Raport Badawczy

Research Report

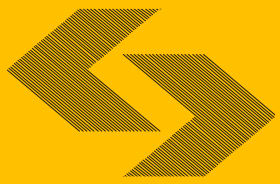
RB/67/2008

Book review: Cluster analysis for data mining and system identification by János Abonyi and Balázs Feil

J. W. Owsiński

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences



Control and Cybernetics

vol. **37** (2008) No. 1

Book review:

CLUSTER ANALYSIS FOR DATA MINING AND SYSTEM IDENTIFICATION

by

János Abonyi and Balázs Feil

The book is, actually, a monograph, whose title could have better been "Fuzzy k-Means Clustering Paradigm in Data Analysis and Modelling". Although, namely, a lot of other material is presented in the book (e.g. elements of ANN, Bayesian classification, PCA and related topics, etc.), the true backbone of the entire presentation and the "narrative" is the generalised fuzzy k-means approach and its various forms as well as applications.

The volume starts from a chapter, containing general introduction to cluster analysis, in which, of course, the algorithms associated with fuzzy k-means, are especially pronounced. This is particularly so with the Gath-Geva algorithm, which is made use of thereafter several times in the book within various contexts. Already in this chapter the situations are treated, in which observations are apparently generated by a process, having a definite and identifiable model. The second chapter is devoted to another classical subject of multidimensional data analysis, namely visualisation (here: of clustering results). Again, here also mainly the techniques associated with fuzzy representation (mapping) of data and their groups are put forward. The two subsequent chapters form the core of the book: fuzzy regression and identification of dynamical systems. In the latter chapter ample use is made of the fuzzy neural networks in the problem of model order selection. In Chapter 5 the authors return to another classical theme of multivariate analysis, namely to construction of classifiers. Fuzzy classifier is developed, also with the use of fuzzy decision tree approach. Finally, in the last chapter of the book segmentation of multivariate time series is taken up, whereby the Gath-geva algorithm returns as one of the fundamental tools.

As mentioned, the book is in fact a monograph on the diverse shapes and applications of the k-means-based fuzzy clustering paradigm, which takes very different forms in dependence upon the field of application and thus the technical and mathematical context. A reader is led along a relatively narrow, but smooth path, winding among various problem-defined landscapes. On this path milestones are constituted by the (somewhat cursory) algorithm descriptions, of which there are indeed many. The book is very well illustrated, with an ample bibliography of 302 positions, and an index.

Due to the formula of the book many issues and methods are, on the way, left untouched. It would also be valuable, if the methods that are not treated were somehow compared with the ones presented (although it is true that this is not always possible in view of the widely differing characteristics of the respective methods). To some extent, one has the impression that the book is based on the principle of having a good, robust and rather general method (a tool) and showing how to use it in every imaginable situation (task). So, in many cases one is left without a convincing answer to the question: why thus? To start with, not in every situation reference to fuzzy rather crisp methods and representations brings positive results. It is obvious that the use of fuzzy methods involves in virtually all cases additional parameters (exponents, cut-off levels) that have to be user-defined ("the assumed degree of fuzziness") and to a large extent determine the character of results.

On the other hand, the book is much more convincing on the subject of "model identification" – even with the reservations mentioned above – than with respect to "data mining". The domain of data mining, namely, developed with two issues as the motive force: the vast volumes of data and the desire to "automatically" reconstruct "knowledge" on the basis of these vast volumes of data, this knowledge taking on a possibly simple (preferably rule-like) form. Out of this conjunction the book misses thorough analysis of computational complexity, which is of paramount importance for the large data sets, which can often be scanned just once, or at most a small number of times.

To sum up, I am glad I had the opportunity to read the book, which presents a rich repository of versatile applications of a powerful paradigm, some of them, especially those related to modelling, having in my opinion quite a special value. Every specialist in advanced data analysis will find the book not only interesting, but also a valuable and well-organised source of additional know-how among the existing multiplicity of techniques.

Jan W. Owsiński

János Abonyi, Balázs Feil, *Cluster Analysis for Data Mining and System Identification*. Birkhäuser Verlag, Basel–Boston–Berlin, 2007. XVIII+303 pages, ISBN-13: 978-3-7643-7987-2. Price: 99 EUR (hardcover)