AstronomicAL: an interactive dashboard for visualisation, integration and classification of data with Active Learning

Grant Stevens[1], Sotiria Fotopoulou[2], Malcolm N. Bremer[2], and Oliver Ray[1]

1 Department of Computer Science, Merchant Venturers Building, University of Bristol, Woodland Road, Bristol, BS8 1UB **2** School of Physics, HH Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol, BS8 1TL

Summary

AstronomicAL is a human-in-the-loop interactive labelling and training dashboard that allows users to create reliable datasets and robust classifiers using active learning. This technique prioritises data that offer high information gain, leading to improved performance using substantially less data. The system allows users to visualise and integrate data from different sources and deal with incorrect or missing labels and imbalanced class sizes. AstronomicAL enables experts to visualise domain-specific plots and key information relating both to broader context and details of a point of interest drawn from a variety of data sources, ensuring reliable labels. In addition, AstronomicAL provides functionality to explore all aspects of the training process, including custom models and query strategies. This makes the software a tool for experimenting with both domain-specific classifications and more general-purpose machine learning strategies. We illustrate using the system with an astronomical dataset due to the field's immediate need; however, AstronomicAL has been designed for datasets from any discipline. Finally, by exporting a simple configuration file, entire layouts, models, and assigned labels can be shared with the community. This allows for complete transparency and ensures that the process of reproducing results is effortless.
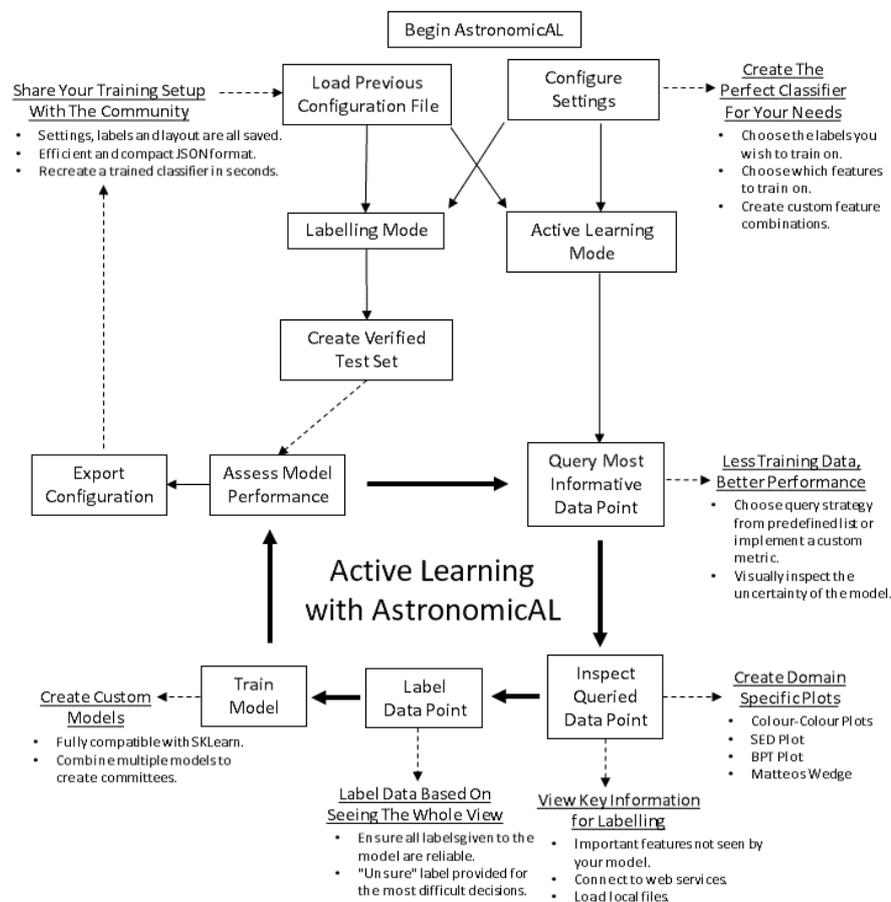
Figure 1: AstronomicAL workflow from the perspective of an astronomy user. Each part of the workflow improves the reliability of labels, leading to enhanced performance of any classifiers produced. AstronomicAL allows the user to tailor this workflow specifically to their domain with its modular and extensible design.

Statement of Need

Active learning has proven to be an effective method in machine learning over the last few decades (Baldridge & Palmer, 2009; Gal et al., 2017; Settles & Craven, 2008). In more recent years, there has been a growing adoption in all domains of scientific research (Eisenstein, 2020; Walmsley et al., 2020). This is evident by the increasing popularity of machine learning frameworks offering an active learning element[1]. However, the purpose of these frameworks is to allow users to integrate active learning into their machine learning workflows. Any additional functionality such as data exploration and interactive labelling is left to the user to implement. Due to each domain's varying requirements, researchers will often need multiple specialised but disjoint tools that hinder effective and efficient interactive labelling. When data from multiple sources is required, presenting all the information in one system is essential. Without a tool that combines these key features, a significant barrier exists to the adoption and effectiveness of active learning in research. AstronomicAL's goal is to remove this barrier to allow more researchers to access the benefits of active learning.

