Data Mining: Multivariate clustering & classification with R

Eric Feigelson

Penn State

Summer School in Astrostatistics for Astronomers XI

June 2015

Astronomical context

Astronomers have constructed classifications of celestial objects for centuries:

- Asteros (fixed stars) vs. planetos (roving stars) [Greece, >2Kyr ago]
- Luminosae, Nebulosae, Occultae [Hodierna, mid-17th c]
- Comet orbits & morphology [Alexander 1850, Lardner 1853, Barnard 1891]
- Stellar spectra: 6 classes (Secchi 1860s), 7 classes (Pickering/Cannon 1900-20s), 10 classes w/ brown dwarfs (Kirkpatrick 2005)
- Galaxy morphology: 6+3 classes (Hubble 1926)
- Supernovae: Ia, Ib, Ic, Iib, IIP, Iin (Turatto 2003)
- Active galactic nuclei: Seyfert gal, radio gal, LINERs, quasars, QSOs, BL Lac, blazars
- Protostars/PMS stars: Class 0, 0/I, I, II, III (Lada 1992, Andre 1993)

In nearly every case, these classes were created by well-argued, but subjective assessment of source properties.

In statistical parlance, the problem is called
unsupervised clustering of heterogeneous multivariate data

Clustering methods in astronomy

Deterministic decision trees

- Abell cluster richness class (Abell 1958)
- Young stellar objects with infrared colors 0.4<[5.8]-[8.0]<1.1 and 0.0<[3.6]-[4.5]<0.8 are classified as *Class II* (Allen 2004)

Percolation or `friends-of-friends' algorithm

1. Plot data points in a 2-dimensional diagram
2. Find the closest pair, and call the merged object a `cluster'
3. Repeat step 2 until some chosen threshold is reached. Some objects will lie in rich clusters, others have one companion, and others are isolated

Statistical approach to clustering

In **unsupervised clustering** of a multivariate *n x p* dataset, the number, location, size and morphology of the data groupings is unknown. There is no `prior knowledge' of classes.

Nonparametric clustering algorithms:

- Agglomerative hierarchical clustering ~ Friends-of-friends
- K-means partitioning
- Density-based clustering

Parametric clustering algorithms:

- Mixture models

Nonparametric unsupervised clustering is a very uncertain enterprise, outcomes depend on algorithms, no likelihood to maximize.

Parametric unsupervised clustering lies on a stronger foundation (MLE, BIC). But it assumes the parametric structure is correct.

Concepts of [supervised] classification

The multivariate dataset under study represents a new ***test set*** that is a mixture of classes that have been defined in advance, either from astrophysical theory or ***training sets***. The prior knowledge of the number, location & morphology of the classes in *p*-space gives a huge advantage over unsupervised clustering.

As with clustering, some classification methods are parametric assuming multivariate normal (MVN) distributions within each class (mixure models), while others are nonparametric. Methods often labeled ***data mining*** or ***machine learning***.

Automated classification techniques are particularly important in ***wide-field astronomical surveys*** which collect a wide variety of astronomical objects: stars, galaxies, active galactic nuclei.

Wide-field surveys include: Optical CRTS, PTF, ASAS, Pan-STARRS, VISTA, DES, LSST, LAMOST; **X-ray** RASS, eROSITA; **Infrared** IRAS, MSX, Akari, WISE; **Radio** NVSS, FIRST, PKS, LOFAR, MWO

Linear classifiers

Linear discriminant analysis (Fisher 1930s) LDA finds a linear combination of variables (a *p*-dimensional hyperplane) that maximally separates two classes with known MVN distributions. The separation is measured by the ratio of the between-cluster variance **B** and the within-cluster variance **W**. The separating plane can be interpreted scientifically and used to classify new objects.

Nonparametric linear classifiers that relax the MVN assumption include the **perceptron algorithm** (1950s) that lies at the foundation of **artificial neural networks**, and the **naïve Bayes classifier.**

Support Vector Machines (SVMs), developed by Vladamir Vapnik from the 1960-1990s, have emerged as extremely powerful generalizations of LDA and the perceptron. It allows nonlinear surfaces to separate curves in *p*-space and `soft' margins to permit misclassifications. SVMs are very powerful and widely used today.

Nonparametric classifiers

Recursive partitioning algorithm (1960s) led to **Classification and Regression Trees or CART** (Leo Breiman, 1970s-2000s). Constructs dendrograms for the training set using sequence of single-variable decision rules to concentrate objects of a single class with mathematical rules for splitting and pruning branches of the dendrogram. Very powerful with **bootstrap aggregation (bagging)** and boosting. Algorithms include ID3, C4.5 an **Random Forests**.

k-Nearest Neighbor (k-NN) classifiers are very simple and computationally efficient: choose a distance metric and integer *k;* locate the *k* nearest neighbors of the training set for each member of the test set, test set point class is set to be the most common class of the *k* neighbors. Variants incude Discriminant Adaptive Nearest Neighbor (DANN) where the distance metric shrinks when the local density of points is high.

Artificial neural networks (ANNs) are algorithms to find heuristic nonlinear rules for distinguishing classes in multivariate training datasets which are then applied to test datasets. Weightings of hidden layers are iteratively reset to improve classification using **back propagation**, a gradient descent procedure. . Many choices regarding topology of hidden layers, optimality criteria, stopping rules & convergence criteria. Widely used method in astronomy.

Remarks

The word `classification' appeared in 1/3 of 2012 astronomy papers.
Astronomers encounter endless problems where patterns are sought in heterogeneous data by placing objects into distinct classes.

Most astronomers still use heuristic procedures for classification, but quantitative methods are increasingly used:

- If no prior knowledge on classes is available, then parametric mixture model or (very uncertain) nonparametric clustering methods
- If prior knowledge is available, then a vast suite of powerful supervised classification methods are available: SVMs, CARTs with boosting & bagging, k-NNs, ANNs

For complex classification problems (e.g. 20 classes in 10-dimensional space with non-MVN structures), parametric models may not be effective while nonparametric methods (CART, k-NN classifiers, ANN) can be successful. Large and reliable training sets are needed for such problems.