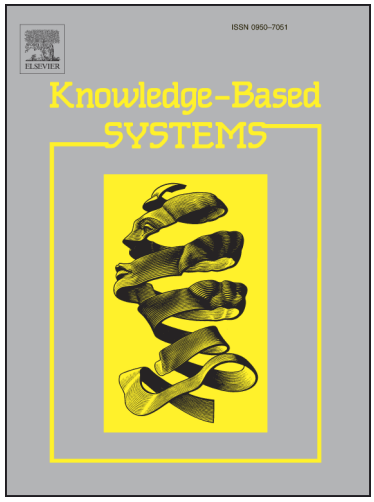

Recent advances and emerging challenges of feature selection in the context of Big Data

V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos

PII: S0950-7051(15)00200-2
DOI: http://dx.doi.org/10.1016/j.knosys.2015.05.014
Reference: KNOSYS 3157

To appear in: *Knowledge-Based Systems*

Received Date: 12 February 2015
Revised Date: 9 April 2015
Accepted Date: 11 May 2015

Please cite this article as: V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Recent advances and emerging challenges of feature selection in the context of Big Data, *Knowledge-Based Systems* (2015), doi: http://dx.doi.org/10.1016/j.knosys.2015.05.014

Recent advances and emerging challenges of feature selection in the context of Big Data

V. Bolón-Canedo*, N. Sánchez-Maroño, A. Alonso-Betanzos

Department of Computer Science, Universidade da Coruña, 15071 A Coruña, Spain

Abstract

In an era of growing data complexity and volume and the advent of big data, feature selection has a key role to play in helping reduce high-dimensionality in machine learning problems. We discuss the origins and importance of feature selection and outline recent contributions in a range of applications, from DNA microarray analysis to face recognition. Recent years have witnessed the creation of vast datasets and it seems clear that these will only continue to grow in size and number. This new big data scenario offers both opportunities and challenges to feature selection researchers, as there is a growing need for scalable yet efficient feature selection methods, given that existing methods are likely to prove inadequate.

Keywords: feature selection, big data, high dimensionality

1. Introduction

The "big data" phenomenon is unfolding before our eyes and its transformational nature is unquestionable. Between the dawn of time up to 2003 humanity generated a total of 5 exabytes of data and by 2008 this figure had tripled to 14.7 exabytes. Nowadays 5 exabytes of data is produced every 2 days —and the pace of production continues to rise. Because the volume, velocity, variety and complexity of datasets is continuously increasing, machine learning techniques have become indispensable in order to extract useful information from huge amounts of otherwise meaningless data. One machine learning technique is feature selection

*Corresponding author.

Email addresses: vbolon@udc.es (V. Bolón-Canedo), nsanchez@udc.es (N. Sánchez-Maroño), ciamparo@udc.es (A. Alonso-Betanzos)