Identification of Incorrect Data Labels Using Conditional Outlier Detection

Charmgil Hong

ABSTRACT

Outlier detection methods help one to identify unusual instances in data that may correspond to erroneous, exceptional, or surprising events or behaviors. This work studies conditional outlier detection, a special instance of the outlier detection problem, in the context of incorrect data label identification. Unlike conventional (unconditional) outlier detection methods that seek abnormalities across all data attributes, conditional outlier detection assumes data are given in pairs of input (condition) and output (response or label). Accordingly, the goal of conditional outlier detection is to identify incorrect or unusual output assignments considering their input as condition. As a solution to conditional outlier detection, this paper proposes the ratio-based outlier scoring (ROS) approach and its variant. The propose solutions work by adopting conventional outlier scores and are able to apply them to identify conditional outliers in data. Experiments on synthetic and real-world image datasets are conducted to demonstrate the benefits and advantages of the proposed approaches.

Key words: Conditional Outlier Detection, Outlier Analysis, Anomaly Detection

1. INTRODUCTION

With the spread of social network services and online media platforms, the availability of user-created data content increases drastically. These data are often accompanied by human-annotated labels, such as tags associated with images or videos, topic words assigned to Web articles, and titles given to sound clips, etc. In terms of data accessibility, such annotations are generally found to be useful in that they provide a brief idea of data content and can be utilized to design another information retrieval service.

As for the data labels annotated by human, however, those tags or keywords sometimes contain errors in that they are mistakenly selected by the user and irrelevantly matched with the data. Apparently, identifying and removing those erratic labels in user-annotated data are one of the important steps in data preprocessing, when one considers utilizing the data for any purpose. One way to analyze and discover such mistakes in data is outlier (or anomaly) detection. Outlier detection is a data analytic task that pursues to find unusual instances in a dataset [1-4]. It has been an active research topic in data science and artificial intelligence communities and frequently used in various applications to identify rare and interesting data patterns that may be associated with either beneficial or malicious events such as fraud identification [5, 6], network intrusion surveillance [7, 8], disease outbreak detection [9], patient monitoring for preventable adverse events (PAE) [10, 11], etc. It is also utilized as a primary data pre-processing step that helps to remove noisy or irrelevant signals in data [12, 13].

* Corresponding Author: Charmgil Hong, Address: (37554) Handong-ro 538, Buk-gu, Pohang, Gyeongbuk, Korea, TEL: +82-10-2663-9088, FAX: +82-54-260-1309, E-mail: charmgil@handong.edu

Receipt date: Jul. 10, 2020, Approval date: Jul. 20, 2020

† School of Computer Science and Electrical Engineering, Handong Global University
This work considers conditional outlier detection (COD) [10, 14] and proposes a novel approach to identify errors or mistakes in the annotation space. COD is a special case of the outlier detection problem that aims to find data objects with unusual values for a subset of variables given the values of the remaining variables. COD is particularly useful when data is considered to manifest conditional dependences among its variables: i.e., a set of variables defines context (input), while the rest are treated as response given the context. Unlike conventional (unconditional) outlier detection methods that seek abnormalities across all data attributes, COD examines and looks for data instances with unusual pairing of input and output values. Accordingly, in the problem of unusual image tags (labels) identification, the COD approach could be especially useful as the approach directly considers the annotation (output) space given associated data instance (input).

Despite its importance and usefulness, COD has received relatively less attention, while the majority of existing research have focused on unconditional outliers [2, 15]. However, as conditional outliers are fundamentally different from unconditional outliers in that conditional outliers reflect unusual responses for a given set of contextual attributes, applications of unconditional outlier detection methods to a COD problem may lead to incorrect results. For instance, consider where one wants to find mistaken image tags in a collection of annotated images. If an unconditional outlier detection method is applied jointly to both images and tags, images with rare themes may get flagged instead of images with mistaken tags (false positives) due to the scarcity of the themes in the dataset. Similarly, images with frequent themes but with unusual annotations may not be detected as outliers due to the abundance of similar themes in the dataset (false negative). Fig. 1 illustrates a COD problem with image-annotation pairs.

