

# Preface

Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait – often proximity according to some defined distance measure.

The aim of this book is to illustrate that advanced fuzzy clustering algorithms can be used not only for partitioning of the data, but it can be used for visualization, regression, classification and time-series analysis, hence fuzzy cluster analysis is a good approach to solve complex data mining and system identification problems.

## Overview

In the last decade the amount of the stored data has rapidly increased related to almost all areas of life. The most recent survey was given by Berkeley University of California about the amount of data. According to that, data produced in 2002 and stored in pressed media, films and electronics devices only are about 5 exabytes. For comparison, if all the 17 million volumes of Library of Congress of the United States of America were digitalized, it would be about 136 terabytes. Hence, 5 exabytes is about 37,000 Library of Congress. If this data mass is projected into 6.3 billion inhabitants of the Earth, then it roughly means that each contemporary generates 800 megabytes of data every year. It is interesting to compare this amount with Shakespeare's life-work, which can be stored even in 5 megabytes. It is because the tools that make it possible have been developing in an impressive way, consider, e.g., the development of measuring tools and data collectors in production units, and their support information systems. This progress has been induced by the fact that systems are often been used in engineering or financial-business practice that we do not know in depth and we need more information about them. This lack of knowledge should be compensated by the mass of the stored data that is available nowadays. It can also be the case that the causality is reversed: the available data have induced the need to process and use them,

e.g., web mining. The data reflect the behavior of the analyzed system, therefore there is at least the theoretical potential to obtain useful information and knowledge from data. On the ground of that need and potential a distinct science field grew up using many tools and results of other science fields: *data mining* or more general, *knowledge discovery in databases*.

Historically the notion of finding useful patterns in data has been given a variety of names including data mining, knowledge extraction, information discovery, and data pattern recognition. The term data mining has been mostly used by statisticians, data analysts, and the management information systems communities. The term knowledge discovery in databases (KDD) refers to the overall process of discovering knowledge from data, while data mining refers to a particular step of this process. Data mining is the application of specific algorithms for extracting patterns from data. The additional steps in the KDD process, such as data selection, data cleaning, incorporating appropriate prior knowledge, and proper interpretation of the results are essential to ensure that useful knowledge is derived from the data. Brachman and Anand give a practical view of the KDD process emphasizing the interactive nature of the process [51]. Here we broadly outline some of its basic steps depicted in Figure 1.

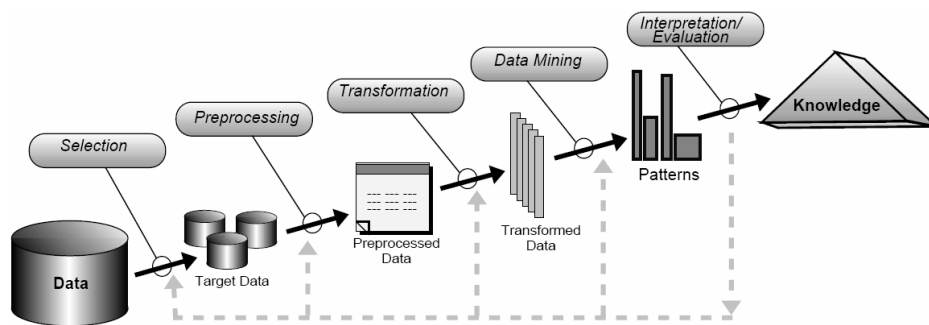


Figure 1: Steps of the knowledge discovery process.

1. *Developing and understanding of the application domain and the relevant prior knowledge, and identifying the goal of the KDD process.* This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives. The first objective of the data analyst is to thoroughly understand, from a business perspective, what the client really wants to accomplish. A business goal states objectives in business terminology. A data mining goal states project objectives in technical terms. For example, the business goal might be “Increase catalog sales to existing customers”. A data mining goal might be “Predict how many widgets a customer will buy, given their purchases over

the past three years, demographic information (age, salary, city, etc.) and the price of the item.” Hence, the prediction performance and the understanding of the hidden phenomenon are important as well. To understand a system, the system model should be as transparent as possible. The model transparency allows the user to effectively combine different types of information, namely linguistic knowledge, first-principle knowledge and information from data.

