Object Classification at the Nearby Supernova Factory

S. Bailey,[1⋆]**, C. Aragon,**[1]**, R. Romano,**[1,2]**, R. C. Thomas,**[1]**, B. A. Weaver**[1,3]**,** and **D. Wong**[1]

[1] Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720
[2] Luis W. Alvarez Fellow, National Energy Research Scientific Computing Center, 1 Cyclotron Road, Berkeley, CA 94720
[3] University of California, Space Sciences Laboratory, Berkeley, CA 94720

The dates of receipt and acceptance should be inserted later

Key words methods: data analysis — methods: statistical — techniques: image processing

We present the results of applying new object classification techniques to the supernova search of the Nearby Supernova Factory. In comparison to simple threshold cuts, more sophisticated methods such as boosted decision trees, random forests, and support vector machines provide dramatically better object discrimination: we reduced the number of non-supernova candidates by a factor of 10 while increasing our supernova identification efficiency. Methods such as these will be crucial for maintaining a reasonable false positive rate in the automated transient alert pipelines of upcoming large optical surveys.

1 Introduction

Upcoming large scale optical surveys such as Pan-STARRS and LSST intend to generate automated rapid-turnaround transient alerts for objects such as supernovae, active galactic nuclei, asteroids, Kuiper Belt objects, and variable stars. Microlensing surveys and gamma ray burst (GRB) detectors have successfully generated automated alerts with a low false-positive rate, but their situation is rather different from that of large scale optical surveys. GRB detectors have significantly lower background events rate than optical surveys, and microlensing events are based upon a trend in a lightcurve rather than a single observation. Future large-scale optical surveys face a fundamentally different problem with their automated alert pipelines, since they attempt to identify optical transients at the time they are first imaged. Since their observation cadence will typically be days instead of hours or minutes, large optical surveys will not be able to wait for a lightcurve of measurements before generating an alert. In comparison to current optical transient programs, future surveys will need to have significantly better signal/background event separation in order to avoid being swamped with false positive alerts.

2 The Nearby Supernova Factory

The Nearby Supernova Factory (Aldering et al. 2002) is a program to discover 100–200 type Ia supernovae in the redshift range $0.03 < z < 0.08$ and spectro-photometrically measure their lightcurve evolution. This dataset will be used for cosmological fits of the expansion of the universe and dark energy; it additionally provides a detailed sample for understanding the underlying physics of type Ia supernovae.

⋆ corresponding author: sjbailey@lbl.gov

The search uses data from the Near Earth Asteroid Tracking (NEAT) program[1] and the Palomar QUEST consortium[2] using the 112 CCD QUEST-II camera (Baltay et al. 2007) on the Palomar Oschin 1.2-m telescope. This search is the largest data volume and sky area supernova search currently operating and thus is a good testing grounds for data processing and object identification issues relevant to future large-scale optical surveys.

3 Classification Methods

Supernova searches typically select objects of interest by applying simple threshold cuts to features such as signal-to-noise, FWHM, object motion between two images, etc. If an object fails any of these cuts, it is rejected. These cuts are easy to understand but do not reflect the subtleties of a multidimensional space, where variables may be correlated and have outliers in their distributions. An object which just barely fails one of the cuts is still rejected the same as an object which fails many cuts. To use threshold cuts, one must find uncorrelated variables without significant outliers such that every cut maintains a high signal efficiency while rejecting background.

Although commonly used in supernova searches, threshold cuts are widely recognized as being a non-optimal method for signal/background separation problems. The following sections describe a variety of more powerful techniques for identifying optical transients in difference images. Further details about these methods may be found in Bailey et al. 2007.

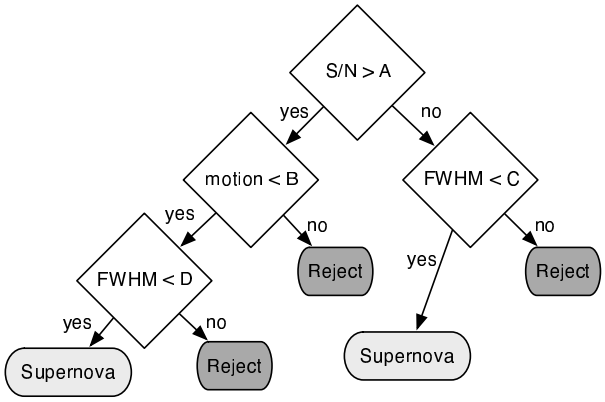

Fig. 1 Example decision tree which would treat high signal-to-noise objects differently than low signal-to-noise objects. In practice, a real decision tree has many more branches and the same variable can be used to branch at many different locations with different cut values.

