European Journal of Operational Research 241 (2015) 161–176

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Computational Intelligence and Information Management

Concurrent multiresponse non-linear screening: Robust profiling of webpage performance

George J. Besseris

Department of Technology Management, City University of Seattle, Bellevue, WA, USA

ARTICLE INFO

Article history:
Received 20 March 2013
Accepted 16 June 2014
Available online 26 June 2014

Keywords:
Robust screening
Non-linear profiling
Non-parametric screening
Non-linear orthogonal array
Webpage performance

ABSTRACT

Profiling engineered data with robust mining methods continues attracting attention in knowledge engineering systems. The purpose of this article is to propose a simple technique that deals with non-linear multi-factorial multi-characteristic screening suitable for knowledge discovery studies. The method is designed to proactively seek and quantify significant information content in engineered mini-datasets. This is achieved by deploying replicated fractional-factorial sampling schemes. Compiled multi-response data are converted to a single master-response effectuated by a series of distribution-free transformations and multi-compressed data fusions. The resulting amalgamated master response is deciphered by non-linear multi-factorial stealth stochastics intended for saturated schemes. The stealth properties of our method target processing datasets which might be overwhelmed by a lack of knowledge about the nature of reference distributions at play. Stealth features are triggered to overcome restrictions regarding the data normality conformance, the effect sparsity assumption and the inherent collapse of the 'unexplainable error' connotation in saturated arrays. The technique is showcased by profiling four ordinary controlling factors that influence webpage content performance by collecting data from a commercial browser monitoring service on a large scale web host. The examined effects are: (1) the number of Cascading Style Sheets files, (2) the number of JavaScript files, (3) the number of Image files, and (4) the Domain Name System Aliasing. The webpage performance level was screened against three popular characteristics: (1) the time to first visual, (2) the total loading time, and (3) the customer satisfaction. Our robust multi-response data mining technique is elucidated for a ten-replicate run study dictated by an $L_9(3^4)$ orthogonal array scheme where any uncontrolled noise embedded contribution has not been necessarily excluded.

1. Introduction

Structured knowledge acquisition remains at the forefront as a key field of intense interest in information sciences as it is exhibited by current reviews on the subject (Liu, Li, Liu, & Chen, 2012; Mariscal, Marban, & Fernandez, 2010). Statistical engineering provides the inference capability for the knowledge gaining effort (Hastie, Tibshirani, & Friedman, 2009). It is statistical engineering that fuels innovation for products and services that succeed in sustaining global demand (Goh, 2010; Pantula, 2011). A fruitful strategy for imbuing cutting-edge information into designing competitive products while building insight in how to improve existing processes has been associated with the area of Design of Experiments (DOE) (Taguchi, Chowdhury, & Taguchi, 2000; Box, Hunter, & Hunter, 2005). DOE studies are generally geared towards screening projects. In harvesting information, DOE provides the profiling apparatus where screened effects undergo a fastidious statistical filtering (Bose, 1961). Classical DOE methodology requires structured data generation engineered through the use of fractional factorial designs (FFDs) followed by a data processing phase (Mukerjee & Wu, 2010). In this work we will spotlight a particular brand of FFDs which is favored by many researchers in a wealth of technological applications. Such FFDs are the ones that belong to the so called orthogonal array (OAs) family. It is noted that OAs constitute part of the standard engineering toolbox for robust on-line design in Taguchi methods (Taguchi, Chowdhury, & Wu, 2004). Accordingly, OAs are applied broadly in information research in diverse studies that range from improving computer vision and medical imaging processing, to enhancing spam-filter performance and software development, and from optimizing vapor deposition processes to screening milling operations and simulated annealing (Besseris, 2009a, 2010b, 2010c; Chang, 2008; Chen & Sun, 2000; Jung & Yum, 2011; Orfanakos & Besseris, 2010; Tansel, Gulmez, Demetgul, & Aykut, 2011). Customarily, the data analysis part in

E-mail address: besseris@cityu.edu

http://dx.doi.org/10.1016/j.ejor.2014.06.021
0377-2217/© 2014 Elsevier B.V. All rights reserved.

