm with d in this certain area.

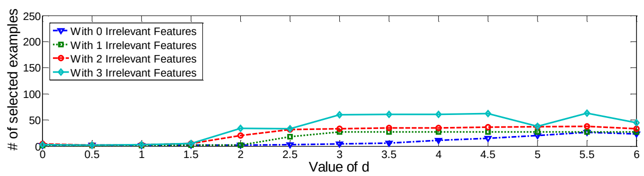

Figure 5: Learning curves with different degree d

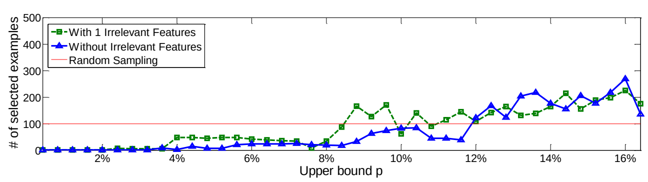

Figure 6: Learning curves with different prior upper bound

With the same data sets, we studied the learning curves of labeling requests by applying *MUVIR*-LI with different upper bound p. In Fig. 6, the X axis represents different values of upper bound proportion and Y axis represents the number of labeling requests. The red line represents the average number of labeling requests by using random sampling. When data sets without irrelevant features, *MUVIR*-LI works well even with upper bound p changing from 1% to 12%. When data sets with irrelevant features, *MUVIR*-LI can still outperforms random sampling with upper bound p changing from 1% to 8.5%. However, when the upper bound exceeds a certain level, the algorithm tends to be random sampling. This might be due to the reason that when the bound is very loose, e.g. the exact proportion of the minority class is 1% and the given upper bound is 10%, the performance of our proposed algorithm may be greatly affected by the introduced noise.

4.2 Real Data Sets

In this subsection, we will demonstrate our algorithm on two real data sets Statlog and Adult. Noted that, before we run our

Views	Features
relevant view 1	education, education years, work class
relevant view 2	age, hours per week, occupation
relevant view 3	martial status, relationship, sex
relevant view 4	race, native country
irrelevant view 1	final weight
irrelevant view 2	capital loss, capital gain

Table 1: Relevant and irrelevant views in Adult Data set.

algorithms, we have preprocessed both data sets in order to keep each feature component has mean 0 and standard deviation 1. In the following experiments, we will compare *MUVIR* and *MUVIR*-LI with the following algorithms: GRADE, GRADE-LI and random sampling.

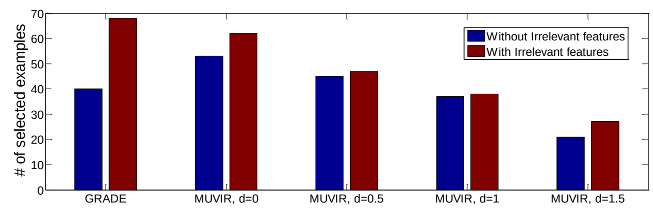

Figure 7: Adult

Adult data set contains 48842 instances and 14 features of each example. It is a binary classes data sets. Considering the original prior of minority class in data sets is around 24.93%. To better test the performance of our model, we keep majority class the same and down sample the minority class to 500 examples. In this way, we generate 24 data sets with minority prior of 1.3%. And we select relevant and irrelevant views based on correlation analysis. Noticed that all the views are fed to all the algorithms without information regarding their relevance. The details about relevant and irrelevant views are represented in Tab.1. Fig. 7 shows the comparison results on real data by applying 5 different approaches. In this experiment, we have not included *MUVIR*-LI, it is because *MUVIR*-LI is mainly developed for multi-class cases and Adult is a binary class data sets. By using random sampling, the average number of selected examples is 76. With irrelevant views, GRADE needs 69 requests, *MUVIR* with $d = 0$ needs 60 requests, *MUVIR* with $d \neq 0$ needs around 30 to 40 requests. The results totally meet our intuition that when dealing data sets with irrelevant views, *MUVIR* with $d \neq 0$ outperforms *MUVIR* with $d = 0$, and *MUVIR* with $d = 0$ outperforms GRADE. However, when dealing with data sets without irrelevant views, GRADE needs less labeling requests than *MUVIR* with $d = 0$, but more labeling requests than *MUVIR* with d around 1.

