LEARNING COMPLEX CLASSIFICATION MODELS FROM LARGE DATA SETS

Julian L. Center, Jr.
Creative Research Corporation, 385 High Plain Rd., Andover, MA 01810
(e-mail: jcenter@ieee.org)

Abstract

Taking a Bayesian approach to designing a classification algorithm, we start with a general model form with adjustable parameters and learn a posterior probability distribution for the model parameters based on a set of data samples. In many applications, classification models can become quite complex. For example, an image recognition algorithm can be based on a mixture model with many components, each with many parameters.

Unfortunately, the computations needed to determine the posterior probability distribution often become overwhelming. Since we are considering models with a large number of parameters, a large number of samples is needed to narrow the range of probable models. Because the model is complex and there are many samples, computing the likelihood of a particular model takes significant computer time. Therefore, exploring the large model-parameter space in detail becomes an intractable problem.

Of course, if the data set is large enough, the information gain will narrow the range of probable models to a very small subset of the parameter space. If we can find this subspace quickly, we can employ our computational power to adequately explore this region. However, searching for this small region can prove difficult because it is so small and because evaluating each point in the search requires evaluating the complete likelihood function.

We present a computationally feasible solution to this problem based on breaking the large data set into several smaller data subsets. We use Bayesian theory to design an algorithm for processing the subsets in stages.

For each stage, we combine a new data subset with a prior distribution that summarizes previous stages. We employ nested sampling to focus our exploration of the parameter space on high probability areas and use slice sampling to draw candidate parameter values from the prior distribution. We then use variational methods to approximate the resulting distribution on the parameter space. This approximation summarizes the results and becomes the prior distribution for the next stage.

The end result is a discrete probability distribution on model parameter space. This leads to a classification algorithm that is a "mixture of experts" combining the classification probabilities of the best models.