Petascale Object Classification of the LSST Event Stream

K. D. Borne (GMU), R. Laher (Caltech), Z. Ivezic (UW), N. Hamam (Caltech), and the LSST Collaboration

We describe a specific example of object classification using ANN (Artificial Neural Networks) applied to data from the SDSS ( Sloan Digital Sky Survey). We test several ANN models for classification Reliability (R) and Completeness (C). The ANN performs better than a deterministic classifier by a factor of 3 in the Unreliability (1-R) metric. These experiments are a precursor to the large-scale object classification that will be required for the LSST. The LSST object database will contain detailed information for 20 billion sources, including approximately 10 billion galaxies, a similar number of stars, and over 1 billion variable sources (optical variables, transients, or moving objects). After 10 years of survey operations, the LSST object database will comprise 10-20 Petabytes of science catalog attributes; over 200 science attributes per object will be available for classification, characterization, and mining. Deep multi-color photometry and long-term time series (photometric and astrometric, at various cadences) will yield an enormous rich potential for new scientific discoveries. LSST’s impressive panoply of precision parameter data will enable the characterization and classification of astronomical objects on a grand scale. Characterization and classification within seconds of each exposure will permit timely knowledge-based follow-up on the most significant and exciting astronomical discoveries of the coming decades.

Controlling the Enormous LSST Data Volume

Deep multi-color photometry and long-term time series (photometric and astrometric, at various cadences) will yield an enormous rich potential for new scientific discoveries. LSST’s impressive panoply of precision parameter data will enable the characterization and classification of astronomical objects on a grand scale. Characterization and classification within seconds of each exposure will permit timely knowledge-based follow-up on the most significant and exciting astronomical discoveries of the coming decades.

ANNs can be used as variable (•) from conventional color-color diagrams of astronomical sources. Exploiting data correlations among astronomical-object categories (•) knowledge hidden within such an enormous data collection. We demonstrate that ANNs not only have superior performance, but ANNs are also capable of automatically discovering and instantiating the discovered correlations explicitly.

Confronting the Enormous LSST Data Volume

While the peak accuracy is competitive with the ANN, the ANN performance is superior, as indicated in Fig. 1. By training an ANN on a smaller data set where a target parameter is known with high confidence (e.g., a quasar), or non-variable (e.g., an H or the white dwarf), ANNs can be used as precursor (initial-pass) classifiers for sorting a large number of source detections before the full dynamic behavior of the associated astronomical object can be fully characterized, over time scales of months to years. Note-pass precursor classifiers may contribute significantly to the near-real-time priorities (e.g., Bayesian) classifications that will accompany each of the 10-100,000 nightly LSST event alert notifications.

Scientific data mining use cases anticipated with the LSST database:

• Hypothesis testing – Identify and discover critical correlations among parameters (•) in the LSST science database, integrated with distributed VO-accessible data
• Compute multi-point multi-dimensional correlation functions over the full panorama of astronomical parameter spaces
• Discover zones of avoidance in interesting parameter spaces (e.g., period-gap)
• Discover new properties of known classes
• Discover new and improved rules for classifying known classes of objects (e.g., photometric–v)
• Identify novel, unexpected temporal behavior in all classes of objects
• Hypothesis testing – verify existing (or generate new) astronomical hypotheses with strong statistical confidence, using millions of training samples
• Serendipity – Discover new one-in-a-billion objects through novelty detection
• Image processing – Identify non-astronomical objects and classify them, and separate from the astronomical catalog inputs
• Provide object data with identification flags, instrument annotations, and pipeline errors from near-real-time deviation detection

The Utility of Artificial Neural Networks:

ANNs discovered correlations among input parameters are hidden within the “black box” of the trained artificial neural network. As a consequence, ANNs thus pose an interpretation challenge – how to understand and to instantiate the discovered correlations explicitly. Nevertheless, ANNs can provide strong indication that correlations among parameters exist. By their inherent generality, ANNs are capable of simultaneously processing diverse astronomical and photometric data streams, such as spitzer, astrometric, and other science streams.

ANN inputs are readily configurable to handle either static or dynamic (time-series) data (e.g., photometric), and the value of the output layer (classification) is a regression on the hidden states. The issue of how this output value can be clarified for the input data.

ANN classification performance for a 6-5-5-1 network after training is shown in Fig. 3a. The discriminator curve for the simple deterministic classifier of U = Unreliability = 1 - R = Reliability in the curve) is significantly lower than the accuracy of the ANN. We believe that this demonstrates that ANNs not only have superior performance, but ANNs are also capable of automatically discovering and instantiating the discovered correlations explicitly.

A simple probabilistic classifier (a quasar), or non-variable (a-b-c, or non-variability in our example), is labeled as

Mathematical formulation of the single ANN layer, where a neural net is trained to classify quasars and white dwarfs. The output layer of the neural net is a regression on the hidden states. The issue of how this output value can be clarified for the input data (e.g., photometric).

The LSST sky survey will yield petascale (petabytes) of data within each night. The scientific goals of the LSST survey are to: (1) detect all massive objects (WDs) in the sugar cube of our Milky Way Galaxy, (2) detect all main-sequence objects in the sugar cube of the Local Group, and (3) detect all main-sequence objects in the sugar cube of our Local Group.

The LSST sky survey will yield petascale (petabytes) of data within each night. The scientific goals of the LSST survey are to: (1) detect all massive objects (WDs) in the sugar cube of our Milky Way Galaxy, (2) detect all main-sequence objects in the sugar cube of the Local Group, and (3) detect all main-sequence objects in the sugar cube of our Local Group.

The LSST sky survey will yield petascale (petabytes) of data within each night. The scientific goals of the LSST survey are to: (1) detect all massive objects (WDs) in the sugar cube of our Milky Way Galaxy, (2) detect all main-sequence objects in the sugar cube of the Local Group, and (3) detect all main-sequence objects in the sugar cube of our Local Group.

The LSST sky survey will yield petascale (petabytes) of data within each night. The scientific goals of the LSST survey are to: (1) detect all massive objects (WDs) in the sugar cube of our Milky Way Galaxy, (2) detect all main-sequence objects in the sugar cube of the Local Group, and (3) detect all main-sequence objects in the sugar cube of our Local Group.