Chapter 7
Robust Statistical Methods for Rapid Data Labelling

Jamie Godwin
University of Durham, UK

Peter Matthews
University of Durham, UK

ABSTRACT

Labelling of data is an expensive, labour-intensive, and time consuming process and, as such, results in vast quantities of data being unexploited when performing analysis through data mining. This chapter presents a new paradigm using robust multivariate statistical methods to encapsulate normal operational behaviour—not failure behaviour—to autonomously derive unsupervised classifier labels for previously collected data in a rapid, cost-effective manner. This enables traditional machine learning to take place on a much richer dataset. Two case studies are presented in the mechanical engineering domain, namely, a wind turbine gearbox and a rolling element bearing. A statistically sound and robust methodology is contributed, allowing for rapid labelling of data to enable traditional data mining techniques. Model development is detailed, along with a comparative evaluation of the metrics. Robust derivatives are presented and their superiority is shown. Example "R" code is given in the appendix, allowing readers to employ the techniques discussed. High levels of agreement between the derived statistical approaches and the underlying condition of the components can be found, showing the practical nature and benefit of this approach.

INTRODUCTION

Many data-driven algorithms require accurate labels in order to encapsulate the various conditions which correspond to their meaning. However, in many real-world applications, deriving these labels is not possible in practice or is not economically viable due to the amount of resources required. As such, although significant quantities of data exist, exploiting this data effectively is not a trivial problem.

In this chapter, an evaluation of the performance of six multivariate distance metrics on two datasets incorporating censored data is presented. Unlike traditional methods, the techniques evaluated in this chapter provide a means

DOI: 10.4018/978-1-4666-6086-1.ch007

of autonomously deriving classifier labels in an unsupervised manner for previously collected data in a rapid, cost-effective way. This can be employed in cases where previously labelled data is either scarce, or highly imbalanced – allowing a greater amount of data to be incorporated into a traditional data mining analysis. To aid in the practicality and demonstrate the soundness of the approaches detailed, 'R' code is provided at all stages of the analysis. This allows the reader to follow the examples, as the techniques are presented on publicly available data for tutorial purposes.

The remainder of this book chapter is organized as follows. Motivation and context for this work is presented in the "background to the problem" section, along with the issues of data scarcity and data imbalance. Next, traditional techniques utilised are presented in the "data intensive techniques" section. The datasets and degradation models used are given in the "dataset description." In total, 6 multivariate distance metrics are introduced and comparatively evaluated on both datasets for their robustness and merit in performing condition assessment; three Minkowski distances (Manhattan, Euclidean and Chebyshev), the Penrose distance and two forms of the Mahalanobis distance are looked at in depth. Multivariate normality testing is then covered. After this, two case studies are presented in the "case study" sections, showing how the metrics can be employed for rapid labelling of data to enable traditional data mining approaches. Conclusions are then presented, with references and 'R' code in the appendices.

BACKGROUND TO THE PROBLEM

Motivation

Data mining can be thought of as an enabler of next generation maintenance techniques within the realm of reliability engineering due to the ability to generate significant cost savings. It is substantially cheaper to perform maintenance before an asset fails (preventive maintenance), rather than after a failure has occurred (corrective maintenance). As the cost of data acquisition technology and storage has reduced significantly, the quantity of data available for analysis has risen substantially. This provides many benefits, however, these are yet to be fully realised.

For instance, by utilising previously collected data in association with data mining techniques, it is possible to determine the current level of wear (or degradation) on an asset such as a turbine or a bearing. This enables preventive maintenance (Iung et al., 2009) to be performed at a significantly reduced cost (Wu & Clements-Croome, 2006) and thus provides a competitive edge to the corporation utilising these techniques (Leger et al., 1999).

However, the uptake of these techniques within the domain of reliability engineering has been slower than expected (Moore & Starr, 2006). This is due to many factors. For example:

- "Black box" expert systems which lack the necessary transparency,
- The financial outlay required to install the infrastructure to enable the data acquisition,
- Inherent uncertainty and varying accuracy present within data-driven processes and techniques,
- Staff training costs,
- No proven track record of the systems in similar domains.

Potentially the largest cause of uncertainty in these systems is caused due to inadequate labelling of the data, required in many cases to extract condition information using traditional data mining techniques. In many real world applications is not possible to know the required information for each data point (for instance, the size of the crack in a rotor shaft or the level of corrosion in a jet engine). Inspection of the equipment can provide this information but fundamentally changes the expected behaviour of the component due to the