This paper proposes a new COD solution that works by comparing (via ratio) two unconditional outlier scores: one score calculated against data with the same observed output value; and another calculated against data with the opposite output value. This new approach offers a couple of merits in addressing the COD problem. First, it allows the user to utilize a wide variety of unconditional outlier scores. Second, it lets one effectively avoid the cases where instances with rare observation but properly associated with the output undesirably receive a high conditional outlier score. Through the paper, the derivation and development of the method is explained. With experiments, the performance and usefulness of the proposed method is demonstrated.

The rest of the paper is organized as follows. Section 2 revisits the background of conditional outlier detection and relevant research work. Section 3 presents the new COD framework and discuss its properties. Section 4 compares and
demonstrates the performance of the proposed solution. Section 5 concludes the paper.

2. BACKGROUND

2.1 Problem Definition: Conditional Outlier Detection (COD)

Given dataset \( D = \{ x^{(i)}, y^{(i)} \}_{i=1}^{N} \), where \( x \in \mathbb{R}^m \) and \( y \in \{0,1\} \), conditional outliers are the outliers that occur in the output space (\( Y \)) conditioned on the input (\( X \)). That is, the goal in conditional outlier detection (COD) is to identify data instances that have unusual output value given their input. This special type of outlier detection problem is challenging because the dependence relations between input and output should be taken into account when identifying outliers. The following subsections review the relevant research and pinpoint the main differences that the proposed approach has.

2.2 Conventional Outlier Detection

Conventionally COD for discrete output space is addressed by modeling the posterior distribution and by identifying data objects that do not fit well the model. In other words, data objects that are assigned a very low posterior probability are considered conditional outliers. Accordingly, depending on how to represent the posterior, a number of methods have been proposed. Hauskrecht et al. [10] is one of the earliest work that introduced the concept and importance of the conditional approach to the outlier problem. The authors used a localized Bayesian networks to represent data with discrete input and output variables. The same group [11, 17] further developed the framework to address COD with mixture of continuous and discrete inputs. Valko and Hauskrecht [18] investigated the instance-specific methods to acquire more accurate predictive models for COD. To select instances that are similar to the target instance, the authors presented a new metric learning algorithm.

Although the abovementioned methods directly tackled the COD problem, they are different from the approach proposed in this paper, in that the existing methods are based on specially designed data models or algorithms for conditional outlier detection. This work proposes a solution that works by adopting existing unconditional outlier detection methods.

2.3 Unconditional Outlier Detection

While the COD problem has been the focus of some recent research efforts, its unconditional counterpart has long been investigated as the main area of the outlier detection research. One way to summarize existing unconditional methods is to organize them according to the assumptions that each method builds upon. Out of many, the discussion conducted in this paper is directly relevant to the following categories of methods.

- Density-based approaches: Density-based approaches assume that the density around a normal data instance is similar to the density around its neighbors, while that of an outlier is relatively lower than its neighbors. Local Outlier Factor (LOF) [19] is one of the most popular methods in this category. LOF has shown a good performance in many applications and is considered as an off-the-shelf outlier detection method. Section 3.1.1 reviews this method more in detail and discusses how to combine LOF with the proposed approach.

- Classification-based approaches: Classification-based approaches are based on a parametric assumption that a function of feature classifying normal and outlier instances can be learned from data. One-class classification strategy is a popular technique falls in this category. It assumes that all training instances are normal, and attempts to learn a discriminative boundary around the training (normal) instances. For testing, instances that fall out of the obtained decision boundary are consid-
considered to be outliers. Support vector machines (SVMs) have been applied to outlier detection using this strategy [20]. Section 3.1.2 reviews this approach and discusses how it is combined in the proposed solution.

2.4 Ensemble Approach to Outlier Detection

The approach proposed in this work can be seen as an ensemble approach in that it combines unconditional outlier scores and weaves them into a conditional outlier score. There have been several works in this direction in the purpose of improving the overall (unconditional) outlier detection performance [21, 22]. However, to the best of knowledge, there has been no conditional outlier detection method that can be based on an ensemble approach.

