
AVVISO DI SEMINARI

Prof. Luís Torgo

Department of Computer Science, Faculty of Sciences, University of Porto (Portugal)
Venerdì 11 giugno 2010 ore 10.30, Aula Gödel, II piano, Dipartimento di Informatica

RESOURCE-BOUNDED OUTLIER DETECTION USING CLUSTERING METHODS

This presentation describes a methodology for the application of hierarchical clustering methods to the task of outlier detection. The methodology is tested on the problem of cleaning Official Statistics data. The goal is to detect erroneous foreign trade transactions in data collected by the Portuguese Institute of Statistics (INE). These transactions are a minority, but still they have an important impact on the statistics produced by the institute. The detection of these rare errors is a manual, time-consuming task. This type of tasks is usually constrained by a limited amount of available resources. Our proposal addresses this issue by producing a ranking of outlyingness that allows a better management of the available resources by allocating them to the cases which are most different from the other and, thus, have a higher probability of being errors. Our method is based on the output of standard agglomerative hierarchical clustering algorithms, resulting in no significant additional computational costs. Our results show that it enables large savings by selecting a small subset of suspicious transactions for manual inspection, which, nevertheless, includes most of the erroneous transactions. In this study we compare our proposal to a state of the art outlier ranking method (LOF) and show that our method achieves better results on this particular application. The results of our experiments are also competitive with previous results on the same data. Finally, the outcome of our experiments raises important questions concerning the method currently followed at INE concerning items with small number of transactions.



Luis Torgo has a degree in Systems and Informatics Engineering and a PhD in Computer Science. He is currently an Associate Professor of the Department of Computer Science of the Faculty of Sciences of the University of Porto. He is also a researcher of the Laboratory of Artificial Intelligence and Data Analysis (LIAAD) belonging to INESC Porto LA. Luis Torgo has been an active researcher in Machine Learning and Data Mining for more than 20 years. He has lead several academic and industrial Data Mining research projects. He is the author of a book on “Data Mining with R, learning by case studies”, CRC Press (2010).

Seminario per: 1) Dottorato in Informatica; 2) Progetto formativo PS_121.