Active learning (Settles, 2012) removes the requirement for large amounts of labelled training data whilst still producing high accuracy models. This is extremely important as with ever-

---

[1]https://github.com/topics/active-learning

growing datasets; it is becoming impossible to manually inspect and verify ground truth used to train machine learning systems. The reliability of the training data limits the performance of any supervised learning model, so consistent classifications become more problematic as data sizes increase. The problem is exacerbated when a dataset does not contain any labelled data, preventing supervised learning techniques entirely. AstronomicAL has been developed to tackle these issues head-on and provide a solution for any large scientific dataset.

It is common for active learning to query areas of high uncertainty; these are often in the boundaries between classes where the expert's knowledge is required. To facilitate this human-in-the-loop process, AstronomicAL provides users with the functionality to fully explore each data point chosen. This allows them to inject their domain expertise directly into the training process, ensuring that assigned labels are both accurate and reliable.

AstronomicAL has been extensively validated on astronomy datasets. These are highly representative of the issues that we anticipate will be found in other domains for which the tool is designed to be easily customisable. Such issues include the volume of data (millions of sources per survey), vastly imbalanced classes and ambiguous class definitions leading to inconsistent labelling. AstronomicAL has been developed to be sufficiently general for any tabular data and can be customised for any domain. For example, we provide the functionality for data fusion of catalogued data and online cutout services for astronomical datasets.

Using its modular and extensible design, researchers can quickly adapt AstronomicAL for their research to allow for domain-specific plots, novel query strategies, and improved models. Furthermore, there is no requirement to be well-versed in the underlying libraries that the software uses. This is due to large parts of the complexity being abstracted whilst allowing more experienced users to access full customisability.

As the software runs entirely locally on the user's system, AstronomicAL provides a private space to experiment whilst providing a public mechanism to share results. By sharing only the configuration file, users remain in charge of distributing their potentially sensitive data, enabling collaboration whilst respecting privacy. The full workflow for AstronomicAL is shown in Figure 1.
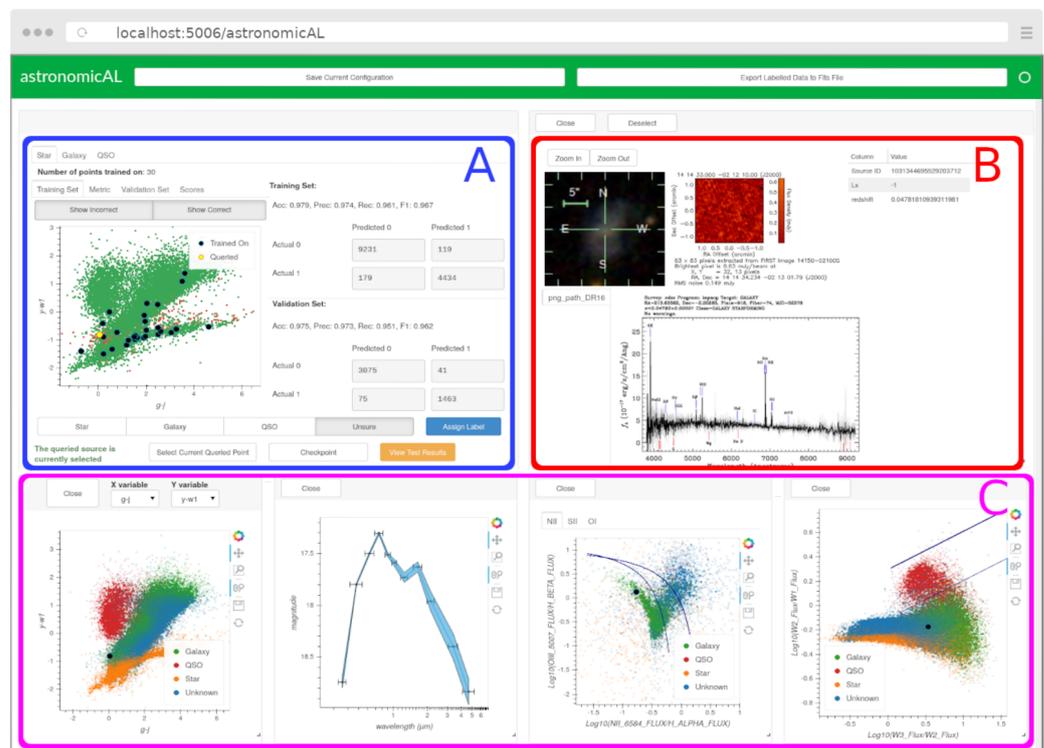
Figure 2: The dashboard is set up in 3 main areas. A) Active learning and labelling panel which controls the machine learning functionality. Users can train separate models for each class and see how each performs as additional points are added to the model. B) User-defined information shown for the currently active data point, such as specific columns, local images and online cutout services. C) Customisable domain-specific plots that can render millions of data points.