DOE is entrusted on mainstream multi-testing techniques such as the Analysis of Variance (ANOVA) or regression-based approaches related to the General Linear Modeling (GLM) principle (Ilzarbe, Alvarez, Viles, & Tanco, 2008). The OAs are utilized optimally in programming the data collection phase of a study only when they are saturated (Besseris, 2009b). OA saturation is synonymous to complete utilization of all available array columns for a selected OA sampling scheme. In the state of saturation, resource usage and research budgeted expenditures are minimized. Thus, from a data mining perspective, while maximum information is recovered from collecting datasets that have been planned on saturated OAs, the cost to gain that information is optimally suppressed at the same time. However, working with saturated OA schemes takes a toll on selecting a proper mainstream technique to handle the output data. This is because saturated OAs are not generally symbiotic with standard multi-factorial contrasting tools such as ANOVA and GLM. Both of these last two methods emerge as inoperative in saturated OA schemes because both approaches require information about the unexplainable error which is practically incalculable under saturation. As a result, inferences drawn with either ANOVA or GLM techniques are not poised to be considered in an objective manner. This phenomenon stems from the fact that at the saturation point, all available DFs become depleted due to their complete allocation to the tested effects, allowing no capacity to be distributed to the estimation of the unexplainable error (Besseris, 2012). Without the ability to pinpoint the contribution arising from the unexplainable error, the required *F*-test ratios in ANOVA and the standard errors in GLM cannot be decoded. In summary, saturation in OA sampling schemes simply equalizes the supply of data with the maximum demand for identifiable information but in the process removes the discovery capability from ANOVA and GLM methods. In spite of lacking any solid basis to sizing-up statistical significance, heuristic approaches exist to assist in interpreting experimental outcomes using the fitting components of ANOVA or GLM results. Response graphs and response tables typify a line of such diagnostics (Box et al., 2005; Taguchi et al., 2004). Squeezing out maximum information at the saturation point during an OA analysis has been a topic of heightened interest for the past half-century where more than thirty techniques have been devised to restore objectivity in the multi-factorial decision making (Besseris, 2009b). Most of the proposed work has been confined to the saturated-unreplicated case addressing the uni-response problem. Addressing the multi-response multi-factorial profiling case has been dealt with recently using nonparametrics which includes the additional option to prioritize the profiled traits for a group of unreplicated responses (Besseris, 2009c). A pure order-statistics single-response methodology has been proposed for a non-linear unreplicated-saturated OA scheme assorted with an illustrative application drawn from the area of improving the quality performance of an information technology system (Besseris, 2010a).

The technique that is presented in this work is intended to extract simultaneous multi-response information induced from multi-factorial contrasting. The main aim of our proposal is to provide a robust tool capable of deciphering process or product behavior in pragmatic modern operational surroundings. As it will become more evident later in this report, the bulk of potential assumptions arising with such types of techniques is maintained rather lean in our approach. The main theme of our method is to stochastically screen for multi-effect status and non-linearity simultaneously. This technique becomes then a convenient 'two-in-one' (combo) profiler. Ordinary profilers test at two preselected operational endpoints taking the risk that in the examined range the response will be a monotonous increasing or decreasing function. This assumption might be valid unless the shape of the curve is concave or convex. If the latter occurs then predictions may be digressed. There are several advantages that are realized with this new approach. Our method integrates the robust data reduction of replicates with the super-ranking concept to downgrade the replicated multi-response problem to its corresponding unreplicated counterpart (Besseris, 2009c, 2010d). The overall method may claim to offer some stealth features because is operable without an assuring knowledge of the various distribution laws that ought to be tagged to the investigated responses. Theoretically, our method possesses a stochastic intelligence that 'flies-by' without being gaged against any statistical entity that might resemble to the concept of the unexplainable error. Nevertheless, the approach intimately conforms to leading edge tendencies in information technology to facilitate forecasting under unknown and unknowable disturbances (Cicerone, Di Stefano, Schachtebeck, & Schobel, 2012).

The type of the sampling schemes that will be adopted in the developments that follow have been in accord to the non-linear OAs insomuch as they are regularly embraced in the recent information and knowledge engineering research (Chang, Chen, & Liao, 2010; Chang & Chen, 2011; Khaw, Lim, & Lim, 1995; Kim & Yum, 2004; Lin & Jules, 2006; Wang & Huang, 2007). Finally, our formalism is distinguishable in another practical aspect that of being indifferent to the so-called sparsity assumption (Besseris, 2009b). Sparsity is a binding notion for quite a few methods that aspire to process saturated OA datasets. In simple terms, this means that without the sparsity assumption several techniques that depend on it are rendered inoperable. Sparsity takes for granted that only a small number of the examined effects will turn out to be statistically significant. From a knowledge discovery standpoint, sparsity discounts in advance the key role of the examined effects. In other words, sparsity is naturally biased against the anticipated extent of influences on the investigated phenomenon. Henceforth, the sparsity postulation has been rescinded in our developments to the benefit of exploring the total content of information without any preset restrictions on how much information should be eventually extracted from the profiling endeavor.

2. Materials/methods

2.1. Measuring web site performance

The measurement of web site performance has been linked directly with corporate overall performance for a wide spectrum of organizational types (Welling & White, 2006). Therefore, it is of paramount importance to acquire and further develop techniques that conveniently and efficiently metamorphose web usage data to propitious knowledge discovery (Liu, 2011; Raju & Satyanarayana, 2008). The ideal tactic to gain knowledge regarding web-page performance issues is to examine weblog transactions (Facca & Lanzi, 2005). Integrating web data which may emanate from multifarious streams, ranging from page-accessing to querying information, offers new grounds for exploratory work on web analytics (May & Lausen, 2004). Several grassroots techniques exist for data-mining website behavior, particularly when the webpage view behavior has been captured through well-planned queries (Mecca, Mendelzon, & Merialdo, 2002; Mena, 1999). Data preparation techniques that consolidate structured sampling and efficient collection schemes are deemed as highly desirable particularly when information is required for characterization of World Wide Web browsing patterns (Cooley, Mobasher, & Srivastava, 1999). Subsequently, the problem of screening and optimizing complex web-site properties has attracted modern data-processing treatments which indicate that there is a great anticipation for further research on this subject (Asllani & Lari, 2007). To assist in setting up screening studies, several web search quality measures have been proposed which lend themselves to efficient data mining