Different from Adult, Statlog contains 58000 examples and 7 classes. Among 7 classes, there are 6 minority classes, with priors varying from 0.02% to 15%. In this experiment, we compare the following 4 methods: GRADE, GRADE-LI with upper bound $p = \max_{c=2}^{m} p^c$, *MUVIR* with $d = 1$, *MUVIR*-LI with $d = 1$ and $p = \max_{c=2}^{m} p^c$. From Fig. 8, we can see that *MUVIR* outperforms all other algorithms for finding all the minority class. With the same upper bound prior, GRADE-

LI needs 272 labeling requests while *MUVIR*-LI only needs 168 labeling requests to discover all the classes. If we apply random sampling, it may needs around 5000 labeling request to only identify the smallest minority class. Compared with Adult, we have better results on Statlog. It is because the distribution of majority class and minority classes are not meshed together as in Adult. Thus, to identify the minority classes in Statlog is a much easier case.

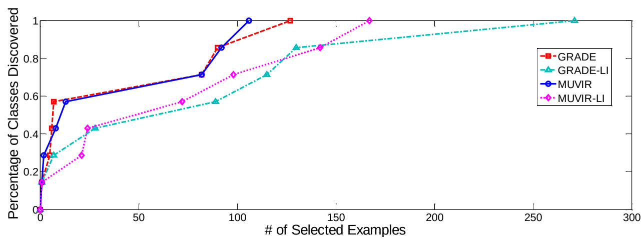

Figure 8: Statlog

5 Conclusion

In this paper, we have proposed a multi-view based method for rare category detection named *MUVIR*. Based on *MUVIR*, we also provided a modified version *MUVIR*-LI for dealing with real applications with less prior information. Different from existing methods, our methods exploit the relationship among multiple views and measure the probability belonging to target class for all examples. Our algorithm works well with multiple special cases: data sets with irrelevant features, data sets with multiple minority class and various correlation levels between minority class and majority class. The effectiveness of our proposed methods is guaranteed by theoretical justification and extensive experiments results on both synthetic and real data sets, especially in the presence of irrelevant views.

Acknowledgment

The authors gratefully acknowledge the support from the National Science Foundation under Grant Numbers IIP-1430144. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[Abney, 2002] Steven P. Abney. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 360–367, 2002.

[Balcan *et al.*, 2004] Maria-Florina Balcan, Avrim Blum, and Ke Yang. Co-training and expansion: Towards bridging theory and practice. In *Advances in neural information processing systems*, pages 89–96, 2004.

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[Chen *et al.*, 2011] Minmin Chen, Yixin Chen, and Kilian Q Weinberger. Automatic feature decomposition for single view co-training. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 953–960, 2011.

[Günnemann *et al.*, 2014] Stephan Günnemann, Ines Färber, Matthias Rüdiger, and Thomas Seidl. Smvc: semi-supervised multi-view clustering in subspace projections. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 253–262. ACM, 2014.

[He and Carbonell, 2007] Jingrui He and Jaime G Carbonell. Nearest-neighbor-based active learning for rare category detection. In *Advances in neural information processing systems*, pages 633–640, 2007.

[He *et al.*, 2008] Jingrui He, Yan Liu, and Richard Lawrence. Graph-based rare category detection. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 833–838. IEEE, 2008.

[Ho, 1998] Tin Kam Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, 1998.

[Long *et al.*, 2008] Bo Long, S Yu Philip, and Zhongfei (Mark) Zhang. A general model for multiple view unsupervised learning. In *SDM*, pages 822–833. SIAM, 2008.

[Muslea *et al.*, 2003] Ion Muslea, Steven Minton, and Craig A Knoblock. Active learning with strong and weak views: A case study on wrapper induction. In *IJCAI*, volume 3, pages 415–420, 2003.

[Muslea *et al.*, 2006] Ion Muslea, Steven Minton, and Craig A Knoblock. Active learning with multiple views. *Journal of Artificial Intelligence Research*, pages 203–233, 2006.

[Sindhwani and Rosenberg, 2008] Vikas Sindhwani and David S Rosenberg. An rkhs for multi-view learning and manifold co-regularization. In *Proceedings of the 25th international conference on Machine learning*, pages 976–983. ACM, 2008.

[Song *et al.*, 2013] Le Song, Animashree Anandkumar, Bo Dai, and Bo Xie. Nonparametric estimation of multi-view latent variable models. *arXiv preprint arXiv:1311.3287*, 2013.

[Yu *et al.*, 2011] Shipeng Yu, Balaji Krishnapuram, Rómer Rosales, and R Bharat Rao. Bayesian co-training. *The Journal of Machine Learning Research*, 12:2649–2680, 2011.