Active Learning and Classification

Active learning utilises methods for predicting which part of the search space of unseen examples would be most likely to improve classification accuracy. By querying these key examples and presenting them directly to the user for inspection and labelling, active learning removes the reliance on non-verified data. In the optimal case, it provides users with simple classifications that have a high impact on model performance, leading to a dramatic reduction in the number of data points required for training a classifier.

By making use of the modAL (Danka & Horvath, 2018) active learning framework, AstronomicAL allows users to take full advantage of active learning techniques, leading to improved classifiers. Scikit-Learn (Pedregosa et al., 2011) models are fully supported in AstronomicAL, and the software will be extended to support Pytorch (Paszke et al., 2019) and Tensorflow (Abadi et al., 2015) in the future.

Astronomy datasets are frequently the result of the fusion of multiple heterogeneous data sources. Often in machine learning, only information that is available across all data points is usable during training. However, in astronomy, where *any* available information could be critical in determining an object's true class, it must be used whenever this information is available. With the flexibility to choose which information is essential for classification, AstronomicAL ensures that users are presented with this critical information whenever available - directly leading to improved confidence and justifiability for any assigned labels. Although this extra information remains unseen by the model, by using AstronomicAL, we inject this information into the training process by labelling the data based on all the available information

rather than just the features the model sees.

In Figure 2 we show an example of a classification task using an extract of the astronomical dataset used in Fotopoulou & Paltani (2018). Figure 2 (A) shows the benefits of active learning where with only 30 data points given as training data to the model, we can classify Stars with 97.5% accuracy. Figure 2 (B) shows the extra information chosen by the user to improve confidence in the assigned labels. This information can be retrieved from web services or local files. In this example, we are showing the SDSS[2] (Ahumada et al., 2020) and FIRST[3] (Becker et al., 1995) cutout services. Figure 2 (C) shows how domain-specific plots can be created and visualised to gain extra insight into the data, aiding the researcher in their ability and confidence in labelling data points. Figure 3 shows the model performance plots available to the user for each classifier being trained. These include training and validation correctness plots, tracking of performance scores as new data points are added, and a visualisation of the model's confidence of each data point.
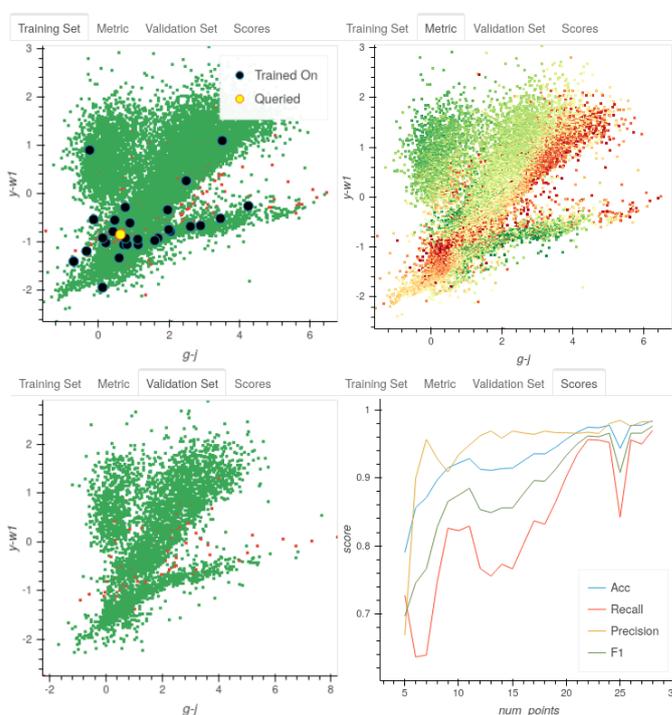

Figure 3: AstronomicAL allows users to view all key information during training. Top Left: Training set showing points that have been trained on, current queried point and correct and incorrect predictions. Top Right: Visualisation of the model's confidence of each data point. Bottom Left: Validation set showing correct and incorrect predictions of the model. Bottom Right: The performance scores of the model as new points are added to the training set.

Interactivity

Using the Panel dashboard library (Rudiger, Artusi, et al., 2021), paired with Holoviews (Rudiger, Stevens, et al., 2021), Bokeh (Bokeh Development Team, 2018) and Datashader (Bednar et al., 2020) visualisation libraries, AstronomicAL provides users with interactive plots, enabling zooming and panning of millions of data points in real-time whilst also giving the ability to rearrange and resize plots dynamically to optimise screen layout.

