Journal of Statistical Planning and Inference 141 (2011) 1307–1312

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Analyzing supersaturated designs with entropic measures

C. Koukouvinos *, E. Massou, K. Mylona, C. Parpoula

Department of Mathematics, National Technical University of Athens, 15773 Zografou, Athens, Greece

ARTICLE INFO

Article history:
Received 28 October 2009
Received in revised form
6 October 2010
Accepted 11 October 2010
Available online 17 October 2010

Keywords:
Supersaturated design
Factor screening
Rényi entropy
Tsallis entropy
Havrda–Charvát entropy
Information gain
Generalized linear models
Error rates

ABSTRACT

A supersaturated design is a design for which there are fewer runs than effects to be estimated. In this paper, we propose a method for screening out the important factors from a large set of potentially active variables, based on an information theoretical approach. Three entropy measures: Rényi entropy, Tsallis entropy and Havrda–Charvát entropy, have been associated with the measure of information gain, in order to identify the significant factors using data and assuming generalized linear models. The investigation of the proposed method performance and the comparison of each entropic measure application have been accomplished through simulation experiments. A noteworthy advantage of this paper is the use of generalized linear models for analyzing data from supersaturated designs, a fact that, to the best of our knowledge, has not yet been studied.

1. Introduction

Technology and science, which are firmly connected with industrial applications, have led, due to their recent progress, to superior and more complicated systems. These large-scale systems involve many potentially important factors which may be varied during experiment design and operation, but commonly only a few of them are expected to be active. The effective factors are however unknown a priori. The situation where many effects are unimportant is called "effect sparsity" and this phenomenon was studied by Box and Meyer (1986).

Hence, it is obvious that experimenters need to investigate methods for a number of factor reductions, in order to succeed in time and cost benefits. In the last decades, researchers have focused on designs for which there are fewer runs than effects to be estimated in a proposed model, called supersaturated designs. Supersaturated designs have been shown to be effective for sifting through a large number of potentially important factors in order to identify those which mainly affect the performance of a system. In such conditions, the experimenter centers on only up to p active factors from the initial set of m factors involved in the experiment.

In such a decision problem, errors of various types and costs must be balanced. In screening designs, there is a cost of declaring an inactive factor to be active (Type I error), and a cost of declaring an active effect to be inactive (Type II error). Type II errors are troublesome as addressed in Lin (1995), as well as Type I errors, since they can cause unnecessary cost in follow-up experiments and they can cause detrimental actions if the experiment has immediate impact in practice. Under circumstances of effect sparsity, Type I errors are very likely.

The idea of SSDs was initiated in the 1950s, when Satterthwaite (1959) proposed this class of designs. Since then, many methods for supersaturated designs construction have been proposed, for example, among others (Lin, 1993; Wu, 1993; Nguyen, 1996; Tang and Wu, 1997).

* Corresponding author. Tel.: +30 210 7721706.
E-mail address: ckoukouv@math.ntua.gr (C. Koukouvinos).

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.10.001

In contrast to the wide study of construction methods of SSDs, their analysis methods are yet in an early research stage, although many different approaches for analyzing SSDs are provided in the statistical literature of the recent years.

Hamada and Wu (1992) used the Plackett–Burman designs in screening experiments for identifying important main effects, whereas a half fraction of Plackett–Burman designs were used by Lin (1993) who suggested and performed the forward selection method for identifying active factors. Wang (1995) applied this analysis on the other half fraction of Plackett–Burman design and a few years later, a Bayesian variable selection method for analyzing experiments with complex aliasing was proposed by Chipman et al. (1997). Abraham et al. (1999) applied stepwise and all-models methods to investigate the active factors, Beattie et al. (2002) proposed a two-stage Bayesian model selection strategy for SSDs, while Li and Lin (2002) proposed a variable selection approach based on penalized least squares. Holcomb et al. (2003) proposed contrast-based methods, while Lu and Wu (2004) proposed a modified stepwise selection based on the idea of staged dimensionality reduction and Zhang et al. (2007) suggested a method based on partial least squares.