3. PROPOSED METHOD

3.1 Ratio of Outlier Scores (ROS)

This paper proposes a new conditional outlier detection approach that works by comparing two unconditional outlier scores and forming a ratio-like score: One score is calculated against data with the same observed output value; Another calculated against data with the opposite output value. This new score is referred to as Ratio of Outlier Scores (or Ratio-based Outlier Scoring; ROS). It offers a couple of important advantages. First, it allows to utilize a wide variety of unconditional outlier scores. Also, it lets one effectively avoid the cases where instances with rare observation but properly associated with the output undesirably receive a high conditional outlier score.

More formally, consider a binary-labeled dataset $D=\{(\mathbf{x}^{(n)},y^{(n)})\}_{n=1}^{N}$ where each instance in $D$ consists of a continuous input vector $\mathbf{x}^{(n)} \in \mathbb{R}^m$, and an associated output value $y^{(n)}=0, 1$. For notational convenience, $D_{\text{agree}}(n)$ and $D_{\text{disagree}}(n)$ are defined to denote subsets of $D$ based on the value of $y^{(n)}$:

$$D_{\text{agree}}(n) = \{\mathbf{x} | y^{\prime} = y^{(n)}\}$$

A subset of $D$ whose output value is equal to $y^{(n)}$ ($D_{\text{agree}}(n)$ does not include the $n$-th instance)

$$D_{\text{disagree}}(n) = \{\mathbf{x} | y^{\prime} \neq y^{(n)}\}$$

A subset of $D$ whose output value is not equal to $y^{(n)}$.

Then, for the $n$-th data instance $(\mathbf{x}^{(n)},y^{(n)})$, $\text{Score}_{\text{ROS}}(y^{(n)}|\mathbf{x}^{(n)})$ is defined as the ratio between two unconditional outlier scores evaluated on $D_{\text{agree}}(n)$ and $D_{\text{disagree}}(n)$, respectively:

$$\text{Score}_{\text{ROS}}(y^{(n)}|\mathbf{x}^{(n)}) := \frac{\sigma_{\mathbf{x}^{(n)}}(\mathbf{x}^{(n)};D_{\text{agree}}(n))}{\sigma_{\mathbf{x}^{(n)}}(\mathbf{x}^{(n)};D_{\text{disagree}}(n))} \quad (1)$$

where $\sigma_{\mathbf{x}^{(n)}}(\mathbf{x}^{(n)};D)$ denotes an unconditional outlier score computed for $\mathbf{x}^{(n)}$ on dataset $D$.

ROS measures the unusualness of the input $\mathbf{x}^{(n)}$ being associated with its output $y^{(n)}$. For normal data instances $\text{Score}_{\text{ROS}}(y^{(n)}|\mathbf{x}^{(n)})$ would be low, which in turn indicates the outlier score based on $D_{\text{agree}}(n)$ is low and/or $D_{\text{disagree}}(n)$ is high. On the other hand, instances with a high ROS score are deemed as outliers, because $\text{Score}_{\text{ROS}}(y^{(n)}|\mathbf{x}^{(n)})$ is high when the outlier score based on $D_{\text{agree}}(n)$ is high and/or $D_{\text{disagree}}(n)$ is low.

Note that Equation (1) easily turns many existing unconditional outlier scores to conditional outlier scores. That is, it can compute and compare the conditional outlier score of data instances by simply applying any unconditional outlier score such as density-based outlier scores [19, 23], classification-based outlier scores [20, 24], etc. to the subsets of $D$ and computing their ratio. Another advantage of this approach is that it can properly handle instances that fall in regions of the input space with low support. For a data instance that does not have enough support (i.e., an instance falls in a sparse neighborhood of $X$), it is not straightforward to come up with an outlier score that is confident. However, the proposed score suffers less from the issue, because both $D_{\text{agree}}(n)$ and $D_{\text{disagree}}(n)$...
will be high in such a sparse region and, as a result, by cancelling each other out in Equation (1), the resulting conditional outlier score will not be high.

