Astronomical Data Analysis Software and Systems XXX
ASP Conference Series, Vol. 532
J. E. Ruiz, F. Pierfederici, and P. Teuben, eds.

Machine-Assisted Discovery Through Identification and Explanation of Anomalies in Astronomical Surveys

Kiri L. Wagstaff, Eric Huff, and Umaa Rebbapragada

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA; `kiri.wagstaff@jpl.nasa.gov`

Abstract. Data volumes in modern astronomical surveys are large, and human attention is comparatively scarce. The most interesting sources are rare and may therefore go permanently buried and unknown in large archives. Many science goals from planned sky surveys (e.g., Roman, SPHEREx, and Euclid) require exquisitely precise measurements taken over billions of galaxies and stars. Existing validation techniques appear unlikely to scale to the next generation of large sky surveys. We propose the use of machine learning to identify, group, and explain anomalies within very large data sets. The goal is to quickly distinguish erroneous measurements and expected patterns in the data from sources and statistical correlations with true astrophysical origins. We illustrate the process of identifying and explaining anomalies in a study conducted on sources observed by the Dark Energy Survey. We found that 96% of automatically identified outliers in a subset of 11M sources were likewise discarded by humans. In addition, several unusual objects led to follow-up spectral observations with the Palomar Observatory. We hypothesize that this discovery process, when applied to other large-scale sky survey data sets, can result in improved science yield and catalog validation.

1. Introduction

Machine learning methods can automate tedious data analysis and classification processes by learning to replicate human decision-making. Can they also help make new discoveries? Current astronomical surveys are generating measurements for billions of sources (e.g., WISE, Gaia, and the Dark Energy Survey), and future projects such as Euclid, the Roman Space Telescope, and the Rubin Observatory will make measurements for at least an order of magnitude more sources. The immense anticipated data volumes pose a challenge for data validation and for mining the catalog for new discoveries. Data validation is the process of filtering out problematic observations and artifacts, and current manual review procedures will not keep up. Manual review also led to discoveries such as quasars (Schmidt 1963), radio pulsars (Hewish et al. 1968), so-called 'green pea' galaxies (Cardamone et al. 2009), and cosmic gamma-ray bursts (Klebesadel et al. 1973). Without a scalable diagnostic process that is functionally similar to human inspection, we will lose this mode of discovery.

Machine learning algorithms now see wide use in large astronomical data sets (Way et al. 2012). They are being used to classify stars and galaxies in large catalogs (Weir et al. 1995), classify supernovae from light curves (Charnock & Moss 2017), detect evidence of exoplanets in light curves (Pearson et al. 2017; Shallue & Vanderburg 2018), and highlight unusual observations or outliers that merit additional study (Nun et al. 2014, 2016; Giles & Walkowicz 2019).

We propose the use of machine learning methods to assist validation and discovery efforts within large sky survey data sets. First, we identify anomalies within the data set. For data sets with billions of sources, the list of anomalous measurements is itself likely to be so large as to defy manual review. Therefore, we employ automated clustering methods to identify groups anomalies that can be reviewed collectively. To accelerate review and interpretation, we use

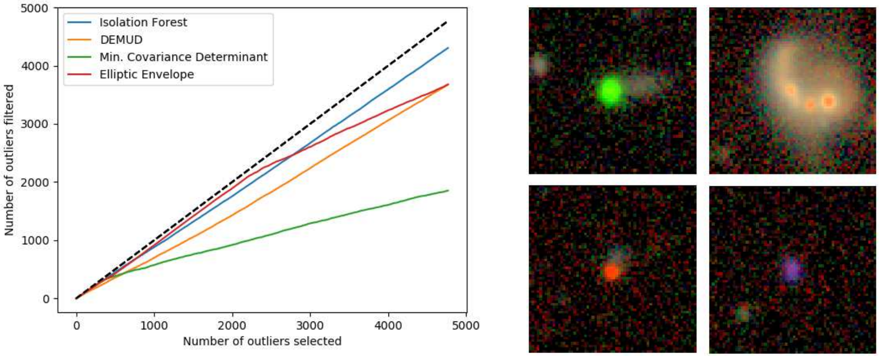

Figure 1. (Left) Comparison of anomaly detection methods against human validation. Dashed black line would show perfect agreement. (Right) Anomalies of scientific interest: likely supernovae (left column), strong lens (top right), and compact star-forming galaxy (lower right).

methods to automatically generate explanations that describe each group. We illustrate this process by applying it to data from the Dark Energy Survey (DES), a cosmological study that has collected multi-wavelength observations of millions of galaxies. The success of machine-assisted anomaly analysis for the DES data encourages the application of the same approach to new and upcoming sky surveys, with great potential for time savings and new discoveries.

