Contextual Outlier Interpretation

Ninghao Liu,[1] Donghwa Shin,[1] Xia Hu[1,2]
[1]Department of Computer Science and Engineering, Texas A&M University
[2]Center for Remote Health Technologies and Systems, Texas A&M Engineering Experiment Station
{nhliu43, donghwa_shin, xiahu}@tamu.edu

Abstract

While outlier detection has been intensively studied in many applications, interpretation is becoming increasingly important to help people trust and evaluate the developed detection models through providing intrinsic reasons why the given outliers are identified. It is a nontrivial task for interpreting the abnormality of outliers due to the distinct characteristics of different detection models, complicated structures of data in certain applications, and imbalanced distribution of outliers and normal instances. In addition, contexts where outliers locate, as well as the relation between outliers and the contexts, are usually overlooked in existing interpretation frameworks. To tackle the issues, in this paper, we propose a Contextual Outlier INterpretation (COIN) framework to explain the abnormality of outliers spotted by detectors. The interpretability of an outlier is achieved through three aspects, i.e., outlierness score, attributes that contribute to the abnormality, and contextual description of its neighborhoods. Experimental results on various types of datasets demonstrate the flexibility and effectiveness of the proposed framework.

1 Introduction

Outlier detection, which is to identify isolated instances that are different from the majority, has become an effective computational tool in real-world applications such as detecting spams [Liu *et al.*, 2017b; Shah, 2017], disease outbreaks [Wong *et al.*, 2002], and mis-behavioral IP sources in networks [Tong and Lin, 2011]. Numerious algorithms have been proposed for outlier detection, including density-based [Breunig *et al.*, 2000; Aggarwal and Yu, 2001; Gao *et al.*, 2010], distance-based [Knorr and Ng, 1999; Liu *et al.*, 2012] and model-based methods [He *et al.*, 2003; Tong and Lin, 2011; Li *et al.*, 2017]. Some other work tackles the curse of dimensionality [Filzmoser *et al.*, 2008], the massive data volumn [Ramaswamy *et al.*, 2000; Lucic *et al.*, 2016] and data heterogeneity [Chen *et al.*, 2016]. However, the essential factors that result in the outliers being detected are usually ignored and cannot be revealed with the detection outcome to end users.

Complementing existing work, enabling interpretability could benefit outlier detection and analysis in several aspects. First, interpretation helps bridge the gap between detecting outliers and identifying domain-specific anomalies. Outlier detection can output data instances with rare and noteworthy patterns, but in many applications we still rely on domain experts to manually select domain-specific anomalies from outliers that they actually care about in the current application. For example, in e-commerce website monitoring, outlier detection can discover users or merchants with rare behaviors, but administrators need to check the results to select those involved in malicious activities such as fraud. Interpretation of the detected outliers, which provides reasons for outlierness, can significantly save the effort of such manual inspection. Second, interpretation can be used in the evaluation process to complement current metrics such as the area under ROC curve (AUC) and nDCG [Davis and Goadrich, 2006] which provide limited information about characteristics of the detected outliers. Third, a detection method that works well in one dataset or application is not guaranteed to have good performance in others. Unlike supervised learning methods, outlier detection is usually performed using unsupervised methods and cannot be evaluated in the same way. Thus, effective outlier interpretation would significantly facilitate the usability of outlier detection techniques in real-world applications.

One straightforward way for outlier interpretation is to apply feature selection to identify a subset of features that distinguish outliers from normal instances [Knorr and Ng, 1999; Micenková *et al.*, 2013; Duan *et al.*, 2014; Vinh *et al.*, 2016; Gao *et al.*, 2017]. However, first it is difficult for some existing methods to efficiently handle datasets of large size or high dimensions, or effectively obtain interpretations from complex data types and distributions. Second, we measure the outlierness score of outliers through interpretation, which is important in many applications where some actions may be taken to outliers with higher priority. Some detectors only output binary labels indicating whether each data instance is an outlier. Sometimes continuous outlier scores are provided, but they are usually in different scales for different detection methods. A unified scoring mechanism by interpretation could facilitate the comparisons among various detectors. Third, besides identifying the notable attributes of outliers, we also analyze the context (e.g., contrastive neighborhood) in which outliers are detected. "It takes two to tango." Discovering the relations between an outlier and its context would provide richer information before taking actions to deal with the detected outliers in real applications.

To tackle the aforementioned challenges, in this paper, we propose a novel Contextual Outlier INterpretation (COIN) framework to provide explanations for outliers. We define the interpretation of an outlier from three aspects: abnormal attributes, the score of outlierness and the contrastive context with respect to the outlier. The first two elements are extracted from the relations between the outlier and its context. COIN can also be applied to existing outlier detection methods which already provide explanations for their results. In addition, prior knowledge about the roles of attributes in specific application scenarios can be easily incorporated with interpretation results, in order to enable end users to filter the given outliers and select the ones that are practically meaningful for the application. The contributions of this paper are summarized as follows:

- We define the interpretation of an outlier as three aspects: abnormal attributes, outlierness score, and the identification of the outlier's local context.
- We propose a novel model-agnostic framework to interpret outliers, as well as designing a concrete model within the framework to extract interpretation information.
- Comprehensive evaluations on interpretation quality, as well as case studies, are conducted through experiments on both real-world and synthetic datasets.

