Source Discovery and Time Series Classification Using Berkeley's Transient Classification Pipeline

The Berkeley Transients Classification Pipeline (TCP) is a parallelized source identification and classification pipeline which ingests several data streams and emits classified science events.

Currently, the TCP ingests and classifies subtracted sources from the Palomar Transient Factory 48" telescope survey. The 7.8 sq-deg, 100 Mpix instrument samples the sky at 120 seconds, producing ~1 Million source detections per night.

Each new observation is first associated with a source in a database containing archived data. Database entries are associated with a source and "interesting or transient" sources are flagged for follow-up observations. Follow-up observations may be made at Berkeley's 48" telescope or at other survey telescopes.

Before training the TCP's classifiers, we re-sample the well sampled lightcurves found in DotAstro.org to better match the PTF survey's observing cadence. This "Noisification" software also adds simulated instrument junk detections without the help of a human.

To train the TCP's science classifiers, the TCP references ~20,000 lightcurves which are contained in the DotAstro.org lightcurve warehouse. DotAstro.org contains 150 variable and transient classes which were derived from archived data.

To train the TCP's science classifiers, the TCP references ~20,000 lightcurves which are contained in the DotAstro.org lightcurve warehouse. DotAstro.org contains 150 variable and transient classes which were derived from archived data.

After training the TCP's classifiers, we re-sample the well sampled lightcurves found in DotAstro.org to better match the PTF survey's observing cadence. This "Noisification" software also adds simulated instrument junk detections without the help of a human.

To train the TCP's science classifiers, the TCP references ~20,000 lightcurves which are contained in the DotAstro.org lightcurve warehouse. DotAstro.org contains 150 variable and transient classes which were derived from archived data.

The resulting dataset from GroupThink is used to train a machine-learning based filter. This filter and it's associated metrics effectively model the user judgments, allowing the TCP to automatically filter out most non-astrophysical detections without the help of a human.

The GroupThink web-app displays thumbnails of astrophysical as well as defective and poorly subtracted sources from the PTF pipeline and allows a dozen users to judge whether the sources are astrophysical or not.

The real-time TCP pipeline is generally distributed over 6 cluster computers, although a 96 core beowulf cluster is occasionally included for re-evaluation purposes.

Additionally, we've recently started reclassification and exploration of classifier and attribute parameter spaces using Hadoop on Yahoo's 4000 core M45 cluster. This was made possible once we ported WEKA to the Hadoop environment.

Each new observation is first associated with a source in a database, then "feature" attributes are regenerated which characterize that source, and finally science classifications are made using those features.

This triplet of steps for a single observation are parallelized over 22-110 CPU cores using IPython.

The TCP is generally distributed over 6 cluster computers, although a 96 core beowulf cluster is occasionally included for re-evaluation purposes.

Additionally, we've recently started reclassification and exploration of classifier and attribute parameter spaces using Hadoop on Yahoo's 4000 core M45 cluster. This was made possible once we ported WEKA to the Hadoop environment.

To train the TCP's science classifiers, the TCP references ~20,000 lightcurves which are contained in the DotAstro.org lightcurve warehouse. DotAstro.org contains 150 variable and transient classes which were derived from archived data.

Confusion Matrix

This confusion matrix shows how well a single machine-learned classifier distinguishes between 30 different science classes. A perfect classifier would contain 100% along it's diagonal.

Here, the classifier's effectiveness was evaluated using cross-validation with the training set data. This means the heat-map colors represent successful classification percentages when given ideal lightcurves. Even when training and testing with the same cross-validated data, we can gain insight on how effective this classifier is with a set of lightcurve attributes.

These are a subset of PTF sources identified and classified by the TCP as potential SN/AGN during a test run. This preliminary test identified 84% of the already known and spectroscopically confirmed Super Novae as an SN/AGN science class. Mis-classifications were primarily faint and nearly-ambiguous sources.

Contacts:
Josh Bloom  (PI)
jbloom@astro.berkeley.edu
Dan Starr  (primary developer)
dstarr@astro.berkeley.edu

Links:
DotAstro.org  http://dotastro.org
(TCP links)  http://is.gd/5mA4
http://tinyurl.com/tcp123

http://DotAstro.org

To train the TCP's science classifiers, the TCP references ~20,000 lightcurves which are contained in the DotAstro.org lightcurve warehouse. DotAstro.org contains 150 variable and transient classes which were derived from archived data.

This triplet of steps for a single observation are parallelized over 22-110 CPU cores using IPython.

The real-time TCP pipeline is generally distributed over 6 cluster computers, although a 96 core beowulf cluster is occasionally included for re-evaluation purposes.

Additionally, we've recently started reclassification and exploration of classifier and attribute parameter spaces using Hadoop on Yahoo's 4000 core M45 cluster. This was made possible once we ported WEKA to the Hadoop environment.