2. Anomaly Detection, Grouping, and Review Methods

We investigated four anomaly detection methods. The Isolation Forest (Liu et al. 2008) identifies a set of global anomalies in terms of how separable each item is from the full data set. The algorithm builds an ensemble of random binary trees where each node employs a randomly chosen feature and threshold. Anomalies are those items that follow very short paths from root to leaf (i.e., items that are quickly isolated). The DEMUD (Wagstaff et al. 2013) method instead seeks to find a set of *diverse* anomalies by incrementally growing a model of known anomalies and repeatedly selecting the item most different from what has already been chosen. DEMUD uses a singular value decomposition (SVD) to model previous anomalies; new items with unusual properties will have high reconstruction error. The Elliptic Envelope fits a multivariate Gaussian model to the data and ranks items by their negative Mahalanobis distance from the mean. The Minimum Covariance Determinant (MCD) estimator refines this approach by using a robust (outlier-tolerant) estimate of the multivariate mean and standard deviation (Rousseeuw & Hubert 2017).

We compared all methods on a data set that consists of 11.9M objects observed by the Dark Energy Survey (DES). The first version of this catalog, released in June 2018, incorporated only cuts on signal-to-noise, resolution, masks against known detector anomalies and data quality indicators, and the automated data quality flags produced during processing. In December 2019, the full catalog was released after 18 months of extensive manual vetting. Therefore, we were able to use the second version of the catalog as a validation set for anomaly detection on the first version. We generated a list of the top anomalies according to each algorithm and tallied how many were also rejected by the independent human review. Figure 1 (left) shows all four algorithms assessed on their first ~5000 selections. The Isolation Forest achieved the best agreement with human filtering. In total, the Isolation Forest selected 14,491 anomalies, of which 13,953 (96%) were also filtered by humans. Therefore, we chose the Isolation Forest as our anomaly detector for this investigation.

Grouping Anomalies for Fast Review. Given the global ranking of anomalies produced by the Isolation Forest, there may be redundancy within the ranked list in which many similar anomalies appear sequentially. To reduce the effort required to review the anomalies, we use data clustering to group anomalies into categories, similar in motivation to the X-PACS method (Macha & Akoglu 2018) but leveraging the output of the Isolation Forest to inform the groupings. In this study, we used the k-means clustering algorithm (MacQueen 1967) to group anomalies into $k = 50$ clusters.

Review of Anomaly Groups. Each anomaly cluster can be inspected to determine whether it consists of data or measurement errors, and therefore should be filtered out, or if it shows evidence of a new category of sky objects. We developed a web-based Anomaly Explorer to facilitate the review process that shows each anomaly as well as observations of the same sky location in other surveys such as GALEX and WISE. We examined the 538 Isolation Forest anomalies that were not filtered in the version 2 data set and found that 38% were the result of measurement error (e.g., the galaxy model fit failed), 33% were due to data corruption (e.g., satellite tracks, moving objects, and transients), just 9% were normal objects (false positives), and 20% were objects of potential scientific interest (e.g., supernovae, strong lensing, and sources with unusual spectra that merit follow-up).

We also tested a framework that uses Causal Graphical Models (Pearl 1995) to generate an explanation for each anomaly group. We expanded the feature space to include measures of data quality, then used the PC algorithm (Spirtes et al. 2000) to generate a directed acyclic graph to highlight which features best explained the anomaly status of a cluster. Clusters with graphs that highlighted data quality features tended to be non-astrophysical anomalies, while graphs where anomaly status was connected to a measured feature (e.g., brightness in a single band) tended to be astrophysical.