In summary, the new conditional outlier detection approach based on the ratio-score defines a general and flexible framework that allows one to plug in an unconditional outlier score and use it to perform conditional outlier detection. The next subsections (Sections 3.1.1 and 3.1.2) introduce how to use ROS in combination with the Local Outlier Factor (LOF) and One-Class Support Vector Machine (OCSVM) methods and their definitions [19, 20].

3.1.1 ROS with Local Outlier Factor (LOF)

Recall that LOF [19] is a non-parametric approach used to detect unconditional outliers based on the density of the local neighborhood of the target data instance. More specifically, it computes the outlier score of an instance by comparing the local density of the instance to the average local density of its \( k \) nearest neighbors:

\[
a_o(x^{(n)}; D) = \text{LOF}(x^{(n)}, k; D) = \frac{\sum_{x \in N(x^{(n)}; D)} \text{lr}(x; D)}{|N(x^{(n)}; D)|}
\]  

(2)

where \( a_o(x^{(n)}; D) \) denotes the unconditional outlier score for instance \( x^{(n)} \) and dataset \( D \), \( |N(x^{(n)}; D)| \) denotes the \( k \)-nearest neighbors of instance \( x^{(n)} \) in dataset \( D \), and

\[
\text{lr}(x; D) = \frac{|N(x; D)|}{\sum_{o \in N(x; D)} \max(\text{dist}_x(o), \text{dist}(x, o))}
\]  

(3)

is the local reachability density, which measures the geometric dispersion of the \( k \)-nearest neighborhood, where \( \text{dist}(x, o) \) denotes the distance between two instances \( x \) and \( o \); and \( \text{dist}_x(o) \) denotes the distance to the \( k \)-th nearest neighbor of \( o \). In this work the Mahalanobis distance is used to compute the pairwise distances.

To come up with the ROS score, Equation (2) is estimated against \( D_{\text{Agree}}(n) \) and \( D_{\text{Disagree}}(n) \), respectively, and the ratio of the two scores is computed (Equation (1)).

3.1.2 ROS with One-Class Support Vector Machine (OCSVM)

OCSVM is a classification-based method originally designed to detect unconditional outliers [20]. It trains an SVM classifier by learning a decision boundary between training data and the origin (zero). Any instances that lie across the boundary are classified as outliers. To compute the ROS score using OCSVM, two SVM models are trained separately on \( D_{\text{Agree}}(n) \) and \( D_{\text{Disagree}}(n) \), and define decision boundaries for unconditional outliers. Then the ratio between the raw scores estimated both models are used for conditional outlier detection.

3.2 ROS on Discriminative Projections

The above newly designed conditional outlier score leverages the ratio of two unconditional outlier scores that are computed in the input space. Hence, any issues affecting the quality of the unconditional scores are likely to be inherited to the new ratio score.

One of the recurring challenges of many unconditional outlier detection approaches is that they tend to exhibit poor performance when the data dimensionality is high. This is because in high-dimensional data spaces, with many (random) dimensions, all data objects appear to be sparse, and many of the distance metrics and density estimators become analytically ineffective and computationally intractable [25-27]. As a result, outliers become hard to define and detect. It is reasonable to assume that these limitations also translate to the conditional score when it comes to the proposed ratio outlier score, and devising a solution to improve the robustness of the method is appropriate.

One common way to resolve the problem of a high-dimensional space (in unconditional settings)
is to reduce the dimensionality of the space via various dimensionality reduction methods, such as principal component or independent component analysis \([28, 29]\), before the detection. However, the conditional outlier detection is different, as the importance of the input space and its individual dimensions depends on how important the dimensions are in defining (or predicting) the output. In such a case, various supervised space transformations or supervised metric learning approaches can be applied.