2 Preliminaries

Interpretation is receiving increasing attention in many machine learning applications. Some recent work gives explanation of the prediction results of classifiers [Ribeiro *et al.*, 2016; Koh and Liang, 2017]. Also, some outlier detection methods provide explanation together with detection results [Perozzi *et al.*, 2014; Liu *et al.*, 2017a; Liang and Parthasarathy, 2016], but they cannot be simply adopted by all detection methods.

Problem Definition Here we formally define the outlier interpretation problem as follows. Given a dataset $\mathcal{X} = \{\mathbf{x}_i \in \mathbb{R}^M | i \in [1, N]\}$ and the query outliers $\mathcal{O}$ detected therefrom, the *interpretation* for each outlier $\mathbf{o}_i \in \mathcal{O}$ is defined as a composite set: $\mathcal{E}_i = \{\mathcal{A}_i, d(\mathbf{o}_i), \mathcal{C}_i = \{\mathcal{C}_{i,l} | l \in [1, L]\}\}$. Here $\mathcal{C}_i$ denotes the *context* (i.e., k-nearest normal instances) of the outlier, $\mathcal{C}_{i,1}, \mathcal{C}_{i,2}, ..., \mathcal{C}_{i,L}$ are clusters in $\mathcal{C}_i$, and L is the number of clusters. $\mathcal{A}_i$ represents the abnormal attributes of $\mathbf{o}_i$ in contrast to $\mathcal{C}_i$. We use "inliers" and "normal instances" interchangeably in this paper. $d(\mathbf{o}_i) \in \mathbb{R}_{\geq 0}$ is the outlierness score of $\mathbf{o}_i$. The reason for clustering the context is illustrated in Figure 1. There are three clusters, each of which represents images of a digit. Red points are the detected outliers. Clusters of digit "2" and "5" compose the context of outlier $\mathbf{o}_1$. The interpretation of $\mathbf{o}_1$ can be obtained by comparing it with the two clusters respectively. However, it would be difficult to explain the outlierness of $\mathbf{o}_1$ if clusters of digit "2" and "5" are not differentiated.

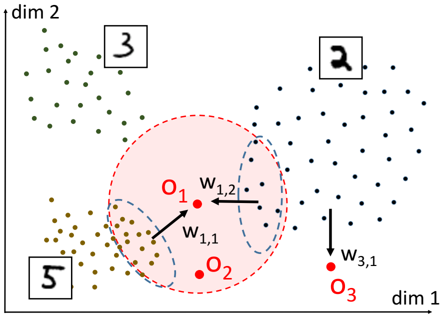


Figure 1: A toy example of outlier interpretation after resolving its context into clusters.

3 Contextual Outlier Interpretation Framework

The general framework of Contextual Outlier INterpretation (COIN) is illustrated in Figure 2. Given a dataset $\mathcal{X}$ and a set of outliers $\mathcal{O}$, we map the interpretation task to a classification problem. Then, the classification problem over the whole data is partitioned to a series of regional problems around each outlier query. Finally, interpretation is obtained from regional classification models.

3.1 Explaining Outlier Detector Using Classifiers

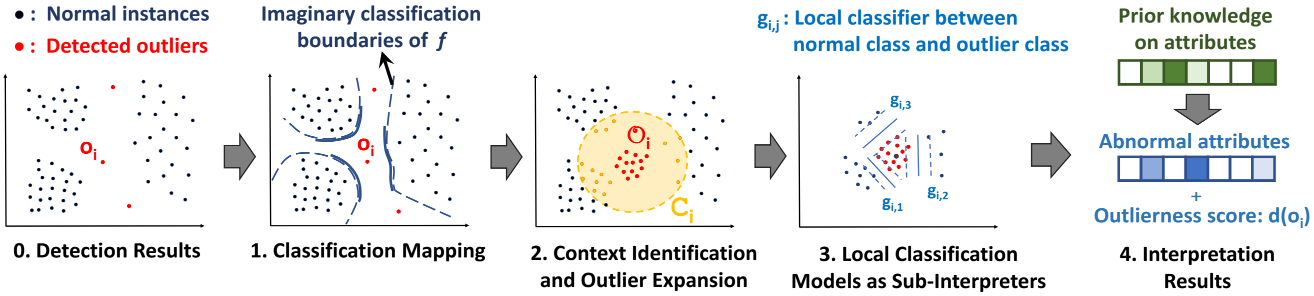
In this subsection, we establish the correlation between outlier detection and traditional supervised classification problems. Formally, an outlier detector can be denoted as $h(\mathbf{x}|\theta, \mathcal{X})$, where θ denotes the parameters. Here $\mathcal{X}$ is also treated as parameters since data instances affect the outlierness of each other. The abnormality of input $\mathbf{x}$ is typically represented by either a binary or continuous score, while the latter case can be easily transformed to the former if a threshold is set to separate inliers and outliers. This motivates us to explain outlier detectors using classification models. Although outlier detection is usually tackled as an unsupervised learning problem, there exists an imaginary hyperplane specified by certain decision function $f(\mathbf{x}|\theta') : \mathbb{R}^M \rightarrow \{0, 1\}$ that separates outliers from normal instances. Here θ' represents the parameters of f. An example can be found in Step 1 of Figure 2. Blue points and red points are normal instances and outliers, respectively, while dotted curves indicate the decision boundaries. The problem of building the decision function f is formulated as below,

$$\min_f \mathcal{L}(h, f; \mathcal{O}, \mathcal{X} - \mathcal{O}), \tag{1}$$

where $\mathcal{L}$ is the loss function including classification error and regularization terms. $\mathcal{O}$ and $\mathcal{X} - \mathcal{O}$ represent outlier class and inlier class, respectively.

