Mathematical and Computer Modelling 57 (2013) 1633–1639

Contents lists available at SciVerse ScienceDirect

Mathematical and Computer Modelling

journal homepage: www.elsevier.com/locate/mcm

Post-processing: Bridging the gap between modelling and effective decision-support. The profile assessment grid in human behaviour

K. Gibert [a,b,*], G. Rodríguez-Silva [b], R. Annicchiarico [c]

[a] Dep. Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona-Tech. C. Jordi Girona 1-3, Barcelona 08034, Spain

[b] Knowledge Engineering and Machine Learning group, Universitat Politècnica de Catalunya, Barcelona-Tech. C. Jordi Girona 1-3, Barcelona 08034, Spain

[c] IRCCS Fondazione Santa Lucia, via Ardeatina 306, 00179 Rome, Italy

ARTICLE INFO

Article history:
Received 10 October 2011
Received in revised form 26 October 2011
Accepted 27 October 2011

Keywords:
Data mining
Knowledge discovery from databases
Clustering
Logistic regression
Profiles assessment grid
Post-processing
Decision-support
Human behaviour

ABSTRACT

The importance of post-processing the results of clustering when using data mining to support subsequent decision-making is discussed. Both the formal *embedded binary logistic regression (EBLR)* and the visual *profile's assessment grid (PAG)* methods are presented as bridging tools for the real use of clustering results. EBLR is a sequence of logistic regressions that helps to predict the class of a new object; while PAG is a graphical tool that visualises the results of an EBLR. PAG interactively determines the most suitable class for a new object and enables subsequent follow-ups. PAG makes the underlying mathematical model (EBLR) more understandable, improves usability and contributes to bridging the gap between modelling and decision-support. When applied to medical problems, these tools can perform as diagnostic-support tools, provided that the predefined set of profiles refer to different stages of a certain disease or different types of patients with a same medical problem, etc. Being a graphical tool, PAG enables doctors to quickly and friendly determine the profile of a patient in the everyday activity, without necessarily understanding the statistical models involved in the process, which used to be a serious limitation for wider application of these methods in clinical praxis. In this work, an application is presented with 4 functional disability profiles.

1. Introduction

Knowledge discovery from data (KDD) is a discipline established by Fayyad in 1989 for: *"The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"* [1]. KDD has quickly grown as a multidisciplinary research field, where advanced techniques from statistics, artificial intelligence, information systems, visualisation ··· are combined to provide effective knowledge acquisition from huge data bases (with dimensions never imagined before the Internet boom). According to Fayyad, KDD refers to high level applications that include concrete methods of *data mining*: *"the overall process of finding and interpreting patterns from data, typically interactive and iterative, involving repeated application of specific data mining methods or algorithms and the interpretation of the patterns generated by these algorithms"*. More than 20 years later, KDD also appears as a powerful methodological framework for modelling very complex phenomena or organisations (such as environmental processes or health-care systems) even when no massive data

* Correspondence to: Knowledge Engineering and Machine Learning group, Universitat Politècnica de Catalunya, Barcelona-Tech. C. Jordi Girona 1-3, Barcelona 08034, Spain. Tel.: +34 669250864.
E-mail address: karina.gibert@upc.edu (K. Gibert).

0895-7177/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.mcm.2011.10.046

sets are available [2]; this might respond to the intrinsic KDD multidisciplinary approach, but also to the importance given to what happens before and after the analysis itself. In fact, KDD is marking the beginning of a new methodological paradigm: *"Most previous work on KDD has focussed on [· · ·] the data mining step. However, the other steps are of considerable importance for the successful application of KDD in practice* [1]". Indeed, prior and posterior analyses are essential to guarantee: (i) *correct and valid results:* proper data cleaning and data preparation is crucial for correctness, while the accurate interpretation of results enables a complete validation process; (ii) *real impact on the target domain:* even when results are correct and valid, intense post-processing is often required to make results understandable and useable by end-users (often lacking strong mathematical skills).

