Multivariate Outlier Detection for Online Experimentation

Problems of Interest

- Develop a multivariate distribution-based outlier filter for a group of metrics
- Metrics: count data with excessive zeros
    - Session, clicks, page views, etcs
- Challenges:
    - The skewness of the data
    - Dependent structure in multivariate count data

1. Parametric-based Method

Type I multivariate zero-inflated generalized Poisson distribution

- In this study, the author extend the univariate ZIGP distribution to Type-I multivariate ZIGP distribution via stochastic representation
- Aim to model positively correlated multivariate zero-inflated count data with over-dispersion or under-dispersion

Definition 1. Let $Z \sim \text{Bernoulli}(1-\phi)$, $\mathbf{x} = (X_1, \ldots, X_m)^\top$, $X_i \sim \text{GP}(\lambda_i, \theta_i)$ for $i = 1, \ldots, m$, and $(Z, X_1, \ldots, X_m)$ are mutually independent. An m-dimensional discrete random vector $\mathbf{y} = (Y_1, \ldots, Y_m)^\top$ is said to have a Type I multivariate ZIGP distribution if

$$\mathbf{y} \stackrel{\mathrm{d}}{=} Z\,\mathbf{x} = \begin{cases} \mathbf{0}, & \text{with probability } \phi, \\ \mathbf{x}, & \text{with probability } 1-\phi, \end{cases} \tag{2.2}$$

where $\phi \in [0,1)$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top \in \mathbb{R}_+^m$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top$, $\max(-1, -\lambda_i/q_i) < \theta_i \leqslant 1$ and $q_i \geqslant 4$ is the largest positive integer for each $\lambda_i + \theta_i q_i > 0$ when $\theta_i < 0$. We write $\mathbf{y} \sim \text{ZIGP}_m^{(\mathrm{I})}(\phi, \boldsymbol{\lambda}, \boldsymbol{\theta})$ or $\mathbf{y} \sim \text{ZIGP}^{(\mathrm{I})}(\phi; \lambda_1, \ldots, \lambda_m, \theta_1, \ldots, \lambda_m)$ and call $\mathbf{x}$ the base vector of the $\mathbf{y}$. ¶

Likelihood-based Statistical Inference

MLEs via the MM algorithm

- Develop a MM algorithm with explicit expressions at each iteration through constructing a Q function to separate the parameter *\phi, \lambda and \theta*

$$\phi^{(t+1)} = \frac{n_0 \phi^{(t)}}{n \beta^{(t)}},$$

$$\lambda_i^{(t+1)} = \frac{n - n_0 - n_{i0} + \sum_{j \in \mathbb{J}_i} \frac{(y_{ij} - 1)\lambda_i^{(t)}}{\lambda_i^{(t)} + \theta_i^{(t)} y_{ij}}}{n - n\phi^{(t+1)}}$$

$$\theta_i^{(t+1)} = \frac{\sum_{j \in \mathbb{J}_i} \frac{\theta_i^{(t)} y_{ij}(y_{ij} - 1)}{\lambda_i^{(t)} + \theta_i^{(t)} y_{ij}}}{\sum_{j \in \mathbb{J}_i} y_{ij}},$$

Implementation

Github link:

https://github.com/linyu2295/Multivariate-Outlier-Detection-for-Online-Experimentation

Challenges

Hard to fit the different metrics and get the results consistently

2. Distance-based Outlier Detection

- Assumption:
    a. Normal objects have a dense neighborhood, thus the outlier is the one furthest from its neighbors.
- Advantages:
    a. Do NOT require to model the underlying probability distribution
- Challenge: scalability
    a. All pairwise distances computation are expensive, take $O(n^2)$ time

Methods for Outlier Detection

One-time sampling-based method

- Score function q:
    a. Assign a real-valued outlierness score to each object x
- Method:
    a. Randomly and independently sample a subset S(X) only once and for each object x define

$$q_{\mathrm{Sp}}(\boldsymbol{x}) := \min_{\boldsymbol{x}' \in S(\mathcal{X})} d(\boldsymbol{x}, \boldsymbol{x}')$$