By utilizing the isolation property of outliers, we can further decompose the problem in Equation (1) into multiple regional tasks of explaining individual outliers:

$$\begin{aligned} &\min_f \mathcal{L}(h, f; \mathcal{O}, \mathcal{X} - \mathcal{O}) \Rightarrow \min_f \sum_i \mathcal{L}(h, f; \mathbf{o}_i, \mathcal{C}_i) \\ &\Rightarrow \sum_i \min_{g_i} \mathcal{L}(h, g_i; \mathbf{o}_i, \mathcal{C}_i) \Rightarrow \sum_i \min_{g_i} \mathcal{L}(h, g_i; \mathcal{O}_i, \mathcal{C}_i). \end{aligned} \tag{2}$$


• : Normal instances
• : Detected outliers
Imaginary classification boundaries of f
g_{i,j} : Local classifier between normal class and outlier class
Prior knowledge on attributes
o_i
o_i
O_i
C_i
g_{i,3}
g_{i,2}
g_{i,1}
Abnormal attributes
+
Outlierness score: d(o_i)

0. Detection Results

1. Classification Mapping

2. Context Identification and Outlier Expansion

3. Local Classification Models as Sub-Interpreters

4. Interpretation Results

Figure 2: The Framework for Contextual Outlier Interpretation

In this way, the original problem is transformed to explaining each outlier $\mathbf{o}_i$ with respect to its context counterpart $\mathcal{C}_i$. Note that it is computationally efficient, given that the number of outliers is usually small. Here g_i represents the local parts of f exclusively for classifying $\mathbf{o}_i$ and $\mathcal{C}_i$. In Figure 2, for example, g_i is highlighted by the bold boundaries around $\mathbf{o}_1$ in Step 1, and $\mathcal{C}_i$ consists of the normal instances enclosed in the yellow circle in Step 2. Since there is a data imbalance between the two classes, we adopt synthetic sampling [He and Garcia, 2009] to expand $\mathbf{o}_i$ to an outlier class $\mathcal{O}_i$ with comparable size to $\mathcal{C}_i$. Local interpretation, encoded in g_i, can be obtained by approximating the local behavior of h between $\mathcal{O}_i$ and $\mathcal{C}_i$.

3.2 Resolving Context for Outlier Explanations

Now we focus on interpreting each single outlier $\mathbf{o}_i$ by solving g_i, from which we can extract interpretation results. Let $p_{\mathcal{O}_i}(\mathbf{x})$ and $p_{\mathcal{C}_i}(\mathbf{x})$ denote the probability density functions of the outlier class and inlier context class, respectively. Since the context $\mathcal{C}_i$ may contain complex cluster structures as shown in Figure 1, it is difficult to directly measure the degree of separation between $\mathcal{O}_i$ and $\mathcal{C}_i$ or to discover the attributes that discriminate the two classes. Therefore, we further decompose $\mathcal{L}(h, g_i; \mathcal{O}_i, \mathcal{C}_i)$ to a set of simpler problems. According to Bayesian decision theory, the error of classifying between $\mathcal{O}_i$ and $\mathcal{C}_i$ is

$$
\begin{aligned}
P^{err}(\mathcal{O}_i, \mathcal{C}_i) &= P(\mathcal{O}_i)\int_{\mathcal{C}_i} p(\mathbf{x}|\mathcal{O}_i)d\mathbf{x} + P(\mathcal{C}_i)\int_{\mathcal{O}_i} p(\mathbf{x}|\mathcal{C}_i)d\mathbf{x} \\
&\approx \Big(\sum_{l\in[1,L]} P(\mathcal{O}_i)\int_{\mathcal{C}_{i,l}} p(\mathbf{x}|\mathcal{O}_i)d\mathbf{x}\Big) + \Big(\sum_{l\in[1,L]} P(\mathcal{C}_{i,l})\int_{\mathcal{O}_i} p(\mathbf{x}|\mathcal{C}_{i,l})d\mathbf{x}\Big) \\
&= \sum_{l\in[1,L]} \Big(P(\mathcal{O}_i)\int_{\mathcal{C}_{i,l}} p(\mathbf{x}|\mathcal{O}_i)d\mathbf{x} + P(\mathcal{C}_{i,l})\int_{\mathcal{O}_i} p(\mathbf{x}|\mathcal{C}_{i,l})d\mathbf{x}\Big) \\
&\approx \sum_{l\in[1,L]} P^{err}(\mathcal{O}_{i,l}, \mathcal{C}_{i,l}). \qquad (3)
\end{aligned}
$$

Suppose we can split the context $\mathcal{C}_i$ into multiple clusters $\{\mathcal{C}_{i,l}|l \in [1, L]\}$ that are well separated from each other, then each term in the summation can be treated as an independent sub-problem without mutual inference. $\mathcal{O}_{i,l}$ is a subset of $\mathcal{O}_i$ close to $\mathcal{C}_{i,l}$. By combining Equation (2) and Equation (3), our final interpretation task is formulated as:

$$
\min_{f} \mathcal{L}(h, f; \mathcal{O}, \mathcal{X} - \mathcal{O}) \Rightarrow \min_{g_{i,l}} \sum_i \sum_l \mathcal{L}(h, g_{i,l}; \mathcal{O}_{i,l}, \mathcal{C}_{i,l}). \qquad (4)
$$

By now we are able to classify $\mathcal{O}_{i,l}$ and $\mathcal{C}_{i,l}$ with a simple and explainable model $g_{i,l}$ such as linear models and decision trees, where the abnormal attributes $\mathcal{A}_{i,l}$ can be extracted from *model parameters*. The overall interpretation for $\mathbf{o}_i$ is obtained by integrating the results across all $\mathcal{C}_{i,l}$, $l \in [1, L]$.

