IDEA GROUP PUBLISHING
701 E Chocolate Avenue, Hershey PA17033, USA
Tel: 717/533-8845; Fax: 717/533-8661; URL: www.idea-group.com
ITB7258

Chapter IX

The Gamma Test

Antonia J. Jones, Dafydd Evans, Steve Margetts and Peter J. Durrant
Cardiff University, UK

Idea Group Inc

The Gamma Test is a non-linear modelling analysis tool that allows us to quantify the extent to which a numerical input/output data set can be expressed as a smooth relationship. In essence, it allows us to efficiently calculate that part of the variance of the output that cannot be accounted for by the existence of any smooth model based on the inputs, even though this model is unknown. A key aspect of this tool is its speed: the Gamma Test has time complexity O(Mlog M), where M is the number of data-points. For data sets consisting of a few thousand points and a reasonable number of attributes, a single run of the Gamma Test typically takes a few seconds.

In this chapter we will show how the Gamma Test can be used in the construction of predictive models and classifiers for numerical data. In doing so, we will demonstrate the use of this technique for feature selection, and for the selection of embedding dimension when dealing with a time-series.

Idea Group Inc

INTRODUCTION

Idea Group Inc

The Gamma test was originally developed as a tool to aid the construction of data-derived models of *smooth* systems, where we seek to construct a model directly from a set of measurements of the system's behaviour, without assuming any *a priori* knowledge of the underlying equations that determine this behaviour. Neural networks may be considered as the generic example of a data-derived modelling technique.

We think of the system as transforming some *input* into a corresponding *output*, so the output is in some way 'determined' by the input. This is a fairly general representation – in the case of a dynamical system ,the current state of the system

This chapter appears in the book, *Heuristic and Optimization for Knowledge Discovery*, edited by Ruhul Sarker, Hussein Abbass and Charles Newton. Copyright 2002, Idea Group Inc.

may be thought of as the input, with the output representing the state of the system after some time interval has elapsed.

One problem in constructing models solely on the basis of observation is that measurements are often corrupted by *noise*. We define noise to be any component of the output that cannot be accounted for by a smooth transformation of the corresponding input.

The Gamma test (Aðalbjörn Stefánsson, Končar & Jones, 1997; Končar, 1997) is a technique for estimating the noise level present in a data set. It computes this estimate *directly from the data* and does not assume anything regarding the parametric form of the equations that govern the system. The only requirement in this direction is that the system is *smooth* (i.e. the transformation from input to output is continuous and has bounded first partial derivatives over the input space).

Noise may occur in a set of measurements for several reasons:

- Inaccuracy of measurement.
- Not all causative factors that influence the output are included in the input.
- The underlying relationship between input and output is not smooth.

The applications of a data-derived estimate of noise for non-linear modelling are clear. In the first instance, it provides a measure of the *quality* of the data – if the noise level is high we may abandon any hope of fitting a smooth model to the data. In cases where the noise level is moderately low, the Gamma test can be used to determine the best time to stop fitting a model to the data set. If we fit a model beyond the point where the mean squared error over the training data falls significantly below the noise level, we will have incorporated some element of the noise into the model itself, and the model will perform poorly on previously unseen inputs despite the fact that its performance on the training data may be almost perfect. Taking our noise estimate as the optimal mean squared error by running the Gamma test for an increasing number of data points and seeing how many points are required for the noise estimate to stabilize, we may also obtain an indication of the number of data points required to build a model which can be expected to perform with this mean squared error .

It is not immediately apparent that a technique for estimating noise levels can be of use in data mining applications. Its usefulness derives from the fact that low noise levels will only be encountered when *all* of the principal causative factors that determine the output have been included in the input. Some input variables may be irrelevant, while others may be subject to high measurement error so that incorporating them into the model will be counter productive (leading to a higher *effective* noise level on the output). Since performing a single Gamma test is a relatively fast procedure, provided the number of possible inputs is not too large, we may compute a noise estimate for each subset of the input variables. The subset for which the associated noise estimate is closest to zero can then be taken as the “best selection” of inputs.

The objectives of this chapter are to provide a clear exposition of what the Gamma test is, to describe how it works, and to demonstrate how it can be used as a data mining tool.

Idea Group Inc

25 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/gamma-test/22154

Related Content

Dynamic Knowledge Representation as a Formalization Conveyor for Manmade Systems With Useful Impulse

Andrey Naumov, Ilya Popov, Igor Bondarenko, Boris Krylov, Roman Timoninand Ivan Ofitserov (2018). *Dynamic Knowledge Representation in Scientific Domains (pp. 270-285).*

www.irma-international.org/chapter/dynamic-knowledge-representation-as-a-formalization-conveyor-for-manmade-systems-with-useful-impulse/200181

Reducing a Class of Machine Learning Algorithms to Logical Commonsense Reasoning Operations

Xenia Naidenova (2008). *Mathematical Methods for Knowledge Discovery and Data Mining (pp. 41-64).*

www.irma-international.org/chapter/reducing-class-machine-learning-algorithms/26132

Neural Networks - Their Use and Abuse for Small Data Sets

Denny Meyer, Andrew Balemiand Chris Wearing (2002). *Heuristic and Optimization for Knowledge Discovery (pp. 169-185).*

www.irma-international.org/chapter/neural-networks-their-use-abuse/22160

Institutional Research Using Data Mining: A Case Study in Online Programs

Constanta-Nicoleta Bodea, Vasile Bodeaand Radu Mogos (2012). *Cases on Institutional Research Systems (pp. 66-102).*

www.irma-international.org/chapter/institutional-research-using-data-mining/60841

User-Centered Maintenance of Concept Hierarchies

Kai Eckert, Robert Meuseland Heiner Stuckenschmidt (2011). *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances (pp. 105-128).*

www.irma-international.org/chapter/user-centered-maintenance-concept-hierarchies/53883