[2] http://skyserver.sdss.org/dr16/en/help/docs/api.aspx#imgcutout
[3] https://third.ucllnl.org/cgi-bin/firstcutout

Plots are also interconnected, with any updates projected to all panels simultaneously, allowing analysis of how the properties of a particular point fit in the context of the whole dataset. This enables researchers to gain insight and identify trends - crucial for reliable labelling.

Figure 4: AstronomicAL enables the creation of curated test sets. The sample is selected according to user-defined criteria and verified visually. This test set can then be used during the active learning training process to ensure that the model will be sufficiently generalisable to future data.

The interactive labelling functionality is not limited to only the training stage; as shown in Figure 4, users can curate a labelled test set to sufficiently demonstrate the validity and generalisability of their model. Furthermore, to facilitate and encourage only assigning labels that the user trusts, AstronomicAL allows users to mark any example as *unsure*. Such examples are excluded from the training set, ensuring that all training data are of high quality.

In summary, all design decisions have been in response to user feedback and were explicitly tailored to improve the visualisation of data, irrespective of the data's specific nature, to ensure that AstronomicAL is a tool for any discipline.

Acknowledgements

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. https://doi.org/10.5281/zenodo.4724125

Ahumada, R., Prieto, C. A., Almeida, A., Anders, F., Anderson, S. F., Andrews, B. H., Anguiano, B., Arcodia, R., Armengaud, E., & Aubert, M. (2020). The 16th Data Release

of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. *249*(1), 3. https://doi.org/10.3847/1538-4365/ab929e

Baldridge, J., & Palmer, A. (2009). How well does active learning actually work? Time-based evaluation of cost-reduction strategies for language documentation. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 296–305.

Becker, R. H., White, R. L., & Helfand, D. J. (1995). The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters. *450*, 559. https://doi.org/10.1086/176166

Bednar, J. A., Crail, J., Crist-Harif, J., Rudiger, P., Brener, G., B, C., Mease, J., Signell, J., Collins, B., Stevens, J.-L., Bird, S., kbowen, maihde, Thorve, A., Ahmadia, A., Jr, B. A. B., Brandt, C. H., Tolboom, C., G., E., … narendramukherjee. (2020). *Holoviz/datashader: Version 0.11.1* (Version v0.11.1) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3987379

Bokeh Development Team. (2018). *Bokeh: Python library for interactive visualization*. https://bokeh.pydata.org/en/latest/

Danka, T., & Horvath, P. (2018). *ModAL: A modular active learning framework for Python*. https://github.com/modAL-python/modAL

Eisenstein, M. (2020). Active machine learning helps drug hunters tackle biology. *Nature Biotechnology*, *38*(5), 512–515. https://doi.org/10.1038/s41587-020-0521-4

Fotopoulou, S., & Paltani, S. (2018). CPz: Classification-aided photometric-redshift estimation. *619*, A14. https://doi.org/10.1051/0004-6361/201730763

Gal, Y., Islam, R., & Ghahramani, Z. (2017). Deep bayesian active learning with image data. *International Conference on Machine Learning*, 1183–1192.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rudiger, P., Artusi, X., Bednar, J. A., Madsen, M. S., B, C., Signell, J., Stevens, J.-L., Liquet, M., Hoxbro, Mease, J., Andrew, Arne, Paprocki, M., kbowen, Jung, E., Amanieu, H.-Y., Winkelmann, J., Talirz, L., A, P., … Thorve, A. (2021). *Holoviz/panel: Version 0.11.0* (Version v0.11.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.4573728

Rudiger, P., Stevens, J.-L., Bednar, J. A., Nijholt, B., Mease, J., Andrew, B, C., Randelhoff, A., Tenner, V., maxalbert, Kaiser, M., ea42gh, Samuels, J., stonebig, Pevey, K., LB, F., Tolmie, A., Stephan, D., Hoxbro, … Barth, L. (2021). *Holoviz/holoviews: Version 1.14.2* (Version v1.14.2) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.4581995

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(1), 1–114. https://doi.org/10.2200/S00429ED1V01Y201207AIM018

Settles, B., & Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 1070–1079. https://doi.org/10.3115/1613715.1613855

Walmsley, M., Smith, L., Lintott, C., Gal, Y., Bamford, S., Dickinson, H., Fortson, L., Kruk, S., Masters, K., Scarlata, C., & others. (2020). Galaxy zoo: Probabilistic morphology through bayesian CNNs and active learning. *Monthly Notices of the Royal Astronomical Society*, *491*(2), 1554–1574. https://doi.org/10.1093/mnras/stz2816