To train the TCP's science classifiers, the TCP references ~20,000 lightcurves which are contained in the DotAstro.org lightcurve warehouse. DotAstro.org contains 150 variable and transient classes which were derived from archived data.

Confusion Matrix

This confusion matrix shows how well a single machine-learned classifier distinguishes between 30 different science classes. A perfect classifier would contain 100% along it's diagonal.

Here, the classifier's effectiveness was evaluated using cross-validation with the training set data. This means the heat-map colors represent successful classification percentages when given ideal lightcurves. Even when training and testing with the same cross-validated data, we can gain insight on how effective this classifier is with a set of lightcurve attributes.

These are a subset of PTF sources identified and classified by the TCP as potential SN/AGN during a test run. This preliminary test identified 84% of the already known and spectroscopically confirmed Super Novae as an SN/AGN science class. Mis-classifications were primarily faint and nearly-ambiguous sources.

Contacts:
Josh Bloom  (PI)
jbloom@astro.berkeley.edu
Dan Starr  (primary developer)
dstarr@astro.berkeley.edu

Links:
DotAstro.org  http://dotastro.org
(TCP links)  http://is.gd/5mA4
http://tinyurl.com/tcp123

http://DotAstro.org

To train the TCP's science classifiers, the TCP references ~20,000 lightcurves which are contained in the DotAstro.org lightcurve warehouse. DotAstro.org contains 150 variable and transient classes which were derived from archived data.

This triplet of steps for a single observation are parallelized over 22-110 CPU cores using IPython.

The real-time TCP pipeline is generally distributed over 6 cluster computers, although a 96 core beowulf cluster is occasionally included for re-evaluation purposes.

Additionally, we've recently started reclassification and exploration of classifier and attribute parameter spaces using Hadoop on Yahoo's 4000 core M45 cluster. This was made possible once we ported WEKA to the Hadoop environment.

To train the TCP's science classifiers, the TCP references ~20,000 lightcurves which are contained in the DotAstro.org lightcurve warehouse. DotAstro.org contains 150 variable and transient classes which were derived from archived data.

Confusion Matrix

This confusion matrix shows how well a single machine-learned classifier distinguishes between 30 different science classes. A perfect classifier would contain 100% along it's diagonal.

Here, the classifier's effectiveness was evaluated using cross-validation with the training set data. This means the heat-map colors represent successful classification percentages when given ideal lightcurves. Even when training and testing with the same cross-validated data, we can gain insight on how effective this classifier is with a set of lightcurve attributes.

These are a subset of PTF sources identified and classified by the TCP as potential SN/AGN during a test run. This preliminary test identified 84% of the already known and spectroscopically confirmed Super Novae as an SN/AGN science class. Mis-classifications were primarily faint and nearly-ambiguous sources.

Contacts:
Josh Bloom  (PI)
jbloom@astro.berkeley.edu
Dan Starr  (primary developer)
dstarr@astro.berkeley.edu

Links:
DotAstro.org  http://dotastro.org
(TCP links)  http://is.gd/5mA4
http://tinyurl.com/tcp123

http://DotAstro.org

To train the TCP's science classifiers, the TCP references ~20,000 lightcurves which are contained in the DotAstro.org lightcurve warehouse. DotAstro.org contains 150 variable and transient classes which were derived from archived data.

This triplet of steps for a single observation are parallelized over 22-110 CPU cores using IPython.

The real-time TCP pipeline is generally distributed over 6 cluster computers, although a 96 core beowulf cluster is occasionally included for re-evaluation purposes.

Additionally, we've recently started reclassification and exploration of classifier and attribute parameter spaces using Hadoop on Yahoo's 4000 core M45 cluster. This was made possible once we ported WEKA to the Hadoop environment.

To train the TCP's science classifiers, the TCP references ~20,000 lightcurves which are contained in the DotAstro.org lightcurve warehouse. DotAstro.org contains 150 variable and transient classes which were derived from archived data.

Confusion Matrix

This confusion matrix shows how well a single machine-learned classifier distinguishes between 30 different science classes. A perfect classifier would contain 100% along it's diagonal.

Here, the classifier's effectiveness was evaluated using cross-validation with the training set data. This means the heat-map colors represent successful classification percentages when given ideal lightcurves. Even when training and testing with the same cross-validated data, we can gain insight on how effective this classifier is with a set of lightcurve attributes.

These are a subset of PTF sources identified and classified by the TCP as potential SN/AGN during a test run. This preliminary test identified 84% of the already known and spectroscopically confirmed Super Novae as an SN/AGN science class. Mis-classifications were primarily faint and nearly-ambiguous sources.

Contacts:
Josh Bloom  (PI)
jbloom@astro.berkeley.edu
Dan Starr  (primary developer)
dstarr@astro.berkeley.edu

Links:
DotAstro.org  http://dotastro.org
(TCP links)  http://is.gd/5mA4
http://tinyurl.com/tcp123

http://DotAstro.org