The estimated time complexity for implementing the framework above is $O(|\mathcal{O}| \times L \times T_g)$, where T_g is the average time cost of constructing $g_{i,l}$. Due to the scarcity of outliers, $|\mathcal{O}|$ is expected to be small. Each $g_{i,l}$ involves $\mathcal{O}_{i,l}$ and $\mathcal{C}_{i,l}$. T_g is also expected to be small since $\mathcal{C}_{i,l}$ and $\mathcal{O}_{i,l}$ are of small sizes. Moreover, the interpretation processes of different outliers are independent of each other, thus can be implemented in parallel to further reduce the time cost.

4 Distilling Interpretation from Models

After introducing the general framework of mapping outlier interpretation into a collection of classification tasks around individual outliers, in this section, we propose a concrete model to explain each outlier, including discovering its abnormal attributes and measuring the outlierness score.

4.1 Context Identification and Clustering

Given an outlier $\mathbf{o}_i$ spotted by the detector h, first we need to identify its context $\mathcal{C}_i$ in the data space. As introduced in Section 2, $\mathcal{C}_i$ consists of the nearest neighbors of $\mathbf{o}_i$. Here we use Euclidean distance as the point-to-point distance measure. The neighbors are chosen only from normal instances. The instances in $\mathcal{C}_i$ are regarded as the representatives for the local background around the outlier. Although $\mathcal{C}_i$ contains only a small number of data instances compared to the size of the whole dataset, they constitute the border regions of the inlier class and thus are adequate to discriminate between inlier and outlier classes, as shown in the Step 2 of Figure 2.

As local context may indicate some interesting structures (e.g., instances with similar semantics are located close to each other in the attribute space), we further segment $\mathcal{C}_i$ into multiple disjoint clusters. To determine the number of clusters L in $\mathcal{C}_i$, we adopt the measure of *prediction strength* [Tibshirani and Walther, 2005] which shows good performance even when dealing with high-dimensional data. After choosing the value of L, common clustering algorithms such as K-means or hierarchical clustering are applied to divide $\mathcal{C}_i$ into multiple clusters as $\mathcal{C}_i = \{\mathcal{C}_{i,1}, \mathcal{C}_{i,2}, \cdots, \mathcal{C}_{i,L}\}$. Clusters of small size, i.e., $|\mathcal{C}_{i,l}| \le 0.03 \cdot |\mathcal{C}_i|$, are abandoned in subsequent procedures.

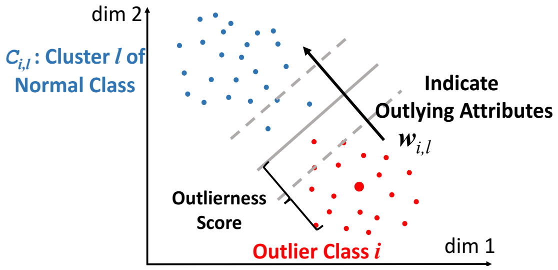

dim 2
$C_{i,l}$: Cluster l of Normal Class
Indicate Outlying Attributes
$w_{i,l}$
Outlierness Score
Outlier Class i
dim 1

Figure 3: Outlier Interpretation from SVM Parameters

4.2 Maximal-Margin Linear Explanations

The concrete type of models chosen for $g_{i,l}$ should have the following properties. First, it is desirable to keep $g \in G$ simple in form. For example, we may expect the number of non-zero weights to be small for linear models, or the rules to be concise in decision trees [Ribeiro *et al.*, 2016]. Here we let $g \in G$ belong to linear models, i.e., $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$. We impose the l_1-norm constraint on $\mathbf{w}$, where attributes a_m that correspond to large $|w[m]|$ values are regarded as abnormal. Second, since outliers are usually highly isolated from their context, there could be multiple solutions all of which could classify the outliers and inliers almost perfectly, but we want to choose the one that best reflects such isolation property. This motivates us to choose l_1 norm support vector machine [Zhu *et al.*, 2004] to build g. The local loss $\mathcal{L}(h, g_{i,l}; \mathcal{O}_{i,l}, \mathcal{C}_{i,l})$ to be minimized in Equation (4) is thus as below:

$$\begin{aligned} &\sum_{n=1}^{N_{i,l}} (1 - y_n g(\mathbf{x}_n) - \xi_n)_+ + c \sum_{n=1}^{N_{i,l}} \xi_n, \\ &\text{s.t.} \quad \xi_n \geq 0, \quad \|\mathbf{w}\|_1 \leq b \end{aligned} \tag{5}$$

where $N_{i,l} = |\mathcal{O}_{i,l} \cup \mathcal{C}_{i,l}|$, $(.)_+$ is the hinge loss, ξ_n is the slack variable, b and c are the parameters. Here $y_n = 1$ if $\mathbf{x}_n \in \mathcal{C}_{i,l}$ and $y_n = -1$ if $\mathbf{x}_n \in \mathcal{O}_{i,l}$.