The benefits of the multidisciplinary approach that is typically used in KDD as well as the synergies of AI methods and statistics for modelling different complex domains, are in [3,4]. Despite the proliferation of new data mining methods, recent research [2] has shown that only a restricted set of DM methods (the more popular and simple) is being used in practice. [2] indicates a preference for applying qualitative models, such as decision trees or rules induction rather than regression or ANOVA. We presume this is more due to the understandability and usability of the final results than to the intrinsic performance of the model. A convenient post-processing of traditional statistical models can bring the results closer to non-expert users and make friendly a set of mathematical equations, often avoided by many decision-makers. It seems clear that effective decision support on the target domain is strongly affected by the understandability of the model. The importance of pre and post-processing steps is clearly recognised in the scientific community [2,5]. However, such steps are currently undertaken in an informal manner in practice, and more research is required to systematise them [2]. The consequences of neglecting preprocessing in clustering applications are analysed in [6,7]. Although this is a general problem, in this work we illustrate the importance of the post-processing step and its impact on the usability of a mathematical model in the context of clustering and logistic regression.

As organizations and systems become larger and more complex, managing them becomes an increasingly difficult activity. Understanding organizations and systems is crucial for management and decision support tools become increasingly important in decision-making processes. In many situations, understanding is very much improved by a profiling model, that can identify typical patterns or entities in the system and associate standard actions, protocols, treatments or decisions to every profile. Many of our previous works involved profiling and, in consequence, clustering techniques. Also, in [8] clustering appears as the most used DM method for KDD in unsupervised contexts.

However, clustering results consist of a list of classes and their object components. The gap between finding the best clustering result and being able to use it for everyday decision-making is enormous. The proposal presented in this work tries to reduce this gap. Usability requires understanding of the profiles, establishing a protocol/model to recognise them, and providing a friendly tool to predict profiles for new objects. For these needs, we have developed a tool that, given a set of profiles in a domain obtained using a clustering process, can easily, and quickly predict the profile of a new entity.

The problem is presented in Section 2. In Section 3 we present the *embedded binary logistic regression (EBLR)* method to post-process clustering as a combination of statistical models to recognise the classes. The *profile's assessment grid (PAG)* is introduced in Section 4 as a visually friendly alternative for non-expert users. It is an interpretative tool that graphically enables the immediate identification of the profile corresponding to a single object. It is useful for predictive purposes (diagnoses or classification), and can be used in clinical praxis. In Section 5 EBLR and PAG are used to find patterns of functional dependency in elderly patients, and Section 6 discusses conclusions and future work.

2. The problem

Once the clustering process is completed a class is associated with every object. A proposal to recognise the class of a new object in a friendly way is presented. The idea is to initially find a formal model for predicting the class that can then be transported into a friendly graphical representation for everyday decision-making. Although modelling is used in first step and visualisation in the second, both are used here as post-processing tasks over the clustering results. The problem can be formulated as follows:

Given:

- a universe $\mathcal{U}$
- a set of objects $\mathcal{I} = \{i_1 \cdots i_n\} \subseteq \mathcal{U}$
- a set of variables $X = (X_1 \cdots X_K)$ describing the object of $\mathcal{I}$
- a set of classes $\mathcal{P} = \{C_1 \cdots C_\xi\}$ such that:
  - $\mathcal{P}$ defines a partition over $\mathcal{I}$
  - R is a total order over $\mathcal{P}$ such that $C_{1R} > C_{2R} > \cdots C_{\xi-1R} > C_\xi$
- The application $Q: \mathcal{I} \longrightarrow \mathcal{P}$ such that $Q(i) = C: i \in C$.

Find

(1) A function f such that: $\hat{Q}(i) = f(X_P, i)$ being $X_P \subseteq X$ a subset of relevant variables to recognise the class of an object and $\hat{Q}(i)$ the estimate of $Q(i)$
(2) The extension of Q to Q^*: $\mathcal{U} \longrightarrow \mathcal{P}$ such that $Q^*(i) = f(X_P, i), \forall i \in \mathcal{U}$
(3) A graphical tool for Q^*.

The *EBLR* method is proposed as a solution for points 1 and 2, under the assumption of *R*. The *PAG* is our proposal for point 3, when $card(\mathcal{P}) \leq 4$.