Advanced Engineering Informatics 27 (2013) 519–536

Contents lists available at SciVerse ScienceDirect

Advanced Engineering Informatics

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

Input variable selection in time-critical knowledge integration applications: A review, analysis, and recommendation paper

S. Tavakoli [a,*], A. Mousavi [b], S. Poslad [c]

[a] School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom
[b] School of Engineering and Design, Brunel University West London, United Kingdom
[c] School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom

ARTICLE INFO

Article history:
Received 14 May 2012
Received in revised form 13 June 2013
Accepted 14 June 2013
Available online 10 July 2013

Keywords:
Input variable selection
Time-critical control
Dimensionality reduction
Sensitivity analysis
Supervisory control and data acquisition

ABSTRACT

The purpose of this research is twofold: first, to undertake a thorough appraisal of existing Input Variable Selection (IVS) methods within the context of time-critical and computation resource-limited dimensionality reduction problems; second, to demonstrate improvements to, and the application of, a recently proposed time-critical sensitivity analysis method called EventTracker to an environment science industrial use-case, i.e., sub-surface drilling.

Producing time-critical accurate knowledge about the state of a system (effect) under computational and data acquisition (cause) constraints is a major challenge, especially if the knowledge required is critical to the system operation where the safety of operators or integrity of costly equipment is at stake. Understanding and interpreting, a chain of interrelated events, predicted or unpredicted, that may or may not result in a specific state of the system, is the core challenge of this research. The main objective is then to identify which set of input data signals has a significant impact on the set of system state information (i.e. output). Through a cause-effect analysis technique, the proposed technique supports the filtering of unsolicited data that can otherwise clog up the communication and computational capabilities of a standard supervisory control and data acquisition system.

The paper analyzes the performance of input variable selection techniques from a series of perspectives. It then expands the categorization and assessment of sensitivity analysis methods in a structured framework that takes into account the relationship between inputs and outputs, the nature of their time series, and the computational effort required. The outcome of this analysis is that established methods have a limited suitability for use by time-critical variable selection applications. By way of a geological drilling monitoring scenario, the suitability of the proposed EventTracker Sensitivity Analysis method for use in high volume and time critical input variable selection problems is demonstrated.

1. Introduction

In a complex interrelated world, industry is faced with ever changing performance metrics. This complexity and relentlessness of change both in substance and in presentation is forcing companies to make ever larger investments in data acquisition and processing technologies. The dilemma of "usefulness" and "relevance" for the investor still prevails [8,38]. In addition, there is a direct relationship between identifying useful and relevant input data and the levels of investment on data acquisition, communication and computational capabilities required for performance measurement. The identification of useful and relevant input data that affects the performance measurement relies on the speed and the quality of the process that separates the useful and relevant data from the non-useful non-relevant input data.

This paper reports on the existing techniques that have been developed in recent years for measuring the degree of usefulness and relevancy of input data in describing the state of system or how sensitivity the state is to specified parameters. These techniques include the traditional Input Variable Selection (IVS), Feature Selection (FS), and Sensitivity Analysis (SA). The authors approach for analyzing and comparing the applicability and practicability of such techniques is to determine if such techniques are capable of dealing with extremely volatile cases in industry which require tracking of real-time events and their importance as they occur. We also briefly refer to the computational restrictions that traditional IVS, FS and SA techniques have that make them irrelevant for the type of systems we deal with.

A case study in the Deep Drilling Industry (Deep Drilling Disaster Prevention) is presented in this article. It demonstrates the

* Corresponding author. Tel.: +44 795 209 4444; fax: +44 207 882 7997.
E-mail address: siamakt@eecs.qmul.ac.uk (S. Tavakoli).

1474-0346/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aei.2013.06.002

ability of a recently proposed real-time sensitivity analysis technique EventTracker [65] in generating accurate disaster knowledge support system for the purpose of Disaster Prevention as a unique solution for dealing with time-critical and unaware situations. The emphasis on "unaware" is to insist that in such cases there is no historical knowledge of the system behavior. In this particular case a drilling machine drilling deep into a geophysical unknown or volatile terrain.

Note that the purpose of the literature review and analysis is reporting on the existing IVS, FS and SA techniques and to demonstrate the impracticality of these techniques in dealing with such situations, because they normally rely on historical knowledge (e.g. statistics and probability), postulation or heuristics to manage and interpret input data and assemble system state information. The example offered in this article demonstrates a real industrial case that no traditional input variable analysis technique can be employed without major compromises to the quality of data or decision making whilst drilling.