From the parameters of the local model $g_{i,l}$, we can find the abnormal attributes and compute the outlierness score with respect to $\mathcal{C}_{i,l}$. Let $\mathbf{w}_{i,l}$ denote the weight vector of $g_{i,l}$, the importance of attribute a_m with respect to the context of $\mathcal{C}_{i,l}$ is thus defined as $s_{i,l}(a_m) = |w_{i,l}[m]|/\gamma_{i,l}^m$. Here $\gamma_{i,l}^m$ denotes the resolution of attribute a_m in $\mathcal{C}_{i,l}$, i.e., the average distance along the m^{th} axis between an instance in $\mathcal{C}_{i,l}$ and its closest neighbors. The overall score of a_m for $\mathbf{o}_i$ is

$$s_i(a_m) = (1/|\mathcal{C}_i|) \sum_l |\mathcal{C}_{i,l}| s_{i,l}(a_m), \tag{6}$$

which is the weighted average score for a_m over all L clusters. Attributes a_m with large $s_i(a_m)$ are regarded as the abnormal attributes for $\mathbf{o}_i$ (i.e., $a_m \in \mathcal{A}_i$). For the outlierness score $d(\mathbf{o}_i)$, we define it as:

$$d_l(\mathbf{o}_i) = |g_{i,l}(\mathbf{o}_i)| / \|\mathbf{w}_{i,l}\|_2. \tag{7}$$

This measure is robust to high dimensional data, as $\mathbf{w}$ is sparse and $d_l(\mathbf{o}_i)$ is calculated in a low dimensional space. An example is shown in Figure 3, where abnormal attributes are indicated from weight vector $\mathbf{w}$ and the outlierness score is shown. The overall outlierness score for $\mathbf{o}_i$ across all context clusters is:

$$d(\mathbf{o}_i) = (1/|\mathcal{C}_i|) \sum_l |\mathcal{C}_{i,l}| \, d_l(\mathbf{o}_i)/\gamma_{i,l}, \tag{8}$$

which is the weighted summation over different context clusters. Here the normalization term $\gamma_{i,l}$ is the average distance from an instance to its closest neighbor in $\mathcal{C}_{i,l}$. Now we have obtained all of the three aspects of interpretation $\mathcal{E}_i = \{\, \mathcal{A}_i, d(\mathbf{o}_i), \mathcal{C}_i = \{\mathcal{C}_{i,l} | l \in [1, L]\} \,\}$.

4.3 Filtering Outliers with Interpretation and Prior Knowledge

In real-world applications, the importance of different attributes varies according to different scenarios [Yang *et al.*, 2011; Ntoulas *et al.*, 2006]. Take social network spammer detection as an example. We have two account attributes: the number of followers (N_{fer}) and the ratio of tweets posted by API (R_{API}). A spammer account tends to have a small N_{fer} value as it is socially inactive, but large R_{API} to conveniently generate malevolent content. However, it is easy for spammers to intentionally increase their N_{fer} by purchasing followers, but manually decreasing R_{API} is more difficult due to expensive human labors. In this sense, R_{API} is more robust than N_{fer} in translating detected outliers as spammers. Therefore, we introduce two vectors $\boldsymbol{\beta}$ and $\boldsymbol{p}$, where $\beta_m \in \mathbb{R}_{\geq 0}$ denotes the prior knowledge about the robustness of a_m, and $p_m \in \{-1, 0, 1\}$ denotes the expected perturbation direction of a abnormal attribute. $p_m = -1$ means we expect outliers to have small value for a_m (e.g., N_{fer}), $p_m = 1$ means the opposite (e.g., R_{API}), while $p_m = 0$ means there is no preference. Thus, the outlierness score of $\mathbf{o}_i$ with respect to $\mathcal{C}_{i,l}$ is refined as:

$$d_l(\mathbf{o}_i) = \| \frac{|g_{i,l}(\mathbf{o}_i)|}{\|\mathbf{w}_{i,l}\|_2} \frac{\mathbf{w}'_{i,l}}{\|\mathbf{w}_{i,l}\|_2} \circ \boldsymbol{\beta} \|, \tag{9}$$

where the operator $\circ$ denotes element-wise multiplication, $w'[m] = \min(0, w[m])$ if $p_m = 1$, and $w'[m] = \max(0, w[m])$ if $p_m = -1$. If we label outliers with 1 and inliers with -1, the sign of $\mathbf{p}$ is reversed. The reason for introducing $\mathbf{w}'$ is that, if interpretation does not conform with the prior knowledge, such as an outlier in spammer detection is interpreted as having low R_{API}, then the outlierness score of the outlier should be deducted.

5 Experiments

In this section, we present evaluation results to assess the effectiveness of our framework. We try to answer the following questions: 1) How accurate is the proposed framework in identifying abnormal attributes of given outliers? 2) Can we accurately measure the outlierness score of outliers? 3) How effective is the prior knowledge of attributes in refining outlier detection results?