To cope with the dimensionality problem, this work adopts a relatively simple supervised dimensionality reduction approach that relies on a discriminative model and its output to define a lower-dimensional projection of the original (high-dimensional) input data. In principle, one can use one or more such discriminative projections. This work focuses on and experiments with one-dimensional projections, in which a high-dimensional input space is reduced to a one-dimensional discriminative space. Such projections can be built with the help of various classification learning methods. For example, by applying the logistic regression model to the dataset \(D\), a probabilistic projection of \((x^{(n)}, y^{(n)})\) representing \(P(Y = y^{(n)} | X = x^{(n)})\) is obtained. Similarly, by taking a raw output of a support vector machines model, non-probabilistic discriminative projections are obtained. Below denotes such a discriminative projection function as \(f\) and the projection of the function as \(\phi\).

\[
f: x^{(n)} \rightarrow \phi^{(n)}
\]

Obtaining a discriminative projection function \(f\) and its projections \(\phi\) is equivalent to training (learning) of a model on the input-output instances in \(D\). Assuming the logistic regression model, the parameters of the projection function are optimized as:

\[
\theta_f = \arg \max \sum_{n=1}^{N} \log P(y^{(n)} | x^{(n)}; \theta)
\]

After \(\theta_f\) is obtained from data, the projection function \(f\) on an observed input \(x^{(n)}\) as follows:

\[
\phi^{(n)} = f(x^{(n)}) = \frac{1}{1 + \exp(-x^{(n)} \theta_f)}
\]

To combine the discriminative projections with the ROS approach, firstly the original data is mapped to the projected discriminative space. After that, the ROS scores are computed on the newly projected space. This extension of ROS is referred to as Ratio of Outlier Scores on Discriminative Projections (ROSDP) approach. To formalize, given the parameters of the discriminative projection \(f\), the ROSDP score is defined as:

\[
\text{Score}_{\text{ROSDP}}(y^{(n)} | x^{(n)}, f) := \frac{a_{\text{f}}(f(x^{(n)}); D_{\text{Agree}(n)})}{a_{\text{f}}(f(x^{(n)}); D_{\text{Disagree}(n)})} = \frac{a_{\text{f}}(\phi^{(n)}; D_{\text{Agree}(n)})}{a_{\text{f}}(\phi^{(n)}; D_{\text{Disagree}(n)})}
\]

where

\[
D_{f: \text{Agree}(n)} = \phi^* | y^* = y^{(n)}
\]

A subset of the projections of \(f\) on \(D\) whose output value is equal to \(y^{(n)}\) \((D_{\text{Agree}(n)}\) does not include \(\phi^{(n)}\))

\[
D_{f: \text{Disagree}(n)} = \phi^* | y^* \neq y^{(n)}
\]

A subset of the projections of \(f\) on \(D\) whose output value is not equal to \(y^{(n)}\)

In the following section, the advantages of the proposed conditional outlier approach based on the ROS score and its combination with discriminative projections (ROSDP) approach are demonstrated through experiments.

4. EXPERIMENTAL STUDY

To validate the correctness and effectiveness of the proposed methods, this section performs an experimental study with synthetic and real world datasets. Recall that both ROS and ROSDP are meta-methods in that they are designed to adopt and interplay with unconditional outlier detection methods. As reviewed and discussed in Sections
3.1.1 and 3.1.2, this work utilizes the local outlier factor (LOF) and one-class support vector machine (OCSVM) methods for experiments. In particular, the experiments conducted in this section compare the following methods:

- Local Outlier Factor on the Joint Space (LOF) [19] - This applies LOF (Equation (2)) to the joint space of all data attributes (both input and output)
- Ratio of Outlier Scores with LOF (ROS+LOF) - This evaluates the ratio–based outlier score (Equation (1)) with LOF calculated on the original input space
- Ratio of Outlier Scores on Discriminative Projection with LOF (ROSDP+LOF) - This evaluates the ratio–based outlier score with LOF on a discriminative projection space (Equation (7))
- One–Class Support Vector Machine on the Joint Space (OCSVM) [20] - This applies OCSVM (Section 3.1.2) to the joint space of all data attributes (both input and output)
- Ratio of Outlier Scores with OCSVM (ROS+OCSVM) - This evaluates the ratio–based outlier score (Equation (1)) with OCSVM applied to the original input space
- Ratio of Outlier Scores on Discriminative Projection with LOF (ROSDP+LOF) - This evaluates the ratio–based outlier score with OCSVM applied to a discriminative projection space (Equation (7))