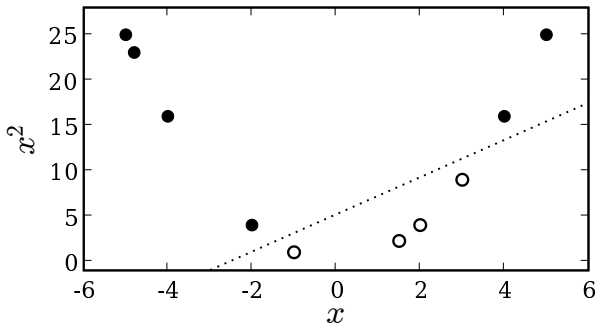

Fig. 2 Support vector machines (SVMs) map an input space of features into a higher dimensional space where the separation of classes becomes easier. In this toy example, the open and filled circles require multiple boundaries to separate them in the original space x, but in the higher dimentional space (x, x^2) they may be separated by a single line.

3.1 Decision Trees

Decision trees (Breiman et al. 1984) separate signal from background events by making a cascading set of event splits as shown in Figure 1. A training procedure automatically selects the features and cut values to generate a tree with maximal separation of signal and background events in the terminal nodes.

For a set of events, define the signal purity as $P = N_S/(N_S + N_B)$, where N_S and N_B are the number of signal and background events in the sample. The "Gini parameter" $P(1 - P)$ is 0 for a sample which is purely signal or purely background, and > 0 otherwise. The training procedure for a decision tree begins with a set of events of known type and considers all features and possible split values to minimize $\mathrm{Gini}_{\text{left child}} + \mathrm{Gini}_{\text{right child}}$. The training sample is split and the procedure is recursively applied to further split the subsamples. The splitting is stopped when the samples have reached a minimum required size (a minimum size requirement prevents overtraining on statistical fluctuations of small samples) or no split can be found which would improve the overall quality of the tree. Terminal nodes with a majority of signal events are called signal leaves; otherwise they are background leaves. When new events are evaluated using the decision tree, their signal/background classification is assigned based upon whether they correspond to a signal or background leaf.

3.1.1 Boosted Trees

Boosting algorithms improve the performance of a classifier by giving greater weight to events that are hardest to classify. In the case of decision trees, a tree is trained on a set of data, misclassified events are identified and their weights are increased, and the process is repeated to form new trees.[3] This iteratively produces a set of increasing quality decision trees. The final classifier uses the weighted ensemble average of all of the trees to make a classification decision. The boosting provides decision trees with better separation power, and the ensemble average washes out the training instabilities associated with single decision trees. In applications with $\sim$20 or more input features, Boosted Decision Trees can provide significantly better results than artificial neural networks (Roe et al. 2005).

3.1.2 Random Forests

Random forests (Breiman 2001) also generate multiple decision trees for a given training set and use a weighted average of the trees as the final decision metric. When training a tree, at each branch the training cycle only considers a random subset of the possible features available to use. This has the effect of washing out the typical training instabilities of decision trees and produces a ensemble classifier which is fast to train and robust against outliers.

3.2 Support Vector Machines

The support vector machine (SVM) algorithm is a classification method that nonlinearly maps data points from the original input space to a higher-dimensional feature space where the separation boundary is a hyperplane (Vapnik 1998; Chen et al. 2005). Figure 2 shows a toy example where open and filled circles require multiple boundaries to separate them in the original space x, but in the higher dimensional space (x, x^2) they may be separated by a single line.

The optimization problem for finding the hyperplane is constructed such that the solution depends only upon the

[1] http://neat.jpl.nasa.gov

[2] http://hepwww.physics.yale.edu/quest/palomar.html

[3] See Bailey et al. 2007 for the generalization of the purity and Gini parameters to the case of weighted events and details of the boosting of those weights.

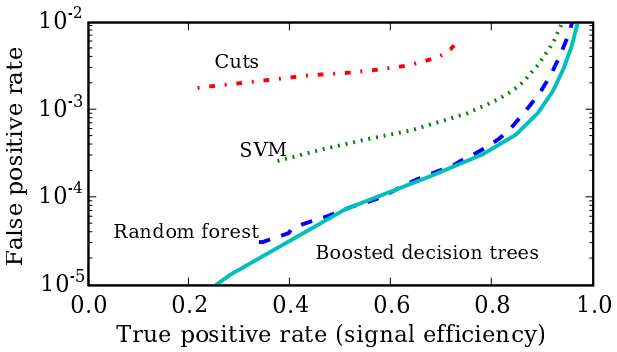

Fig. 3 Comparison of boosted trees, random forest, SVM, and threshold cuts for false positive identification fraction *vs.* true positive identification fraction.

events that are closest to the boundary (the "support vectors"). These events are identified and the hyperplane parameters are determined by the optimization. The dependence of the solution only upon the support vectors is conceptually similar to boosting algorithms which give greatest weight to difficult to classify events.

