Data Mining of Virtual Campus Data

Alfredo Vellido[1], Félix Castro[1, 2], Terence A. Etchells[3], Àngela Nebot[1], and Francisco Mugica[4]

[1] Dept. Llenguatges i Sistemes Informàtics, LSI, Universitat Politècnica de Catalunya, Campus Nord, C. Jordi Girona 1-3, Barcelona 08034, Espanya {avellido, fcastro, angela}@lsi.upc.edu

[2] Centro de Investigación en Tecnologías de Información y Sistemas, CITIS, Universidad Autónoma del Estado de Hidalgo, Ciudad Universitaria, Carretera Pachuca-Tulancingo km. 4.5, Hidalgo, México

[3] School of Computing and Mathematical Sciences, Liverpool John Moores University, Byrom St., L3 3AF, Liverpool, UK t.a.etchells@livjm.ac.uk

[4] Instituto Latinoamericano de la Comunicación Educativa (ILCE), Calle del Puente 45, México D F 14380, México fmugica@ilce.edu.mx

Summary. As mentioned elsewhere in this book, e-learning offers "a new context for education where large amounts of information describing the continuum of the teaching–learning interactions are endlessly generated and ubiquitously available". But raw information by itself may be of no help to any of the e-learning actors. The use of Data Mining methods to extract knowledge from this information can, therefore, be an adequate approach to follow in order to use the obtained knowledge to fit the educational proposal to the students' needs and requirements. This chapter provides a case study in which several advanced Data Mining techniques are employed to extract different types of knowledge from virtual campus data concerning students system usage behaviour. The diverse palette of Data Mining problems addressed here include data clustering and visualization, outlier detection, classification, feature selection, and rule extraction. They concern diverse e-learning problems, such as the characterization of atypical students' behaviour and the prediction of students' performance. Different Data Mining techniques from the areas of Statistical and Machine Learning; Fuzzy Logic, and Inductive Reasoning are

A. Vellido et al.: *Data Mining of Virtual Campus Data,* Studies in Computational Intelligence (SCI) **62**, 223–254 (2007)
www.springerlink.com

employed to tackle these problems. Strong emphasis is placed on the interpretability of the results, obtained through rule extraction, so that they can be fed back to the e-learning system in a practical an efficient manner.

1 Introduction

Any e-learning system is, by its own nature, likely to generate large amounts of information describing the continuum of the teaching-learning interactions almost in real time. All this information, gathered from diverse and usually heterogeneous sources, may be of no help by itself to any of the e-learning actors in its raw form. Actually, an excess of such information can become a liability for e-learning tutors and managers unless it is processed according to reasonable goals. Data Mining can provide the adequate tools for such processing; obtaining actionable patterns from large data repositories. The use of Data Mining methods to extract knowledge from the e-learning system available information can, therefore, be an adequate approach to follow in order to use the obtained knowledge to fit the educational proposal to the students' needs and requirements.

Virtual campus environments, such as the one that is the subject of this case study, are fastly becoming a mainstream alternative to traditional distance higher education. The Internet medium they use to convey content, also allows the gathering of information on students' online behaviour. Here, we focus on e-learning systems improvement through the analysis of the data generated by the virtual campus students, aiming to discover their system usage patterns.

The amount of research concerned with the mining of data generated by the usage of e-learning systems is still somehow scarce on the ground (see Castro et al., this book). In this study, we address two main problems concerning virtual campus students' behaviour: the characterization of atypical behaviour and the prediction of students' performance. Several Data Mining problems are concerned, namely: students' data clustering and visualization, behavioural outlier detection, students' classification according to course marks, data feature selection, and rule extraction. The latter becomes of paramount importance as we aim to place a strong emphasis on the interpretability of the obtained results; no matter how accurate these might be: unless they are translated into practical and efficient rules that system managers and tutors can act upon, it will be extremely difficult to feed them back to the e-learning system.

The rest of the chapter is structured as follows: First, in Sect. 2, we provide a detailed description of the data under analysis; they correspond to