International Journal of Innovative
Computing, Information and Control
Volume 4, Number 6, June 2008

ICIC International ©2008 ISSN 1349-4198
pp. 1507–1526

TECHNIQUES FOR HANDLING MISSING DATA: APPLICATIONS TO ONLINE CONDITION MONITORING

Fulufhelo Vincent Nelwamondo and Tshilidzi Marwala

School of Electrical and Information Engineering
University of the Witwatersrand
Johannesburg, South Africa
{ f.nelwamondo; t.marwala }@ee.wits.ac.za

Received May 2007; revised October 2007

Abstract. *The use of inferential sensors is a common task in online fault detection in various control applications. A problem arises when sensors fail while the control system is designed to make a decision based on the data from those sensors. Various techniques to handle missing data are discussed in this paper. Firstly, a novel algorithm that classifies and regresses in the presence of missing data is proposed. The algorithm is tested for both classification and regression problems. Secondly, an estimation algorithm that uses an ensemble of regressors is proposed. Hybrid genetic algorithms and fast simulated annealing are used to predict the missing values and their results are compared. Results show that fast simulated annealing is slightly faster than the hybrid GA for the problem investigated. Results provide a valuable insight into dealing precisely with missing data.*
Keywords: Ensemble, Fast simulated annealing, Hybrid genetic algorithms, Missing data

1. **Introduction.** One of the biggest problems hindering the performance of online condition monitoring is dealing with missing data. Missing data often result due to sensor failure in online condition monitoring systems. This paper looks at a problem of condition monitoring where computational intelligence techniques are used to classify and regress in the presence of missing data. The research in this paper is largely motivated by the question: *How do we perform classification or regression with an incomplete input vector during online operation ?* The biggest challenge is that the standard neural networks are not able to process input data with missing values and hence, can not perform classification or regression when some input data are missing. Online condition monitoring uses time series data. More often,there is limited time between the readings depending on how frequently the sensor is sampled. As a result, missing data become a huge obstacle in deciding the condition of the machine being monitored because there is a small and limited time between readings. In both classification and regression tasks, all decisions concerning how to proceed must be taken during this finite time period.

There are three general ways that have been used to deal with the problem of missing data [1]. The simplest method is known as 'listwise deletion' and this method simply deletes instances with missing values [1]. The major disadvantage of this method is the dramatic loss of information in data sets. In their research, Kim and Curry [2] found that when 2% of the features are missing and the complete observation is deleted, up to 18% of the total data may be lost. Another disadvantage of 'listwise deletion' is that it assumes that the observation with missing values is not important and can be ignored. As a result, the condition of the machine being monitored can not be detected should one or more sensors fail. The second common technique imputes the data by finding estimates of the values and missing entries are replaced with these estimates. The estimates vary with