Machine Learning to Recognize Phenomena in Large Scale Simulations

Jeff Schneider
Barnabas Poczos
Liang Xiong
Alex Szalay

Executions of large-scale simulations that generate enormous amounts of data are now ubiquitous across the sciences. Low-level differential equations are used to drive the simulations and scientists hope to see and understand the larger scale phenomena that arise from them. Traditionally, simulations would be visualized by scientists or summarized by a few simple statistics. As these data sets increase in size, however, scientists are increasingly unable to easily answer even basic questions: Did anything interesting or unusual happen in the simulation? When and where did it happen? What are the most common phenomena comprising the simulation? What are the spatial and temporal distributions of those phenomena? Do they interact with each other? In this work, we propose new machine learning based methods of answering these questions.

Classification, clustering, anomaly detection, low-dimensional embedding, and manifold learning, are among the most common and important problems in machine learning. Existing methods for doing these tasks treat the individual data point as the object of consideration. Each point may be given a class label, detected as an anomaly, or embedded into a different space. In this work, we change the object of consideration to be a group of data points. For example, in a turbulence simulation, any individual particle or grid point is unlikely to have much significance, but a spatial group of them might represent an interesting vortex.

In this new setting, we assume each object (now a group of data points) is associated with an underlying continuous probability distribution over features (e.g. velocities, pressures, temperatures). These distributions are unknown, but the data points comprising the group are assumed to be i.i.d. samples from the distributions. We develop both parametric and non-parametric methods for analyzing these groups.

In the non-parametric approach, we estimate the distance (or divergence) between pairs of distributions from their i.i.d. samples. We propose divergence estimators based on simple k nearest neighbor statistics. We have shown that the estimators are consistent and can be computed quickly with classic tree-based caching structures. Once we have distances between objects (groups of points), we use those distances to create classification, clustering, and anomaly detection algorithms that use them.

In the parametric approach we work from topic models, such as latent dirichlet allocation (LDA), where documents are commonly treated as an unordered group of words. At the lowest level of the models, we substitute the multinomials usually used to draw words from a dictionary with mixtures of Gaussians that generate the continuous feature vectors appearing in our data. The models provide full generative models for the groups appearing in a simulation. Using the generative models, we can assign a likelihood to each group and thus detect anomalies. We can also learn separate models for each class and perform classification by finding the highest likelihood model for a new object of interest. The higher level part of the model can be configured to automatically do clustering as part of the modeling process.

We demonstrate our methods using data from the JHU Turbulence Database Cluster (TDC). TDC simulates fluid motion through time on a 3-dimensional grid. At each time step and each vertex of the grid, TDC records the 3-dimensional velocity of the fluid as well as the pressure. We consider the grid vertices in a local cubic region as a group, and the goal is to find groups of vertices whose velocity distributions (i.e. moving patterns) are unusual and potentially interesting. In our empirical study, we found that we were able to identify both individual vortices and higher level phenomena such as a sheet of vortices arising from the sheer of two fluid masses rubbing across each other. We also demonstrate the broader applicability of our methods by testing them on image data sets.