Supervised detection of anomalous light-curves in massive astronomical catalogs

Isadora Nun, Karim Pichara, Pavlos Protopapas, and Dae-Won Kim

-- Summary --

Ryan Lee

September 29th, 2014

Overview:

Nun et al. develop a new supervised machine learning algorithm that uses random forests and Bayesian networks to identify rare astronomical light curves. Such algorithms become increasingly important as the amount of astronomical data continues to grow.

Intuition:

Starting from a training set of data points (each of which could contain multiple features) with known labels, a random forest (RF) classifier is trained. A RF is a large number of binary classification trees trained on bootstrap data sets - samples chosen randomly with replacement that are the same size as the training set. In each tree, at every node, a random subset of features are used to split the data. The RF output for a particular input is the tally of votes from each tree in the forest that classifies that data point into labels present in the training set. After normalization, for each data input, a vector is generated that indicates the probability that particular data point belongs to each of the labels. (i.e. If there are L unique labeled categories in the training set, the output vector would be of length L.) In this way, a set of N data points is converted to a set of N probability vectors.

In the second step, a Bayesian network (BN) is constructed with the data set of probability vectors. A bayesian network is a directed acyclic graph that describes assumed probabilistic dependencies in the data. The structure of the tree can be learned using an algorithm explained by Cooper & Herskovits (1992). The joint probability of each vector of probabilities from the RF can be estimated as a product of more tractable probabilities by going up the tree. To further ease computation, the continuous probability vectors from the RF is discretised. If the joint probability as determined by the BN is low, this suggests that the chance of encountering a similar combination of probabilities in the training set was low. Thus, a low joint probability means that the data point is rare, and they have defined a outlierliness score which is the inverse.

Benefits:

1. The RF and BN need to be trained only once on a training set. Classifying unknown data points afterwards is fast.
2. The RF method has been shown to be one of the best classification methods, exploiting the diversity of random samples to improve its performance.
3. The BN elegantly allows for the determination of joint probability, which is a proxy for outlierliness. This is advantageous over distance methods or density methods in identifying potential outliers.

Questions for Application to Bank Fraud:

1. How do we create labels based on transaction data? Would we use the meta data as labels, or “learn” clusters from a training set?
2. Would a random forest perform better if, in each tree, at each node, it used a linear combination of features to split the data?