Publications of the Astronomical Society of Australia

The Role of Citizen Science in Future Surveys in Astrophysics

Abstract

...

Keywords: keyword1 – keyword2 – keyword3 – keyword4 – keyword5

1 INTRODUCTION

Citizen science in the form of distributed data analysis has made significant contributions to astronomical science through projects such as Galaxy Zoo which allow many hundreds of thousands of people to make contributions. As astronomers continue to develop surveys which will produce ever growing volumes of (often) publicly accessible data, it seems likely that further opportunities for such collaboration will occur, though the increasing sophistication of machine learning and the need in some cases for rapid classification may well requiring increasingly sophisticated frameworks to be put in place. This paper is the result of discussions at a workshop held at St Catherine's College in Oxford from April 15th-17th 2015, hosted by the Zooniverse and sponsored by the Kavli Foundation, at which participants sought to explore the possibilities for citizen science with the crop of surveys currently under development.

2 THE PROBLEM: SURVEYS AND BIG DATA

3 ZOONIVERSE ARCHITECTURE

4 REAL TIME CLASSIFICATION

What 'real time' means depends on the science case. Radio transients observed within minutes of discovery have been found to be past their peak brightness, supernovae may need follow-up in days whereas surveys such as DES think of follow-up on timescales of months in deciding which objects should be targeted by fibres. However, for most of these use cases then *classification* is still urgent so that decisions can be taken, even if action can wait.

It is useful to think about problems involving streams of data for classification as being divided by their cadence. Projects can usefully be divided into the following categories :

- **Batch projects:** Provide data on, for example, monthly timescales. Examples include projects using the main (annual) LSST data release.
- **Sparse projects:** Alerts at irregular intervals when intervention is required. Examples include supernova identification with pre-filtering.
- **Continuous projects:** These projects would require substantial datasets to be reviewed, with enough data to keep volunteers busy all the time (rather than requiring a small number of actions after an alert).
- **Campaign projects:** Similar to continuous projects, but running for a short while; an example was the Snapshot Supernova project which reviewed Skymapper data while PESSTO follow-up was available.

In all cases, the common problem is the identification of several different categories of interesting object alongside a variety of artefacts; we essentially face a filtering problem. This filtering is a classification problem, but the performance of classification is also effected by the bias of astronomers affected by time constraints. A striking example was the recent discovery of superluminous supernovae, of which ~ 10 are known. These objects are identified by their long decay times, and have been found in archival data in SDSS, but were not recognised as a category of interest until recently, despite many searches for transients in these data sets.

A similar story emerges from the identification, by citizen scientists, of a set of extreme star forming dwarf galaxies known as the 'Green Peas'; identified by visual inspection in the first instance, they can be identified through judicious cuts in colour/colour space. The citizen science task may be less repeated classification

but rather the identification of categories of interest. Machine learning can then be used to 'amplify' categorization provided by inspection in a series of filters or, perhaps, a random forest populated by classifiers.

An example project could proceed by starting with a simple set of questions about expected common object types. As classification proceeds we should expect the performance of algorithms to improve as a large training set is built up. This will allow the filtering of the dataset to remove common object types, increasing the potential for rarer objects to be found. Classification instructions can then be adjusted to concentrate on these new targets, until they too obtain training sets large enough for significant algorithmic filtering. The advantages of such a scheme include the fact that it is responsive; it is capable of adjusting to new categories of interest and to improvements in algorithmic performance.

In essence, this proposal recognises that whether or not real time observing is happen what is important is on-the-fly learning. Training sets developed dynamically could respond to changes in conditions or instrument performance, but more importantly also to the changing characteristics of a filtered dataset. This is especially important as training sets for rarer objects can only be produced once more common ones have been removed. An ambitious system - one which is truly efficient - would spend about the same amount of time assembling a training set for an algorithm as it takes to run it.

At present, responsive classification is seen in Zooniverse projects such as Space Warps, Planet Hunters and Disk Detective as taking place through the discussion forums provided. In these venues, subsets of volunteers have proved capable of developing detailed and complex classifications of problematic objects, but using this information in a programatic way is hard. Future projects should probably look to design modes of interaction that lie between hard-wired classification and free discussion if this sort of project is to be truly useful.

5. Machine Learning =======

5 MACHINE LEARNING

6. Education

6 RECOMMENDATIONS AND CONCLUSIONS

1. Progress is slowed or prevented when data is not shared freely; in many cases this is an inconvenience but in the case of monitoring for transients in, for example, large radio surveys, a failure to rapidly share data or alerts is fatal to many interesting investigations. Commensal observing should be part of the standard plan for such surveys.
2. Citizen science projects themselves can be built in an open and commensal way. Sharing code (or allowing volunteers to clone projects for their own use) will allow for rapid iteration during project development. This is not only useful, but points to a mode of working - projects set up to find one set of objects could quickly be adapted to find rarer ones as the project matures.
3. Despite the progress of computer vision, we expect there to still be a meaningful role for citizen scientists in the era of LSST and SKA-scale data. Routine classification of common objects will rapidly be tractable automatically, but anomaly detection and the identification of more unusual objects will still require some degree of human intervention.
4. Simplicity is good. While sophisticated versions of projects can be built with significant investment in complex weighting of users, task assignment and workflow, when these things are not necessary they should be avoided. This produces output where the systematics are more easily understood, and often a more enjoyable experience for volunteers.
5. Of the taxonomy of data flows shown above, the hardest is the case where continuous inspection of near-real time data is required. In this case, dynamic combinations of human and machine classifications will be needed and projects which can explore how to do this well should be prioritised in the next few years.
6. Citizen science projects have enormous potential for education, and can act as a shop window for the science being carried out by surveys and by universities. The fact that projects can now easily be created - supported by reuse of datasets - makes bringing project design into the classroom plausible, but work and documentation will be needed to allow educators to quickly understand how to make use of a complex set of tools.