5.1 Datasets

We use both real and synthetic datasets in experiments. We follow the procedures in [Keller *et al.*, 2012] and create two synthetic datasets with ground-truth abnormal attributes for

	SYN1	SYN2	WBC	Twitter	MNIST
N	405	405	458	11,000	42,000
M	15	15	9	16	150
$\|\mathcal{O}\|$	30	30	25	1,000	1,000

Table 1: Details of the datasets in experiments

each outlier. In the first synthetic dataset, each outlier is close to only one normal cluster and far away from the others. In the second synthetic dataset, each outlier is in the vicinity of several normal clusters simultaneously, so the scenario is more complicated. The real-world datasets used in our experiments include Wisconsin Breast Cancer (WBC) dataset [Asuncion and Newman, 2007], MNIST dataset and Twitter spammer dataset [Yang *et al.*, 2011]. WBC dataset records the measurements for breast cancer cases with two classes, i.e. benign and malignant. The former class is considered as normal, while we downsampled 25 malignant cases as the outliers. MNIST dataset includes a collection of 28×28 images of handwritten digits. Here we use the training set which contains 42,000 examples. Instead of using raw pixels as attributes, we build a Restricted Boltzmann Machine (RBM) with 150 latent units to map images to a low-dimensional space which is more proper for interpretation than raw pixels. A multi-label logistic classifier is then built to classify digits, and the ground-truth outliers are selected as the misclassified instances downsampled to $1,000$. The Twitter dataset contains information of normal users and spammers crawled from Twitter. Following [Yang *et al.*, 2011], we divide attributes into two categories according to whether they are robust to the spammers in disguise. Attributes of low robustness refer to those which can be easily controlled by spammers to avoid being detected, while attributes of high robustness are the opposite.

5.2 Baseline Methods

We include some recent outlying-aspect mining and classifier interpretation methods as baseline methods:

- CA-lasso (CAL) [Micenková *et al.*, 2013]: An interpretation method that analyzes the separability between outlier and inliers as a linear classification problem solved with LASSO, without further clustering the context of outliers.
- IPS-BS [Vinh *et al.*, 2016]: An interpretation method that applies isolation path score to measure outlierness. Beam Search is then applied to look for the abnormal attributes.
- LIME [Ribeiro *et al.*, 2016]: A global classifier is first constructed to classify outliers and inliers. Then the abnormal attributes for each outlier is identified by locally interpreting the classification model around the outlier. A neural network is used as the global classifier for MNIST data, and SVMs with RBF kernel are used for other datasets.

5.3 Abnormal Attributes Evaluation

The goal of this experiment is to verify that the identified attributes indeed explain the abnormality. Since ground-truth abnormal attributes of real-world datasets are not available, we append M Gaussian-noise attributes to all real-world data instances. Noise attributes are not expected to be identified as abnormal as they are of small magnitudes. In our experiments, we choose $0.08 \times N$ nearest neighbors of an outlier $\mathbf{o}_i$ as its context $\mathcal{C}_i$. The radius of synthetic sampling for building the outlier class $\mathcal{O}_i$ is set as half of the average distance to the inlier class $\mathcal{C}_i$ to avoid overlap between $\mathcal{O}_i$ and $\mathcal{C}_i$. The parameters of SVMs are tuned by validation, where some samples from $\mathcal{O}_i$ and $\mathcal{C}_i$ are randomly selected as the validation set. The same parameter values are used for all outliers in the same dataset. We report the Precision, Recall and F_1 score averaged over all the outliers in Table 2. Besides finding that COIN shows relatively better performance, some observations can be made as follows:

	COIN			CAL			IPS-BS			LIME		
	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
SYN1	**0.97**	**0.89**	**0.93**	0.89	0.81	0.84	0.87	0.44	0.58	0.82	0.79	0.80
SYN2	**0.99**	**0.90**	**0.94**	0.92	0.70	0.80	**1.00**	0.37	0.54	0.91	0.70	0.79
WBC	0.86	0.37	**0.52**	0.84	0.37	0.51	**0.90**	0.15	0.26	0.35	**0.39**	0.37
Twitter	**0.91**	0.33	0.48	0.75	0.34	0.47	0.72	0.29	0.41	0.60	**0.67**	**0.63**

Table 2: Performance of abnormal attributes identification

- In general, the Recall value of SYN2 is lower than that of SYN1, because the context of each outlier in SYN2 has several clusters, and the true abnormal attributes vary among different clusters. In this case, retrieving all ground-truth attributes is more challenging.
- IPS-BS is more cautious in making decisions. It tends to stop early if the discovered abnormal attributes already make the outlier query well isolated. Therefore, IPS-BS has high Precision, but only a small portion of true attributes are discovered (low Recall).
- The Recall scores are low for real-world data since we treat all original attributes to be the ground truth, so low Recall values do not necessarily mean bad performances.

5.4 Outlierness Score Evaluation

We evaluate if interpretation methods are able to accurately measure the outlierness score of outlier queries. For each dataset, we randomly sample the same number of inliers as the outliers, and use them together as queries to interpreters. The label is 1 for each true outlier, and 0 for each inlier. For each query, interpreters are asked to estimate its outlierness score. After that, we rank the instances in a descending order with respect to their outlierness scores. Since true outliers are more isolated, an effective interpreter should convert such isolation degree to larger scores.