3. Results: Anomalous Galaxies within the Dark Energy Survey

Figure 1 (right) shows four anomalies of scientific interest. The first column shows two sources that only appear in one band (green, red) which suggests that the source brightened only during the time that band was observed. Since they are also overlaid on diffuse whitish objects, we interpret these transient bright sources as supernovae within distant galaxies. The upper right observation with multiple sources may be a strong lensing system. The lower right (purple) source has an anomalous SED, corresponding to a strong r-band excess. Follow-up optical spectroscopy of this source with Palomar Observatory (thanks to Daniel Stern) showed that the r-band excess was indeed astrophysical, arising from extreme [OIII] emission in a compact star-forming galaxy at redshift ~0.35. We have identified a small population of similar anomalous emission-line sources with high apparent star formation rates, and analysis is ongoing.

4. Conclusions

We found that 96% of the DES anomalies identified via machine learning were identical to those found by humans, and the remaining 4% of anomalies yielded new kinds of problematic observations that should have been removed as well as some unusual objects with genuinely interesting properties. In other words, an uncalibrated outlier detection algorithm correctly identified the bulk of the sources removed by exhaustive human validation, while also identifying a substantial additional population of sources that were missed during this enormous work effort. We are not proposing that this procedure replace the normal work of human validation on survey data. However, evidence suggests that this process could improve the quality of science catalogs while dramatically reducing the work required to produce them.

Acknowledgments. The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Copyright 2020 California Institute of Technology. U.S. Government sponsorship acknowledged.

References

Cardamone, C., Schawinski, K., Sarzi, M., Bamford, S. P., Bennert, N., Urry, C. M., Lintott, C., Keel, W. C., Parejko, J., Nichol, R. C., Thomas, D., Andreescu, D., Murray, P., Raddick, M. J., Slosar, A., Szalay, A., & Vandenberg, J. 2009, 399, 1191

Charnock, T., & Moss, A. 2017, The Astrophysical Journal Letters, 837, L28

Giles, D., & Walkowicz, L. 2019, 484, 834

Hewish, A., Bell, J., Pilkington, P., & Scott, R. 1968, Nat, 217, 709

Klebesadel, R. W., Strong, I. B., & Olson, R. A. 1973, The Astrophysical Journal Letters, 182, L85+

Liu, F. T., Ting, K. M., & Zhou, Z.-H. 2008, in Proceedings of the IEEE International Conference on Data Mining, 413

Macha, M., & Akoglu, L. 2018, Data Mining and Knowledge Discovery, 32, 1444

MacQueen, J. B. 1967, in Proceedings of the Fifth Symposium on Math, Statistics, and Probability (Berkeley, CA: University of California Press), vol. 1, 281

Nun, I., Pichara, K., Protopapas, P., & Kim, D.-W. 2014, The Astrophysical Journal, 793, 23

Nun, I., Protopapas, P., Sim, B., & Chen, W. 2016, The Astronomical Journal, 152, 71

Pearl, J. 1995, Biometrika, 82, 669

Pearson, K. A., Palafox, L., & Griffith, C. A. 2017, 474, 478

Rousseeuw, P. J., & Hubert, M. 2017, WIREs Data Mining and Knowledge Discovery, 8, e1236

Schmidt, M. 1963, Nat, 197, 1040

Shallue, C. J., & Vanderburg, A. 2018, The Astronomical Journal, 155, 94

Spirtes, P., Glymour, C., & Scheines, R. 2000, Causation, Prediction and Search (MIT Press), 2nd ed.

Wagstaff, K. L., Lanza, N. L., Thompson, D. R., Dietterich, T. G., & Gilmore, M. S. 2013, in Proceedings of the Twenty-Seventh Conference on Artificial Intelligence, 905

Way, M. J., Scargle, J. D., Ali, K. M., & Srivastava, A. N. 2012, Advances in Machine Learning and Data Mining for Astronomy (Chapman & Hall/CRC), 1st ed.

Weir, N., Fayyad, U. M., & Djorgovski, S. 1995, The Astronomical Journal, 109, 2401