2. *Creating target data set.* This phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
3. *Data cleaning and preprocessing.* The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modelling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modelling tools. Basic operations such as the removal of noise, handling missing data fields.
4. *Data reduction and projection.* Finding useful features to represent the data depending on the goal of the task. Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representation of data. Neural networks, cluster analysis, and neuro-fuzzy systems are often used for this purpose.
5. *Matching the goals of the KDD process to a particular **data mining method**.* Although the boundaries between prediction and description are not sharp, the distinction is useful for understanding the overall discovery goal. The goals of data mining are achieved via the following data mining tasks:
  - *Clustering:* Identification a finite set of categories or clusters to describe the data. Closely related to clustering is the method of probability density estimation. Clustering quantizes the available input-output data to get a set of prototypes and use the obtained prototypes (signatures, templates, etc.) as model parameters.
  - *Summation:* Finding a compact description for subset of data, e.g., the derivation of summary for association of rules and the use of multivariate visualization techniques.
  - *Dependency modelling:* finding a model which describes significant dependencies between variables (e.g., learning of belief networks).
  - *Regression:* Learning a function which maps a data item to a real-valued prediction variable based on the discovery of functional relationships between variables.

- *Classification*: learning a function that maps (classifies) a data item into one of several predefined classes (category variable).
  - *Change and Deviation Detection*: Discovering the most significant changes in the data from previously measured or normative values.
6. *Choosing the data mining algorithm(s)*: Selecting algorithms for searching for patterns in the data. This includes deciding which model and parameters may be appropriate and matching a particular algorithm with the overall criteria of the KDD process (e.g., the end-user may be more interested in understanding the model than its predictive capabilities.) One can identify three primary components in any data mining algorithm: model representation, model evaluation, and search.
- *Model representation*: The natural language is used to describe the discoverable patterns. If the representation is too limited, then no amount of training time or examples will produce an accurate model for the data. Note that more flexible representation of models increases the danger of overfitting the training data resulting in reduced prediction accuracy on unseen data. It is important that a data analysts fully comprehend the representational assumptions which may be inherent in a particular method.  
For instance, rule-based expert systems are often applied to classification problems in fault detection, biology, medicine etc. Among the wide range of computational intelligence techniques, fuzzy logic improves classification and decision support systems by allowing the use of overlapping class definitions and improves the interpretability of the results by providing more insight into the classifier structure and decision making process. Some of the computational intelligence models lend themselves to transform into other model structure that allows information transfer between different models (e.g., a decision tree mapped into a feedforward neural network or radial basis functions are functionally equivalent to fuzzy inference systems).
  - *Model evaluation criteria*: Qualitative statements or fit functions of how well a particular pattern (a model and its parameters) meet the goals of the KDD process. For example, predictive models can often be judged by the empirical prediction accuracy on some test set. Descriptive models can be evaluated along the dimensions of predictive accuracy, novelty, utility, and understandability of the fitted model. Traditionally, algorithms to obtain classifiers have focused either on accuracy or interpretability. Recently some approaches to combining these properties have been reported
  - *Search method*: Consists of two components: parameter search and model search. Once the model representation and the model evaluation criteria are fixed, then the data mining problem has been reduced

to purely an optimization task: find the parameters/models for the selected family which optimize the evaluation criteria given observed data and fixed model representation. Model search occurs as a loop over the parameter search method.

The automatic determination of model structure from data has been approached by several different techniques: neuro-fuzzy methods, genetic-algorithm and fuzzy clustering in combination with GA-optimization.

7. *Data mining*: Searching for patterns of interest in a particular representation form or a set of such representations: classification rules, trees or figures.
8. *Interpreting mined patterns*: Based on the results possibly return to any of steps 1–7 for further iteration. The data mining engineer interprets the models according to his domain knowledge, the data mining success criteria and the desired test design. This task interferes with the subsequent evaluation phase. Whereas the data mining engineer judges the success of the application of modelling and discovery techniques more technically, he/she contacts business analysts and domain experts later in order to discuss the data mining results in the business context. Moreover, this task only considers models whereas the evaluation phase also takes into account all other results that were produced in the course of the project. This step can also involve the visualization of the extracted patterns/models, or visualization of the data given the extracted models. By many data mining applications it is the user whose experience (e.g., in determining the parameters) is needed to obtain useful results. Although it is hard (and almost impossible or senseless) to develop totally automatical tools, our purpose in this book was to present as data-driven methods as possible, and to emphasize the transparency and interpretability of the results.
9. *Consolidating and using discovered knowledge*: At the evaluation stage in the project you have built a model (or models) that appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying “live” models within an organization’s decision making processes, for example in real-time personalization of Web pages or repeated scoring of marketing databases. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the

enterprise. In many cases it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models.