For every instance of LOF (LOF, ROS+LOF, and ROSDP+LOF), we set the number of neighbors to \( k = 50 \); and use Mahalanobis distance to measure the distance between pairs of instances. To obtain data models/discriminative projection functions in ROSDP, we use \( L_2 \)-regularized logistic regression and choose the regularization coefficient using the internal cross validation.

4.1 Experiment Setup

The evaluation and comparisons of COD methods could be very challenging because outlier validation may be ambiguous and may require additional human feedback. For the purpose of the comparative evaluation, this work conducts experiments on simulated outliers as follows:

1. In each experiment, select 1.0% of instances uniformly at random (\textit{outlier ratio} = 1.0%)
2. For each selected instance, invert the output value (\( y_{\text{outlier}} = |y_{\text{outlier}} - 1| \))

The resulting outliers can be interpreted as contextually abnormal output signals, that could be considered as errors or mistakes. For example, in an annotated image dataset, the outliers would be perceived as images with incorrect labels.

It should be stressed out that all methods in below experiments (including their model building and outlier scoring stages) are run on data with simulated outliers. In other words, the models are not trained on the clean (i.e., outlier–free) data before they perform outlier detection on the data with simulated outliers. Such a setup is indeed impractical since in real world applications outliers in data are not a priori known nor cannot be excluded.

4.2 Evaluation Metrics

This work uses the Area Under the Precision–Recall Curve (AUPRC) as the primary performance measure. AUPRC summarizes the precision–recall tradeoff in all range of outlier scores. In the context of outlier detection, precision measures how much outlier calls are correct out of all outlier calls made, Recall measures how much outliers are correctly detected out of all outliers in data.

This work also evaluates the Area Under the Receiver Operating Characteristic Curve (AUROC). AUROC summarizes the specificity–recall tradeoff over all possible outlier detection thresholds. Although AUROC has been a widely used metric for outlier detection performance, it is worthwhile to note that the metric is not precisely reflecting the outlier detection performance as AUROC puts true negatives (normal instances that are identified to be normal) in evaluation. As normal (non–outlying)
data instances are usually dominant in a dataset, AUROC tends to overestimate the outlier detection performance.

For both AUPRC and AUROC, higher means better performance.

4.3 Experiments

4.3.1 Synthetic Data (Identification of Unusual Data Labels)

**Data** Four synthetic datasets are used for the evaluation of outlier detection performance. Each artificial dataset contains 2,000 instances of 2-dimensional continuous input \((X \in \mathbb{R}^2)\) and binary \((Y = 0, 1)\) output pair that form a unique shape as displayed in Fig. 2. In Fig. 2, the inputs (context) are the coordinates, while the binary outputs (responses) are denoted in shapes (0=blue circle, 1=orange cross). For convenience, the datasets are referred to as Chessboard (Fig. 2(a)), Donut (Fig. 2(b)), Dots (Fig. 2(c)), and Cross (Fig. 2(d)).

**Simulating Outliers** Conditional outliers are simulated by selecting 1.0% of the dataset instances and by inverting their output labels (recall Section 4.1). Given that each dataset contains two classes representing the membership, the simulated outliers can be interpreted as incorrect response in the output space (wrongful class assignments).

**Results** Tables 1 and 2 summarize the results on the four synthetic datasets. All the results are averaged over five simulation runs. The numbers shown in boldface in the tables indicate the best results (by paired t-test at \(a = 0.05\)) on each experiment set.

