Identification of Outliers through Clustering and Semi-supervised Learning
In All Sky Automated Survey Data Set

Sharmodeep Bhattacharyya¹, Joseph W. Richards¹,², John Rice¹, Dan L. Starr², Nathaniel R. Butler² and Joshua S. Bloom²
¹Department of Statistics, University of California, Berkeley
²Department of Astronomy, University of California, Berkeley

Data

- Recently there has been a huge surge of data in astronomy, making outlier or novelty detection a crucial step in analyzing these data.
- We have 25-class labeled light curves of 1542 well-studied variable light sources from Hipparcos and Optical Gravitational Lensing Experiment (OGLE) surveys.
- A set of 64 periodic and non-periodic features are extracted from the light curves using the method in Richards et. al. [1].
- We consider a set of unlabeled light curves of ~ 50,000 variable light sources from All Sky Automated Survey (ASAS). We select 11375 low noise variable light sources for our analysis.
- We compute the same set of 64 features for these 11375 light curves.

Goal

Identify interesting light curves in the ASAS data set. The goal is to identify outlier data points in the ASAS data set that can not be categorized into any of the 25 classes of the labeled data set with high confidence.

Statistical Framework and Method

- We have \( (X_i, Y_i), \ldots, (X_n, Y_n) \) (where, \( n = 1542 \)) as the labeled training data, with, \( X_i \in \mathbb{R}^p (p = 64) \) as the features and \( Y_i \in \mathbb{R} \) as the labels.
- We have \( Z_1, \ldots, Z_m \) (where, \( m = 11,375 \)) as the test data, where, \( Z_i \in \mathbb{R}^p \).
- We call a small set of data points ‘outlier’, if it is away from a large set of data points in terms of a specific metric. Also, we want this ‘outlier’ to have some peculiarity. For figures 2, 3 and 5 unfolded, where as, for others folded.

Metric

The first issue to consider in a clustering approach is the metric to use. We consider two different metrics.

- Weighted Euclidean Metric: Metric \( d_w \), defined by,
  \[
d_w(x_i, x_j) = \sum_{k=1}^{p} w_k (X_{ik} - X_{jk})^2
  \]
where, weights \( w_k (k = 1, \ldots, p) \) being the Random Forest Importance measure of each feature. An example of such weights is given in figure 1(a).

- Proximity Metric: The proximity matrix is a similarity measure computed by fitting an unsupervised model to the unlabeled data. An example of proximity matrix structure is given in figure 1(b).

Semi-supervised Approach

- The method has parameters \( (\alpha, L_1, L_2, C) \).
- We call a cluster ‘\( \alpha \)-outlier’ if it contains less than \( 100(\alpha)% \) of the data points from which it has been separated at a given iteration.
- Parameters \( L_1, L_2, C \) and \( K \) control how to determine, whether an outlier is ‘interesting’ or not. If the data points of outlier clusters can be labeled easily, then, usually we do not consider those outliers as ‘interesting’.

Here is the method -

Step 1 We consider the scaled version of feature space of both labeled and unlabeled data set together, that is, consider \( X = (X_1, \ldots, X_n, Z_1, \ldots, Z_m) \).

Step 2 Now, cluster the \( (n + m) \) data using hierarchical divisive clustering. At any iteration, cluster \( S \) divides into \( S_1 \) and \( S_2 \).

Step 3 At each iteration, if \( \min(|S_1|, |S_2|) \leq |S| \) or \( \min(|S_1|, |S_2|) \leq 10 \), we flag the smaller cluster. We stop after a large number of iterations, say \( K \).

Step 4 Now consider each flagged cluster \( S \) and the set of labeled data points in \( S \) be \( S_L \). If \( L_1 |S_L| \leq |S_1| \), then remove flag of \( S \). If \( L_2 |S_L| \leq |S_2| \) \( (L_2 < L_1) \), but, more than \( C|S_1| \) has same labels, then also, remove the flag of the cluster.

Step 5 Consider all the data points from the unlabeled data set in the flagged clusters as the ‘interesting outliers’.

Results

- We apply our method to the ASAS data set, broken into 4 parts, for computational constraints.
- We take \( \alpha = 0.01, L_1 = 0.5, L_2 = 0.25, C = 0.75 \) and \( K = 200 \).
- Below, we present the light curves of few ‘interesting’ outliers, with a note on their peculiarity. For figures 2, 3 and 5 unfolded, where as, for others folded light curves are shown.

Summary

So, we see that through our hierarchical semi-supervised approach, we identify some interesting light sources. However, still our method need improvement, possibly by introducing better semi-supervised learning, as it still picks up many ‘un-interesting’ outliers. Also, we have to perform a more thorough follow up analysis on ‘interesting’ outlier light sources.

References