

# Chapter 9

## Robust Dimensionality Reduction: A Resistant Search for the Relevant Information in Complex Data

**Jan Kalina**

 <https://orcid.org/0000-0002-8491-0364>

*Institute of Computer Science, The Czech Academy of Sciences, Czech Republic*

### **ABSTRACT**

*With the increasing availability of massive data in various fields of applications such as engineering, economics, or biomedicine, there appears an urgent need for new reliable tools for obtaining relevant knowledge from such data, which allow one to find and interpret the most relevant features (variables). Such interpretation is however infeasible for the habitually used methods of machine learning, which can be characterized as black boxes. This chapter is devoted to variable selection methods for finding the most relevant variables for the given task. After explaining general principles, attention is paid to robust approaches, which are suitable for data contaminated by outlying values (outliers). Three main approaches to variable selection (prior, intrinsic, and posterior) are explained, and their recently proposed examples are illustrated on applications related to credit risk management and molecular genetics. These examples reveal recent robust approaches to data analysis to be able to outperform non-robust tools.*

DOI: 10.4018/978-1-6684-5264-6.ch009

## **INTRODUCTION**

As the current society is without any doubt oversupplied with data (information), it belongs to crucial tasks of informatics to contribute to a transform of the vast amount information to a small amount of relevant knowledge. In other words, the fight against redundant (useless) information belongs to key tasks of informatics, or the whole current science in general. Only with the help of tools for transforming information to knowledge, the society can expect its shift towards the ideas of the knowledge society (Tegmark, 2017). Artificial intelligence (AI) tools can be expected to contribute to such complexity reduction of the omnipresent information. Tools of computational intelligence (CI), which represents a subset of artificial intelligence, can be especially helpful in this respect. When computational intelligence needs to obtain practically useful knowledge from available information while accounting for uncertainty, machine learning with its statistical algorithms comes into play.

A plethora of innovative tools is nowadays available for obtaining relevant knowledge from (possibly big) data in a variety of tasks. To give only a single application, a number of promising artificial intelligence tools has been engaged in the fight against the COVID-19 pandemic (Lalmuanawma et al., 2020). The role of scientific computations has acquired increasing attention of practitioners as well as among statisticians (Quarteroni, 2018) and the quickly growing field of scientific computations, exploiting advanced computing for analyzing scientific problems, has been denoted as computational science (Holder & Eichholz, 2019). The key pillars of computational intelligence are generally acknowledged to include neural networks, fuzzy logic methods, or evolutionary computation algorithms; still, probabilistic methods allowing to evaluate results under randomness (uncertainty) have their irreplaceable role within computational intelligence as well.

While habitually used methods of machine learning applicable within scientific computations can be characterized as black boxes, practical applications often require to understand why a particular conclusion (e.g. decision) was made, or which are the most relevant variables contributing to explaining a given response variable. If the methods allow such clear interpretation, we speak about explainable artificial intelligence or explainable machine learning. Naturally, understanding limitations of artificial intelligence belongs to ethical issues and the impossibility to explain rigorously why given algorithms yield particular results represents an important ethical issue as such (Jacobson et al., 2020). Two approaches (or in fact aims) for improving the explainability of machine learning tools, which may be used at the same time, are dimensionality reduction and robustness to outlying values (outliers), where the latter allows to reduce the influence of individual outliers and to evaluate (quantify) the influence of individual observations.

23 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: [www.igi-global.com/chapter/robust-dimensionality-reduction/314342](http://www.igi-global.com/chapter/robust-dimensionality-reduction/314342)

## Related Content

---

### Embedded System for Heart Disease Recognition using Fuzzy Clustering and Correlation

Helton Hugo de Carvalho Júnior, Robson Luiz Moreno and Tales Cleber Pimenta (2012). *Multidisciplinary Computational Intelligence Techniques: Applications in Business, Engineering, and Medicine* (pp. 327-350).

[www.irma-international.org/chapter/embedded-system-heart-disease-recognition/67300](http://www.irma-international.org/chapter/embedded-system-heart-disease-recognition/67300)

### Enhanced Global Best Particle Swarm Classification

Nabila Nouaouria, Mounir Boukadoumand Robert Proulx (2014). *International Journal of Software Science and Computational Intelligence* (pp. 1-15).

[www.irma-international.org/article/enhanced-global-best-particle-swarm-classification/127350](http://www.irma-international.org/article/enhanced-global-best-particle-swarm-classification/127350)

### Cooperative Encoding Strategy for Gate Array Placement

Hong-Bo Wang, Qing-Dong Su and Ruolei Zeng (2018). *International Journal of Software Science and Computational Intelligence* (pp. 29-43).

[www.irma-international.org/article/cooperative-encoding-strategy-for-gate-array-placement/223493](http://www.irma-international.org/article/cooperative-encoding-strategy-for-gate-array-placement/223493)

### Recent Survey on Medical Image Segmentation

Mohammed A.-M. Salem, Alaa Atef, Alaa Salah and Marwa Shams (2017). *Handbook of Research on Machine Learning Innovations and Trends* (pp. 424-464).

[www.irma-international.org/chapter/recent-survey-on-medical-image-segmentation/180956](http://www.irma-international.org/chapter/recent-survey-on-medical-image-segmentation/180956)

### Machine Learning Applications in Breast Cancer Diagnosis

Syed Jamal Safdar Gardezi, Mohamed Meselhy Eltoukhy and Ibrahima Faye (2017). *Handbook of Research on Machine Learning Innovations and Trends* (pp. 465-490).

[www.irma-international.org/chapter/machine-learning-applications-in-breast-cancer-diagnosis/180957](http://www.irma-international.org/chapter/machine-learning-applications-in-breast-cancer-diagnosis/180957)