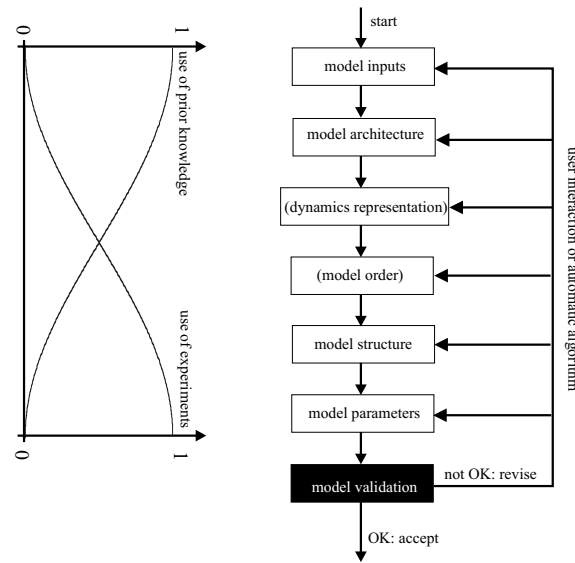


Figure 2: Steps of the knowledge discovery process.

Cross Industry Standard Process for Data Mining ([www.crisp-dm.org](http://www.crisp-dm.org)) contains (roughly) these steps of the KDD process. However, the problems to be solved and their solution methods in KDD can be very similar to those occurred in system identification. The definition of system identification is the process of modelling from experimental data by Ljung [179]. The main steps of the system identification process are summarized well by Petrick and Wigdorowitz [216]:

1. Design an experiment to obtain the physical process input/output experimental data sets pertinent to the model application.
2. Examine the measured data. Remove trends and outliers. Apply filtering to remove measurement and process noise.
3. Construct a set of candidate models based on information from the experimental data sets. This step is the model structure identification.
4. Select a particular model from the set of candidate models in step 3 and estimate the model parameter values using the experimental data sets.

5. Evaluate how good the model is, using an objective function. If the model is not satisfactory then repeat step 4 until all the candidate models have been evaluated.
6. If a satisfactory model is still not obtained in step 5 then repeat the procedure either from step 1 or step 3, depending on the problem.

It can be seen also in Figure 2 from [204] that the system identification steps above may roughly cover the KDD phases. (The parentheses indicate steps that are necessary only when dealing with dynamic systems.) These steps may be complex and several other problem have to be solved during one single phase. Consider, e.g., the main aspects influencing the choice of a model structure:

- What type of model is needed, nonlinear or linear, static or dynamic, distributed or lumped?
- How large must the model set be? This question includes the issue of expected model orders and types of nonlinearities.
- How must the model be parameterized? This involves selecting a criterion to enable measuring the closeness of the model dynamic behavior to the physical process dynamic behavior as model parameters are varied.

To be successful the entire modelling process should be given as much information about the system as is practical. The utilization of prior knowledge and physical insight about the system are very important, but in nonlinear black-box situation no physical insight is available, we have ‘only’ observed inputs and outputs from the system.

When we attempt to solve real-world problems, like extracting knowledge from large amount of data, we realize that there are typically ill-defined systems to analyze, difficult to model and with large-scale solution spaces. In these cases, precise models are impractical, too expensive, or non-existent. Furthermore, the relevant available information is usually in the form of empirical prior knowledge and input-output data representing instances of the system’s behavior. Therefore, we need an approximate reasoning systems capable of handling such imperfect information. computational intelligence (CI) and soft computing (SC) are recently coined terms describing the use of many emerging computing disciplines [2, 3, 13]. It has to be mentioned that KDD has evolved from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, artificial intelligence, and more recently it gets new inspiration from computational intelligence. According to Zadeh (1994): “. . . in contrast to traditional, hard computing, soft computing is tolerant of imprecision, uncertainty, and partial truth.” In this context Fuzzy Logic (FL), Probabilistic Reasoning (PR), Neural Networks (NNs), and Genetic Algorithms (GAs) are considered as main components of CI. Each of these technologies provide us with complementary *reasoning* and *searching* methods to solve complex, real-world problems. What is important to note is that soft

computing is not a melange. Rather, it is a partnership in which each of the partners contributes a distinct methodology for addressing problems in its domain. In this perspective, the principal constituent methodologies in CI are complementary rather than competitive.

Because of the different data sources and user needs the purpose of data mining and computational intelligence methods, may be varied in a range field. The purpose of this book is not to overview all of them, many useful and detailed works have been written related to that. This book aims at presenting new methods rather than existing classical ones, while proving the variety of data mining tools and practical usefulness.