This paper is organized as follows. In Section 2, we present the entropic measures used and we discuss how to apply these measures in the proposed method. In Section 3, we perform some simulation experiments to evaluate the suggested method comparing the use of each measure. In Section 4 the emergent results are discussed and some concluding comments are given.

2. Analysis of supersaturated designs via entropic measures application

Generalized linear models (GLMs) have been used for the implementation of the proposed method, especially a logistic regression model, where the response variable has only two possible outcomes, denoted by 0 and 1, has been taken into consideration.

Hence, we may consider now a logistic regression model, whose general form is the following:

$$y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k,$$

where $\mathbf{x}_k=[1,x_{k1}, x_{k2},\ldots,x_{km}]$, $\boldsymbol{\beta}'=[\beta_0,\beta_1,\beta_2,\ldots,\beta_m]$ and the response variable y_k takes on the value either 0 or 1. We will assume that the response variable y_k is a Bernoulli random variable with probability distribution $P(y_k=1)=\pi_k$, where $\pi_k=\exp(\mathbf{x}_k'\boldsymbol{\beta})/(1+\exp(\mathbf{x}_k'\boldsymbol{\beta}))$. For more details on logistic regression model, we refer the interested reader to McCullagh and Neider (1989) and Montgomery et al. (2006).

The suggested method is an information theoretical approach of variable selection issue, as it is based on three entropies: Rényi entropy, Tsallis entropy and Havrda–Charvát entropy.

Shannon (1948) introduced the concept of entropy which has a central role in information theory and may be considered a measure of the uncertainty associated with a random variable, since it is defined in terms of its probability distribution. In other words, Shannon entropy is a measure of the average information content that is missing when the value of the random variable is unknown and is a way for quantifying the information. Shannon entropy has been used widely in many applications, one of which is in the top-down decision tree algorithm C5.0, the foundation of most data mining packages.

Rényi (1961) generalized Shannon entropy to a one-parameter family of entropies by defining an entropy of order α which is called the Rényi entropy. The concept of Rényi entropy has a number of applications in coding theory, statistical mechanics, statistics and other related fields. See for instance Bercher (2008), Jenssen and Eltoft (2008), and Zografos (2008)). In 2009, Golshani et al. (1992) defined the conditional Rényi entropy, giving a definition different from the one that Cachin (1997) gave in 1997 and hence the validity of the chain rule for Rényi entropy is now proved.

More precisely, the Rényi entropy of a probability distribution $P=(p_1,\ldots,p_n)$ or of a random variable X, with probability distribution $P(X=x_i)=p_i$, $i=1,2,\ldots,n$, is defined as

$$H_\alpha(X)=H_\alpha(P)=\frac{1}{1-\alpha}\log\sum_{i=1}^{n}p_i^\alpha,\quad \alpha>0,\ \alpha\neq 1.$$

The Rényi entropy tends to Shannon entropy as $a\to 1$.

Considering a random vector (X,Y) with probability distribution $P(X=x_i,Y=y_j)=p_{ij}$, $i=1,2,\ldots,n$ and $j=1,2,\ldots,m$, then the joint Rényi entropy is given by the form

$$H_\alpha(X,Y)=\frac{1}{1-\alpha}\log\sum_{i,j=1}^{n}p_{ij}^\alpha,\quad \alpha>0,\ \alpha\neq 1.$$

According to that obtained by Golshani et al. definition, the conditional Rényi entropy of random variable Y, given X, is defined as

$$H_\alpha(Y|X)=\frac{1}{1-\alpha}\log\frac{\sum_{i,j}p_{ij}^\alpha}{\sum_i p_i^\alpha},$$

which implies that

$$H_\alpha(Y|X)=\frac{1}{1-\alpha}\log\frac{\sum_{i,j}p_{ij}^\alpha}{\sum_i p_i^\alpha}=\frac{1}{1-\alpha}\left[\log\sum_{i,j}p_{ij}^\alpha-\log\sum_i p_i^\alpha\right]=H_\alpha(X,Y)-H_\alpha(X).$$