Computationally, the the higher-dimensional mapping is never explicitely calculated since the solution only depends upon dot products of vectors in that space. Instead, a "kernel-trick" is used to map dot products in the high dimensional space to a kernel function: $k(\mathbf{x}_1, \mathbf{x}_2) = \phi(\mathbf{x}_1) \cdot \phi(\mathbf{x}_2)$. Kernel functions can provide quite general mappings to higher dimensional spaces while keeping the problem computationally tractable. For this analysis we used a soft-boundary SVM method called C-SVM, which handles noisy data by allowing (but penalizing) events on the "wrong" side of the hyperplane while solving for the optimal hyperplane parameters. This seeks to prevent overfitting to genuinely noisy/overlapping data. For SVM implementation details, see Vapnik 1998.

4 Comparison of Classification Methods

Most classifiers produce a single score to describe each event. By adjusting a threshold cut on this single value, one may tune the trade-off between purity (fraction of selected events which really are signal), completeness (fraction of signal events which were selected) and total sample size. A useful way of visualizing these tradeoffs is to plot the fraction of background events selected vs. the fraction of signal events selected while adjusting the threshold cut.

Figure 3 shows a comparison between several classification methods as applied to the Nearby Supernova Factory dataset: cuts, support vector machine, random forest, and boosted decision trees. Overall, boosted decision trees provided the best performance, with random forests providing very similar results. Support vector machines considerably outperformed cuts, but did not provide as much signal/background separation power as the decision tree based methods.

Overall, for a fixed signal efficiency, boosted trees and random forests provided $\sim$30 times better background rejection than threshold cuts. After introducing boosted decision trees to the SNfactory search pipeline, we chose to operate at a selection point with $\sim$10 times less background but with an improved signal efficiency.

5 Discussion

In general, supernova seaches have not attempted to completely automate candidate selection and announcement, and thus have only improved their classification methods to the point of achieving a reasonable workload, which tends to involve vetting 10–100 false positives for each good candidate selected. Significantly more improvement will be needed to make completely automated rapid-turnaround optical transient alerts viable. At the Nearby Supernova Factory, further improvements could be readily obtained by improving the set of features used for classification. For example, one of our primary remaining backgrounds is a subtraction dipole which results from subtracting slightly misaligned stars. Any classifier of our data would be improved by adding a feature that specifically measures this effect.

In practice, there is no substitute for high quality data and a well understood detector; any problems with background events should first be addressed at the level of the detector and image processing pipeline if possible. But future optical surveys will require both high quality input data and powerful classifiers in order to maintain reasonable false positive rates in their automated transient alert pipelines. The methods presented in this work provide dramatically better object discrimination than methods otherwise employed by current supernova searches.

Acknowledgements. We would like to thank G. Aldering for useful conversations and the SNfactory collaboration for providing the search images for this analysis and followup spectroscopy of discovered supernovae. SB would like to thank the organizers and hosts of the Hotwiring the Transient Universe workshop. This work was supported in part by the Director, Office of Science, Office of High Energy and Nuclear Physics, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and by a grant from the Gordon & Betty Moore Foundation.

References

Aldering, G. et al. (SNfactory): 2002, Proc. SPIE 4836, 61
Bailey, S. et al.: 2007, ApJ 665, 1246.
Baltay, C. et al.: 2007 PASP, submitted (astro-ph/0702590)
Breiman, L.: 2001, "Random Forests," University of California, Berkeley, Dept. of Statistics, technical report 567
Breiman, L., Friedman, J.H., Stone, C.J., and Olshen, R.A.: 1984, Classification and Regression Trees (Belmont, CA: Wadsworth International Group)
Chen, P.-H., Lin, C.-J., and Scholkopf, B.: 2005, Applied Stochastic Models in Business and Industry 21(2), 111
Roe, B.P. et al.: 2005 Nucl. Instrum. Meth. A543, 577
Vapnik, V.N.: 1998, Statistical Learning Theory (New York: Wiley)