The remainder of this article is structured as follows. First, a cross-conceptual survey of system 'variable' and 'feature' extraction is given in Section 2. In Section 3, a comprehensive spectrum of perspectives is provided to summarize IVS methodologies. In Section 4, the authors focus on how sensitivity analysis supports IVS. This is followed in Section 5 by a description of application of EventTracker sensitivity analysis in a detailed case study and the demonstration of its feasibility based upon the results. Finally, the conclusions of the research project are provided in Section 6.

2. IVS and feature selection

Although the concepts of Input Variable Selection (IVS) and feature selection [49] appear similar, there are key differences between them. Here an analysis of these key aspects such as dimensionality reduction and the associated computational overhead is carried out.

Feature Selection (FS) is a well-known problem addressed by much research [13,16,24,26,27,30,53,71,72]. The objective of this paper is to propose a technique to decide which input data sources are useful and relevant to determine the state of a given system, e.g. output. For this purpose we distinguish between FS and IVS with respect to two aspects; first the nature or substance of the input variables and features, and second, how they are selected. Both aspects are explained as follows.

2.1. Input variables and features

Although the terms "variable" and "feature" are sometimes regarded to be quite similar, input variables differ in nature from features. Input variables and raw input data series generally provide information about the system, whereas features are used to represent a model of the system. A feature may be locally and temporarily created to support the understanding of some specific behavior in the system. Based on the problem in hand, non-sequential data, i.e. snapshots, may be sufficient for a given data mining task. However, continuous information of input variables is time-critical in certain applications when the system state changes rapidly in time, i.e. in Volatile Systems [28]. This conceptual difference is shown in Fig. 1.

Input variables represent knowledge about the direct result of aggregation – and pre-filtering - of raw input data from data sources. Variable construction resembles data fusion as an overlay task in terms of aggregating raw input data sources. Features, however, are meant to add to the aggregated knowledge when input variables do not provide the required knowledge with adequate certainty, efficiency or based upon other domain-specific objectives.

Features are created in two ways:

- They are created either by application of specific transfer functions based on the input variables, raw input data, or a combination of both. Feature construction extends the knowledgebase.
- Or they are created through finding patterns in a data series.

With the help of an example, we describe feature construction and feature extraction as follows.

2.1.1. Features derived from input variables (feature construction)

When a system is monitored, the key features that describe the system status are separated from the initial input variables that are generated by situated sensors. Each constitutes a distinct layer in the data acquisition architecture, i.e. the sensor level is taken to be the primary layer and the feature level occurs at an intermediate layer [1]. For example, the input variables for a typical production line are specified to assemble the initial feature candidates that in turn are interpreted as indicators of the shop floor status and the job characteristics [41]. Transformations and combinations of the primary data and intermediate features are then used to extract the key performance indicators [41].

In another example, in order to detect faults in rechargeable batteries i.e. state, two performance indicators are required, capacity and life cycle [52]. These two indicators are measured based upon the amount of charge and the number of completed charge/discharge cycles (i.e. input variables) prior to the nominal capacity falling below a specified value. In order to reduce the measurement cycles and to facilitate the detection process, the paper reports on devising a new set of variables (features) by combining the derivatives of the two variables with different orders.

Fig. 2 illustrates the feature construction from input variables. The curved arrows symbolize the extra effort spent in converting input variable type knowledge into feature type knowledge. Depending on the type of process, one can expect increases in the computational overhead. The cost of the computational overhead may affect subsequent control, monitoring, and decision making processes, particularly from a time-critical perspective.

2.1.2. Features based upon data mining (feature extraction)

The key challenge in any data mining and analysis process is to decrease the uncertainty associated with the relationship between input variables and the interpretation of system behavior. It is not always the case that input variables can provide a sufficiently clear view of a system's behavior [12,20]. Neither can input variables always lead to an adequate model that can be built and evaluated. An analysis of the time series of the acquired data (and input variables) can enable features of data sources, that are not previously known and actionable, to be revealed. For example, in order to understand the complex characteristics of bioprocesses and enhance production robustness, both descriptive (e.g., frequent pattern discovery, clustering) and predictive (e.g. classification, regression) pattern recognition methods have been applied [15]. As a result, significant trends in processing data sourced from archived temporal records of physical parameters and production scale data were discovered [15].

Furthermore, a Virtual Metrology system capable of predicting each wafer's metrology measurements based upon production equipment data and metrology results collected for statistics, such as mean, variance, minimum, and maximum values from each sensor during two etching processes as used by a Korean semiconductor manufacturing company is presented [36].