We report the results in Table 3 with AUC as the evaluation metric. The proposed method achieves better performance than the baseline methods especially on SYN2 and MNIST. This can be explained by the more complex structures in these datasets, where an outlier may be close to several neighboring clusters. COIN resolves the contextual clusters around each outlier, so it can handle such scenario. This also explains why IPS_BS is also more effective in complex datasets than the other two baseline methods. The isolation tree used in IPS_BS can handle complex cluster structures.

AUC	SYN1	SYN2	WBC	Twitter	MNIST
COIN	0.78	0.93	0.96	0.85	0.87
CAL	0.71	0.63	0.94	0.81	0.76
IPS_BS	0.69	0.91	0.90	0.79	0.82
LIME	0.74	0.62	0.94	0.83	0.78

Table 3: Outlierness score ranking performance

5.5 Filtering Outliers with Prior Knowledge

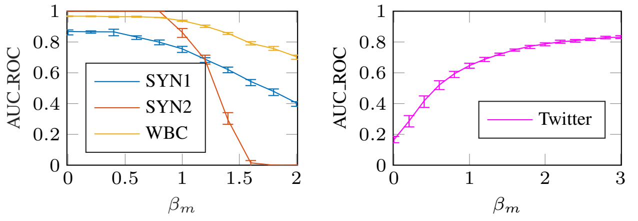
In this experiment, we discuss if interpretations, together with prior knowledge, can help filtering existing outliers to satisfy the demand of specific applications.

The experiment has two parts. In the first part, we append M new noise attributes to data instances, so each instance is augmented to $\mathbf{x} \in \mathbb{R}^{2M}$. Different from the noise attributes in Section 5.3 that are of small magnitude, the attributes here may turn inliers to "outliers". However, these new outliers are irrelevant to the ground truth. We sample $0.5 \times |\mathcal{O}|$ inliers, together with ground-truth outliers, as queries fed into COIN. We set $\boldsymbol{p}$ to be zero and run COIN with different $\boldsymbol{\beta}$ values. The weights corresponding to original attributes are fixed to 1 ($\beta_m = 1, m \in [1, M]$), and we only vary the weights of noise attributes ($\beta_m = \beta, m \in [M+1, 2M]$). Similar to Section 5.4, we obtain the outlierness score for all queries and rank them in a descending order according to the score. Ground-truth outliers are expected to rank higher. The ranking performance is reported in Figure 4a. The plot indicates that as we increase the weights of noise attributes, the performance of the interpreter degrades for all datasets, because it is more difficult to distinguish between real outliers and noisy instances. From the opposite perspective, assigning large weights to important attributes will filter out mis-detected outliers.

The second part of the experiment uses the Twitter dataset in which features extracted from user profiles, posts and graph structures are used as attributes. According to [Yang *et al.*, 2011], the robustness level varies for different attributes. Some attributes, such as the number of followers, hashtage ratio and reply ratio, can be easily controlled by spammers to avoid being captured, so they are of low robustness, while some other attributes such as account age, API ratio and URL ratio have high robustness. In this experiment, we fix the weight of low-robustness attributes to 1, and vary the weight β_m of high-robustness attributes. The remaining procedures are the same as first part of the experiment discussed above. The result of outlierness ranking is reported in Figure 4b. The rising curve shows that as more emphasis is put on high-robust attributes, we are able to refine the performance of spammer identification. The experiment result indicates that by resorting to the interpretation of outliers, we can gain more insights on their characteristics, and adaptively select those that are in accordance with the specific application.

5.6 Case Studies

We conduct some case studies to illustrate interpretation results on MNIST. The attributes are the hidden features extracted by the RBM instead of raw pixels. The case study results are shown in Figure 5. There are three query outlier images shown in the first row. We choose two neighboring clusters for each query, and compute the average image of each cluster, as shown in the second row. The average images can be seen as part of the contexts of outliers. Clear handwritten digits can be seen from the average images, so that the clusters are internally coherent. The third and fourth rows together indicate the noteworthy attributes of the query image with respect to the corresponding average images. The black strokes enclosed by red circles in third-row images represent positive abnormal attributes, i.e., the query image is regarded as an outlier instance because it possesses these attributes. The strokes enclosed by blue circles in fourth-row images are negative abnormal attributes, as the query outlier digit does not include them. These negative attributes, however, commonly appear in the neighbor images of the outlier. The positive and negative attributes together explain why the outlier image is different from its nearby normal images.



(a) Data with noise attributes

(b) Twitter spammer data

Figure 4: The influence of the prior knowledge on outlierness score. Results averaged over 20 runs, bars depict 25-75%.

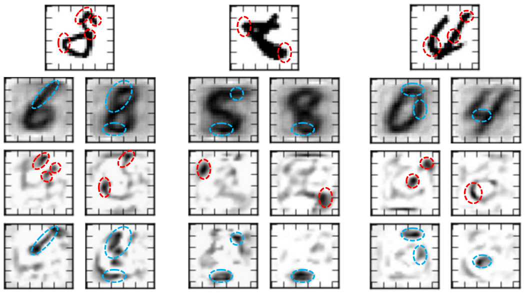


Figure 5: Visualization of outlier interpretation on MNIST dataset.