In terms of AUPRC (Table 1), ROS+LOF and ROSDP+LOF are the clear winners. Across all experiments, ROS+LOF and ROSDP+LOF detect conditional outliers with statistically superior performances. Compared to LOF, the base outlier

![Fig. 2](image-url)

*Fig. 2.* Synthetic datasets for experiments. The inputs (context) are the coordinates, while the binary outputs (responses) are denoted in shapes (0=blue circle, 1=orange cross). According to the shape, the datasets are referred to as: (a) Chessboard, (b) Donut, (c) Dots, and (d) Cross.

**Table 1. Area Under the Precision–Recall Curve (AUPRC)**

<table>
<thead>
<tr>
<th></th>
<th>AUPRC</th>
<th>LOF based</th>
<th>OCSVM based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ROS + LOF</td>
<td>ROSDP + LOF</td>
</tr>
<tr>
<td>Chessboard</td>
<td>0.429±0.016</td>
<td>0.945±0.018</td>
<td>0.938±0.020</td>
</tr>
<tr>
<td>Donut</td>
<td>0.131±0.022</td>
<td>0.464±0.060</td>
<td>0.463±0.059</td>
</tr>
<tr>
<td>Dots</td>
<td>0.098±0.016</td>
<td>0.376±0.051</td>
<td>0.395±0.051</td>
</tr>
<tr>
<td>Cross</td>
<td>0.145±0.021</td>
<td>0.575±0.067</td>
<td>0.559±0.068</td>
</tr>
</tbody>
</table>

Numbers shown in bold indicate the best results on each experiment set (by paired t-test at \(a=0.05\)).
Table 2. Area Under the Receiver Operating Characteristic Curve (AUROC)

<table>
<thead>
<tr>
<th>AUROC</th>
<th>LOF based</th>
<th>OCSVM based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOF</td>
<td>ROS + LOF</td>
</tr>
<tr>
<td>Chessboard</td>
<td>0.974±0.009</td>
<td>0.987±0.009</td>
</tr>
<tr>
<td>Donut</td>
<td>0.842±0.021</td>
<td>0.906±0.020</td>
</tr>
<tr>
<td>Dots</td>
<td>0.748±0.024</td>
<td>0.867±0.025</td>
</tr>
<tr>
<td>Cross</td>
<td>0.868±0.021</td>
<td>0.934±0.020</td>
</tr>
</tbody>
</table>

Numbers shown in bold indicate the best results on each experiment set (by paired t-test at α=0.05).

score estimator of the former, there is a huge improvement that highlights the usefulness of the proposed ratio-based approach. Although not as good as ROS+LOF and ROSDP+LOF, the ROS methods combined with OCSVM also produce strong results. On two of the synthetic datasets, ROS+OCSVM and ROSDP+OCSVM show superior performances. On the Donut and Dots datasets, however, the improvement is not as evident. This is because the input regions of each output value are not as separate as in the other dataset and, in turn, OCSVM applied to each class does not yield distinctive scores for the ratio-based estimation.

A similar performance pattern is observed in terms of AUROC (Table 2), while the performance difference between methods are not as apparent as with AUPRC. Across all four datasets, ROS+LOF and ROSDP+LOF always outperform the other methods, and ROS+OCSVM and ROSDP+OCSVM follow. On the other hand, the base estimators, LOF and OCSVM, applied to the joint space of input and output, do not produce meaningful results in terms of conditional outlier detection. Overall, the experimental results support the proposed framework for conditional outlier detection.

4.3.2 Real World Image Data (Identification of Incorrect Image Annotations)

**Data** The data used in this section is the MNIST dataset [16]. This dataset contains images of hand-written digits (as in Fig. 3), along with the ground truth labels telling what digits are scanned. In particular, six subsets of MNIST are used for the experiments: each subset includes 2,000 randomly sampled images of two pre-selected digits (1,000 images per digit). The pre-selected pairs of digits are:

- 0 and 6
- 1 and 7
- 2 and 3
- 3 and 5
- 4 and 8
- 6 and 9

Notice that the above digit pairs share a certain degree of similarity in their visual patterns and, hence, their distinction could be confusing. In each subset, a new binary output label is created according to given image, such that the label indicates the matching digit, e.g., in the first subset, output values 0 and 1 indicate 0 and 6, respectively.