The aim of the book is to illustrate how effective data mining algorithms can be generated with the incorporation of fuzzy logic into classical cluster analysis models, and how these algorithms can be used not only for detecting useful knowledge from data by building transparent and accurate regression and classification models, but also for the identification of complex nonlinear dynamical systems. According to that, the new results presented in this book cover a wide range of topics, but they are similar in the applied method: fuzzy clustering algorithms were used for all of them. Clustering within data mining is such a huge topic that the whole overview exceeds the borders of this book as well. Instead of this, our aim was to enable the reader to take a tour in the field of data mining, while proving the flexibility and usefulness of (fuzzy) clustering methods. According to that, students and unprofessionals interested in this topic can also use this book mainly because of the Introduction and the overviews at the beginning of each chapter. However, this book is mainly written for electrical, process and chemical engineers who are interested in new results in clustering.

## Organization

This book is organized as follows. The book is divided into six chapters. In Chapter 1, a deep introduction is given about clustering, emphasizing the methods and algorithms that are used in the remainder of the book. For the sake of completeness, a brief overview about other methods is also presented. This chapter gives a detailed description about fuzzy clustering with examples to illustrate the difference between them.

Chapter 2 is in direct connection with clustering: visualization of clustering results is dealt with. The presented methods enable the user to see the  $n$ -dimensional clusters, therefore to validate the results. The remainder chapters are in connection with different data mining fields, and the common is that the presented methods utilize the results of clustering.

Chapter 3 deals with fuzzy model identification and presents methods to solve them. Additional familiarity in regression and modelling is helpful but not required because there will be an overview about the basics of fuzzy modelling in the introduction.



Chapter 4 deals with identification of dynamical systems. Methods are presented with their help multiple input – multiple output systems can be modeled, *a priori* information can be built in the model to increase the flexibility and robustness, and the order of input-output models can be determined.

In Chapter 5, methods are presented that are able to use the label of data, therefore the basically unsupervised clustering will be able to solve classification problems. By the fuzzy models as well as classification methods transparency and interpretability are important points of view.

In Chapter 6, a method related to time-series analysis is given. The presented method is able to discover homogeneous segments in multivariate time-series, where the bounds of the segments are given by the change in the relationship between the variables.

## Features

The book is abundantly illustrated by

- Figures (120);
- References (302) which give a good overview of the current state of fuzzy clustering and data mining topics concerned in this book;
- Examples (39) which contain simple synthetic data sets and also real-life case studies.

During writing this book, the authors developed a toolbox for MATLAB® called Clustering and Data Analysis Toolbox that can be downloaded from the File Exchange Web site of MathWorks. It can be used easily also by (post)graduate students and for educational purposes as well. This toolbox does not contain all of the programs used in this book, but most of them are available with the related publications (papers and transparencies) at the Web site: [www.fmt.vein.hu/softcomp](http://www.fmt.vein.hu/softcomp).

## Acknowledgements

Many people have aided the production of this project and the authors are greatly indebted to all. These are several individuals and organizations whose support demands special mention and they are listed in the following.

The authors are grateful to the Process Engineering Department at the University of Veszprem, Hungary, where they have worked during the past years. In particular, we are indebted to *Prof. Ferenc Szeifert*, the former Head of the Department, for providing us the intellectual freedom and a stimulating and friendly working environment.

Balazs Feil is extremely grateful to his parents, sister and brother for their continuous financial, but most of all, mental and intellectual support. He is also

indebted to all of his roommates during the past years he could (almost) always share his problems with.

Parts of this book are based on papers co-authored by *Dr. Peter Arva*, *Prof. Robert Babuska*, *Sandor Migaly*, *Dr. Sandor Nemeth*, *Peter Ferenc Pach*, *Dr. Hans Roubos*, and *Prof. Ferenc Szeifert*. We would like to thank them for their help and interesting discussions.

The financial support of the Hungarian Ministry of Culture and Education (FKFP-0073/2001) and Hungarian Research Funds (T049534) and the Janos Bolyai Research Fellowship of the Hungarian Academy of Sciences is gratefully acknowledged.



<http://www.springer.com/978-3-7643-7987-2>

Cluster Analysis for Data Mining and System  
Identification

Abonyi, J.; Feil, B.

2007, XVIII, 306 p., Hardcover

ISBN: 978-3-7643-7987-2

A product of Birkhäuser Basel