6 Conclusion and Future Work

In this paper, we propose a model-agnostic outlier interpretation framework by resolving outliers' local context. We define the interpretation of an outlier from three aspects including the abnormal attributes, outlierness score and the outlier's context. Interpretation is distilled from the results of a series of classification tasks. Prior knowledge in different applications can be incorporated with interpretation results to refine the outlier detection result. Interesting extensions include applying hierarchical clustering to accurately partition the whole data space, considering heterogeneous data sources and incoporating deep models [He *et al.*, 2017].

Acknowledgments

The work is, in part, supported by DARPA (#N66001-17-2-4031) and NSF (#IIS-1657196, #IIS-1718840). The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

[Aggarwal and Yu, 2001] Charu C Aggarwal and Philip S Yu. Outlier detection for high dimensional data. In *ACM Sigmod Record*, volume 30, pages 37–46. ACM, 2001.

[Asuncion and Newman, 2007] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.

[Breunig *et al.*, 2000] Markus Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, 2000.

[Chen *et al.*, 2016] Ting Chen, Lu-An Tang, Yizhou Sun, Zhengzhang Chen, and Kai Zhang. Entity embedding-based anomaly detection for heterogeneous categorical events. In *IJCAI*, 2016.

[Davis and Goadrich, 2006] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *ICML*, 2006.

[Duan *et al.*, 2014] Lei Duan, Guanting Tang, Jian Pei, James Bailey, Guozhu Dong, Akiko Campbell, and Changjie Tang. Mining contrast subspaces. In *PAKDD*, 2014.

[Filzmoser *et al.*, 2008] Peter Filzmoser, Ricardo Maronna, and Mark Werner. Outlier identification in high dimensions. *CSDA*, 2008.

[Gao *et al.*, 2010] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, and J. Han. On community outliers and their efficient detection in information networks. *KDD*, 2010.

[Gao *et al.*, 2017] Jun Gao, Ninghao Liu, Mark Lawley, and Xia Hu. An interpretable classification framework for information extraction from online healthcare forums. *Journal of healthcare engineering*, 2017.

[He and Garcia, 2009] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *TKDE*, 2009.

[He *et al.*, 2003] Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9):1641–1650, 2003.

[He *et al.*, 2017] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *WWW*, 2017.

[Keller *et al.*, 2012] Fabian Keller, Emmanuel Muller, and Klemens Bohm. Hics: high contrast subspaces for density-based outlier ranking. In *ICDE*. IEEE, 2012.

[Knorr and Ng, 1999] Edwin M Knorr and Raymond T Ng. Finding intensional knowledge of distance-based outliers. In *VLDB*, 1999.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*, 2017.

[Li *et al.*, 2017] Jundong Li, Harsh Dani, Xia Hu, and Huan Liu. Radar: Residual analysis for anomaly detection in attributed networks. In *IJCAI*, 2017.

[Liang and Parthasarathy, 2016] Jiongqian Liang and Srinivasan Parthasarathy. Robust contextual outlier detection: Where context meets sparsity. In *CIKM*, 2016.

[Liu *et al.*, 2012] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *TKDD*, 2012.

[Liu *et al.*, 2017a] Ninghao Liu, Xiao Huang, and Xia Hu. Accelerated local anomaly detection via resolving attributed networks. In *IJCAI*, 2017.

[Liu *et al.*, 2017b] Yuli Liu, Yiqun Liu, Ke Zhou, Min Zhang, and Shaoping Ma. Detecting collusive spamming activities in community question answering. In *WWW*, 2017.

[Lucic *et al.*, 2016] Mario Lucic, Olivier Bachem, and Andreas Krause. Linear-time outlier detection via sensitivity. In *IJCAI*, 2016.

[Micenková *et al.*, 2013] Barbora Micenková, Raymond T Ng, Xuan-Hong Dang, and Ira Assent. Explaining outliers by subspace separability. In *ICDM*, 2013.

[Ntoulas *et al.*, 2006] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting spam web pages through content analysis. In *WWW*, 2006.

[Perozzi *et al.*, 2014] Bryan Perozzi, Leman Akoglu, Patricia Iglesias Sánchez, and Emmanuel Müller. Focused clustering and outlier detection in large attributed graphs. *KDD*, 2014.

[Ramaswamy *et al.*, 2000] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *ACM SIGMOD Record*, 2000.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?": Explaining the predictions of any classifier. In *KDD*, 2016.

[Shah, 2017] Neil Shah. Flock: Combating astroturfing on livestreaming platforms. In *WWW*, 2017.

[Tibshirani and Walther, 2005] Robert Tibshirani and Guenther Walther. Cluster validation by prediction strength. *JCGS*, 2005.

[Tong and Lin, 2011] Hanghang Tong and Ching-Yung Lin. Non-negative residual matrix factorization with application to graph anomaly detection. *SDM*, 2011.

[Vinh *et al.*, 2016] Nguyen Xuan Vinh, Jeffrey Chan, Simone Romano, James Bailey, Christopher Leckie, Kotagiri Ramamohanarao, and Jian Pei. Discovering outlying aspects in large datasets. *DMKD*, 2016.

[Wong *et al.*, 2002] W. Wong, A. Moore, G. Cooper, and M. Wagner. Rule-based anomaly pattern detection for detecting disease outbreaks. In *AAAI/IAAI*, 2002.

[Yang *et al.*, 2011] Chao Yang, Robert Chandler Harkreader, and Guofei Gu. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *International Workshop on RAID*, 2011.

[Zhu *et al.*, 2004] Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. *NIPS*, 16(1):49–56, 2004.