**Simulating Outliers** Conditional outliers are simulated by selecting 1.0% of the dataset instances and by inverting their output labels (recall Section 4.1). Given that each dataset contains two classes of handwritings with subtle differences, the simulated outliers can be interpreted as errors or

![Fig. 3. The MNIST dataset [16].](image-url)
mistakes in image labels.

Results. Tables 3 and 4 present the conditional outlier detection results on the six digit pairs. All the results are averaged over five simulation runs. The numbers shown in boldface in the tables indicate the best results (by paired t-test at α = 0.05) on each experiment set.

The results in terms of AURPC (Table 3) demonstrate a similar performance pattern as in Tables 1 and 2. Namely, ROS+LOF and ROSDP+LOF show consistently outstanding performance, while ROS+OCSVM and ROSDP+OCSVM follow. Compared with each base outlier score (LOF and OCSVM), the advantages of ROS and ROSDP methods are clearly revealed through the results. In terms of AUROC, the results mostly agree with the rest, while ROSDP also records statistically superior performances. Albeit the improvement is not the most explicit, ROS+OCSVM also shows a consistent enhancement over the plain OCSVM method.

All in all, the experimental results reported in this section support the validity and usefulness of the ROS and ROSDP approaches in addressing the COD problem.

5. CONCLUSION

This work studied conditional outlier detection in the context of incorrect data label identification. Through the discussions, the definition of conditional outlier and its detection was clearly defined. Previous research addressing both conditional and unconditional outlier problems were also revisited. As the main contribution of this work, the ratio-based outlier scoring (ROS) approach and its variant, (ROSDP), were proposed. The propose

<table>
<thead>
<tr>
<th>Table 3, Area Under the Precision–Recall Curve (AUPRC)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AUPRC</strong></td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>0 and 6</td>
</tr>
<tr>
<td>1 and 7</td>
</tr>
<tr>
<td>2 and 3</td>
</tr>
<tr>
<td>3 and 5</td>
</tr>
<tr>
<td>3 and 8</td>
</tr>
<tr>
<td>8 and 9</td>
</tr>
</tbody>
</table>

Numbers shown in bold indicate the best results on each experiment set (by paired t-test at α=0.05).

<table>
<thead>
<tr>
<th>Table 4, Area Under the Receiver Operating Characteristic Curve (AUROC)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AUROC</strong></td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>0 and 6</td>
</tr>
<tr>
<td>1 and 7</td>
</tr>
<tr>
<td>2 and 3</td>
</tr>
<tr>
<td>3 and 5</td>
</tr>
<tr>
<td>3 and 8</td>
</tr>
<tr>
<td>8 and 9</td>
</tr>
</tbody>
</table>

Numbers shown in bold indicate the best results on each experiment set (by paired t-test at α=0.05).
solutions work by adopting existing (unconditional) outlier scores and are able to apply them to identify conditional outliers in data. The experiments on synthetic and real world image datasets demonstrated the benefits and advantages of the proposed approaches. The proposed solutions in this work have a unique merit that they can bridge the gap between unconditional and conditional outlier detection studies, which used to be pursued in orthogonal efforts. In the future, this work will extend to embrace a wider range of unconditional outlier scores and demonstrate its applicability to data from a variety of domains.

REFERENCE

[17] M. Hauskrecht, I. Batal, M. Valko, S. Viswes-


Charmgil Hong
received the B.S. degree in computer science and engineering from Handong Global University, Pohang, Korea, in 2010 and the Ph.D. degree in computer science from the University of Pittsburgh, Pittsburgh, PA, in 2017. He worked as a Postdoctoral Associate with the School of Computing and Information, University of Pittsburgh. Since 2019, he has been an Assistant Professor with the School of Computer Science and Electrical Engineering, Handong Global University, Korea. His research interests include machine learning and its applications to healthcare.