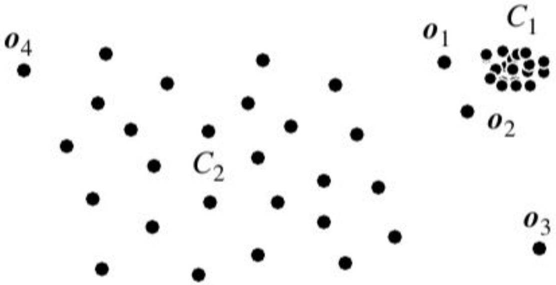
- Evaluation criterion:
    a. Area under the precision-recall curve (AUPRC, average precision)

Advantages

- Scalable
  a. The time complexity is linear in the number of data points
- Effective
  a. It is empirically shown to be the most effective on average among existing distance-based outlier detection methods
- Easy to use
  a. Require only one parameter, the number of samples ***s*** and small sample size (default value is 20)

3. Density-based Outlier Detection

- Assumption:
  a. An object is an outlier if its density relatively much lower than that of its neighbors
- Advantages:
  a. Objects may be considered outliers with respect to their local neighborhoods, rather than with respect to the global data distribution



Methods for Outlier Detection

LOF-based method

- Reachability distance measure
    a. Reachdist_k (o, o’) = max[dist_k (o), dist (o, o’)], not symmetric
    b. k: smoothing effect, and it specifies the minimum neighborhood to be examined to determine the local density of an object
- Local reachability density of an object o:
    a. Lrd_k (o) = ||N_k (o)|| / sum_{o’ in N_k (o)} reachdist_k (o, o’)
- Local outlier factor:
    a. $$LOF_k(\mathbf{o}) = \frac{\sum_{\mathbf{o}' \in N_k(\mathbf{o})} \frac{lrd_k(\mathbf{o}')}{lrd_k(\mathbf{o})}}{\|N_k(\mathbf{o})\|} = \sum_{\mathbf{o}' \in N_k(\mathbf{o})} lrd_k(\mathbf{o}') \cdot \sum_{\mathbf{o}' \in N_k(\mathbf{o})} reachdist_k(\mathbf{o}' \leftarrow \mathbf{o}).$$

Methods for Outlier Detection

LOF-based method

- Local outlier factor:
    a. The average of the ratio of the local reachability density of o and those of o's k-nearest neighbors.
    b. The lower local reachability density of o and the higher the local reachability densities of the k-nearest neighbors of o, the higher the LOF value is.
    c. A local outlier has relatively low local density compared to the local densities of its k-nearest neighbors.

Challenges

Due to the pairwise distance calculation among all the data points, it is impossible to implement this method at large-scale company data

4. Isolation Forest for Outlier Detection

- Assumption:
    a. Anomalies are ‘few and different’, which make them more susceptible to isolation than normal points.
- Advantages over model-based, distance-based, and density-based methods:
    a. Utilizes no distance or density measures to detect anomalies, which eliminates major computational cost
    b. Linear time complexity with a low constant and a low memory requirement
    c. Capable to scale up to handle extremely large data size and high-dimensional problems with a large number of irrelevant attributes

Methods for Outlier Detection

Isolation Trees

- Idea:
    a. Anomalies are more susceptible to isolation under random partitioning (partitions are generated by randomly selecting an attribute and then randomly selecting a split value of the selected attribute)
    b. The number of partitions required to isolate a point = the path length from the root node to a leaf node
- Characteristic:
    a. Identifies anomalies as points having shorter path lengths
    b. Has multiple trees acting as 'experts' to target different anomalies
    c. Build a partial model by sub-sampling which incidentally alleviates the effects of swamping and masking
        i. Better isolate examples of anomalies
        ii. Each isolation tree can be specialised, as each sub-sample includes different set of anomalies or even no anomaly

Challenges

This method requires to pre-specify the ratio of outliers in the dataset, it is hard to implement it in practice.