論 文 の 内 容 の 要 旨

論文題目 Statistical Modeling as a Search for Randomness Deficiencies
（和訳：乱数欠陥の探索による統計的モデル化)

氏　　名　Burfoot Daniel Crary
（バールフット、ダニエル クレーリー ）

Abstract

The problem of estimating a real distribution from a limited set of empirical data is a widely important problem in statistical modeling and recognition. This thesis proposes an algorithm for the automatic construction of complex models. This algorithm works by dividing the data set into subsets, and improving the current model ensemble for each subset, and then continuing recursively. The special features of the method are that it avoids the combinatorial explosion caused by combining several context functions; that it works when the context functions are not independent, and that its performance scales well when applied to very large data sets. The thesis is divided into 11 chapters as described below, which present an overview of the basic concepts, a formal description of the method, an empirical evaluation on real world data, as well as an academic discussion of the philosophical basis of the idea.

Chapter 1 (Introduction) presents the main concepts of the proposed approach to statistical modeling. The "Algorithmic Information Theory (AIT) view" of statistical modeling is developed and compared to the view of traditional statistics and the view of information theory. Two key conceptual barriers to the AIT view are discussed, and solutions are proposed. The basic operation of the PITA algorithm is discussed, and two example problems are mentioned where PITA can be used. These example problems illustrate two important challenges faced by any feature-based modeling algorithm: the feature combination problem and the feature selection problem. A brief motivation section argues that a new way of statistical thinking is needed for new applications of statistics to fields such as computer vision and speech recognition.

Chapter 2 (Model Ensemble Update Method) presents a method for applying a simultaneous update to an ensemble of statistical models. The method uses the CDF F(x) of the model Q(x) to transform the observed data points X: Y=F(X). This technique is called the Probability Integral Transform. If the model Q(x) matches the real distribution P(x) that generated the points, then the distribution of Y values will be uniform. If the Y values are non-uniform (randomness deficiency), this indicates an imperfection in the model Q(x) and indicates a way to update Q(x) to bring it closer to P(x). The parameters of the model update are simply the bin counts of a histogram of Y values (called PIT values). Also, the update and the improvement resulting from the update are independent of the original data and models, and depend only on the histogram bin counts. Finally, it is shown that the PIT histogram update can be applied to an ensemble of models.

Chapter 3 (PIT Value Analysis Algorithm) presents the core idea of the thesis, the PIT Value Analysis (PITA)

algorithm. A key idea of the algorithm is to maintain a separate model distribution for each data point. Each data point is also assumed to have its own context. A binary context function can be used to separate the data set into two subsets. The PIT histogram based model ensemble update can then be applied separately to each subset. The PITA algorithm uses a battery of context functions, and in each round selects the best context function by scoring the PIT histograms of the corresponding subsets. At the end of each round, the subsets corresponding to the best context function are updated separately using the PIT histogram model ensemble update method. The computational characteristics of the algorithm are discussed, with special reference to its ability to handle large data sets. The key ability of the algorithm to be layered over some other initial model is discussed. Finally the so-called "Bin Overlap" problem is analyzed and methods for dealing with it are presented.

Chapter 4 (Alternative CDF Transformation Techniques) discusses a set of alternative model ensemble update methods. Each of these depends on a different set of statistics related to the PIT values, and has a method for predicting the codelength savings achieved by the update.

Chapter 5 (Related Work) compares the PITA algorithm to other well-known ideas in the machine learning and statistics literature, including AdaBoost and other boosting methods, the Maximum Entropy framework, and various Goodness of Fit tests in statistics.

Chapters 6-10 present applications of the PITA algorithm to various tasks, including image compression, binary pattern classification, word morphology modeling, speech modeling, and motion recognition. The rigorous test of image compression indicates that the theory is basically sound. Encouraging results are achieved for image compression rates. In the chapter on binary pattern classification, a modified form of the PITA algorithm is applied to a well-known machine learning benchmark dataset. The results achieved are competitive with AdaBoost, a popular boosting algorithm. The word morphology chapter discusses problems that arise when dealing with non-numeric data, and presents some solutions to those problems. The speech modeling chapter shows how complex models for speech can provide much better descriptions of the speech data than standard models such as the Laplacian or Gaussian. On the motion recognition task, the PITA algorithm achieves much better recognition performance than an HMM.

Chapter 11 (Conclusion) provides some concluding remarks. It is emphasized that the PITA algorithm is just one example of a statistical modeling algorithm based on the search for randomness deficiencies. Other types of algorithms, using other techniques to search for randomness deficiencies, can be defined. The specific advantages of the PITA algorithm are reiterated, including its ability to construct high-performance complex models, its property that it can be layered over other models, and its ability to scale up to large data sets. Various issues and lessons coming from each specific application are discussed. A particularly important lesson comes from the speech modeling application, where a model of about one million bits complexity was used to save more than eighty million bits when encoding the data. This shows that when attempting to model raw data (in this case speech data), highly complex models can be constructed without overfitting. Finally, some pieces of future work are mentioned. One goal is to develop a version of PITA that can exploit the idea of "hidden" abstractions. Another goal is to use the AIT view of statistical modeling to develop a theory of cortical function.

The specific contributions made by the thesis are as follows. (i) The realization that statistical modeling can be approached as a search for randomness deficiencies in encoded data, i.e. a bit string in information transmission, indicating the imperfectness of data compression. (ii) The realization that the encoded data does

not need to be represented as a bit string, and in particular can be represented as a stream of [0,1]-distributed numbers called PIT values, and the process of searching for randomness deficiencies will still work. (iii) The derivation of a method for updating an *ensemble* of models based on a histogram of PIT values. (iv) A formula for predicting the codelength savings achieved by applying the model update which depends only on the PIT histogram bin counts. (v) A method for scoring a context function in terms of the scores associated with the PIT histograms it generates. (vi) An algorithm called PITA that uses the model ensemble update method, a set of context functions, and the method for scoring the context functions, to construct complex conditional models Q(x|c), where x is an outcome and c is an arbitrary context value. (vii) The derivation of a number of other model ensemble update methods, along with different ways of predicting the savings achieved. (viii) The realization that the PITA algorithm can be *layered* on top of some other arbitrary model, resulting in an improved hybrid model. (ix) Experimental demonstrations of the PITA algorithm on a variety of tasks including image compression, binary pattern classification, word morphology modeling, speech modeling, and motion modeling and recognition.

In summary, this thesis develops an approach for the construction of complex models using large data sets, using a view of the problem of statistical modeling that is based on algorithmic information theory. The soundness of the method is demonstrated on a wide number of representative applications. Also, because of the algorithm's ability to be layered over an existing model to achieve a performance improvement, it has very wide applicability.

For the above reasons, the thesis makes several contributions to the field of information sciences. Thus, it is accepted for the degree of doctor of philosophy.