Chapter 1
Feature Selection for Data and Pattern Recognition: An Introduction

Urszula Stańczyk and Lakhmi C. Jain

Abstract Surrounded by data and information in various forms we need to characterise and describe objects of our universe using some attributes of nominal or numerical type. Selection of features can be performed basing on domain knowledge, executed through dedicated approaches, driven by some particular inherent properties of methodologies and techniques employed, or governed by other factors or rules. This chapter presents a general and brief introduction to topics of feature selection for data and pattern recognition. Its main aim is to provide short descriptions of the chapters included in this volume.

Keywords Feature · Feature selection · Pattern recognition · Data mining

1.1 Introduction

Some say that our earliest memories form when, as children, we learn to describe the world we live in, and express verbally what we feel and think, how we perceive other people, objects, events, abstract concepts. While we grow older, we learn to detect and recognise patterns [20], and our discriminating skills grow as well. We develop associations, preferences and dislikes, which are employed, consciously and subconsciously, when choices are made, actions taken.

Imagine opening an unknown thick book and finding in it a whole page dedicated to a line of thought of some character, jumping from one topic to another, along with connecting ideas, feelings and memories, digressions. Without looking at the cover or the title page, by similarity to a stream of consciousness, one instantly thinks

U. Stańczyk (✉)
Institute of Informatics, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland
e-mail: urszula.stanczyk@polsl.pl

L.C. Jain
Faculty of Education, Science, Technology and Mathematics,
University of Canberra, Canberra, ACT 2601, Australia
e-mail: lakhmi.jain@unisa.edu.au

U. Stańczyk and L.C. Jain (eds.), *Feature Selection for Data and Pattern Recognition*, Studies in Computational Intelligence 584,
DOI 10.1007/978-3-662-45620-0_1

about James Joyce as the author. A painting with a group of posed ballet dancers upon a stage we would associate with Degas, and water lilies in a pond with Monet. Hearing rich classical organ music we could try to guess Bach as the composer. In each of these exemplary cases we have a chance of correct recognition basing on some characteristic features the authors are famous for. Our brains recognise lily flowers or organ tunes, yet to make other people or machines capable of the same we need to explain these specific elements, which means describing, expressing them in understandable and precise terms.

Characterisation of things is a natural element of life, some excel at it while others are not so good. Yet anybody can make basic distinctions, especially with some support system. Some of how these characteristics play into problems we need to tackle, tasks waiting to be solved, comes intuitively, some we get from observations or experiments, drawn conclusions. Some pointers are rather straightforward while others indirect or convoluted.

According to a dictionary definition a *feature* is a distinctive attribute or aspect of something and it is used as a synonym for characteristic, quality, or property [29, 38]. With such meaning it is employed in general language descriptions but also in more confined areas of technical sciences, computer technologies, in particular in the domain of data mining and pattern recognition [24, 30, 39].

For automatic recognition and classification [11, 27] all objects of the universe of discourse need to be perceived through information carried by their characteristics and in cases when this information is incomplete or uncertain the resulting predictive accuracies of constructed systems, whether they induce knowledge from available data in supervised or unsupervised manner [28], relying on statistics-oriented calculations [8, 19] or heuristic algorithms, could be unsatisfactory or falsified, making observations and conclusions unreliable.

The performance of any inducer depends on the raw input data on which inferred knowledge is based [21], exploited attributes, the approach or methodology of data mining applied, but also on the general dimensionality of the problem [40]. Contemporary computer technologies with their high computational capabilities aid in processing, but still for huge data sets, and very high numbers of variables the process, even if feasible, can take a lot of time and effort, require unnecessary or impractically large storage.

Typically the primary goal is to achieve the maximal classification accuracy but we need to take into account practical aspects of obtained solutions, and consider compromises with trade-offs such as some loss in performance for much shortened time, less processing, lower complexity, or smaller structure of the system.

Feature selection is an explicit part of most knowledge mining approaches—some attributes are chosen over others while forming a set of characteristic features in the first place [10, 18]. Here the choice can be supported by expert knowledge. Once some subset of variables is available, using it to construct a rule classifier, a rule induction algorithm leads to particular choices of conditions for all constituent rules, either usual or inhibitory. In a similar manner in a decision tree construction specific attributes are to be checked at its nodes, and artificial neural networks through their

learning rule establish the degrees of importance or relevance of features. Such examples can be multiplied.

Even for working solutions it is worthwhile to study attributes as it is not out of realm of possibility that some of them are excessive or repetitive, even irrelevant, or there exist other alternatives of the same merit, and once such variables are discovered, different selection can improve the performance, if not with respect to the classification accuracy, then by better understanding of analysed concepts, possibly more explicit presentation of information [23].

With all these factors and avenues to explore it is not surprising that the problem of feature selection, with various meanings of this expression, is actively pursued in research, which has given us the motivation for dedicating this book to this area.

1.2 Chapters of the Book

The 13 chapters included in this volume are grouped into four parts. What follows is a short description of the content for each chapter.

Part I Estimation of Feature Importance

Chapter 2 is devoted to a review of the field of all-relevant feature selection, and presentation of the representative algorithm [5, 25]. The problem of all-relevant feature selection is first defined, then key algorithms are described. Finally the Boruta algorithm is explained in a greater detail and applied both to a collection of synthetic and real-world data sets, with comments on performance, properties and parameters.

Chapter 3 illustrates the three approaches to feature selection and reduction [17]: filters, wrappers, and embedded solutions [25], combined for the purpose of feature evaluation. These approaches are used when domain knowledge is unavailable or insufficient for an informed choice, or in order to support this expert knowledge to achieve higher efficiency, enhanced classification, or reduced sizes of classifiers. The classification task under study is that of authorship attribution with balanced data.

Chapter 4 presents a method of feature ranking that calculates the relative weight of features in their original domain with an algorithmic procedure [3]. The method supports information selection of real world features and is useful when the number of features has costs implications. It has at its core a feature extraction technique based on effective decision boundary feature matrix, which is extended to calculate the total weight of the real features through a procedure geometrically justified [28].

Chapter 5 focuses on weighting of characteristic features by the processes of their sequential selection. A set of all accessible attributes can be reduced backwards, or variables examined one by one can be selected forward. The choice can be conditioned by the performance of a classification system, in a wrapper model, and the observations with respect to selected variables can result in assignment

of weights. The procedures are employed for rule [37] and connectionist [26] classifiers, applied in the task of authorship attribution.

Part II Rough Set Approach to Attribute Reduction

Chapter 6 discusses two probabilistic approaches [44] to rough sets: the variable precision rough set model [43] and the Bayesian rough set model, as they apply to data dependencies detection, analysis and their representation. The focus is on the analysis of data co-occurrence-based dependencies appearing in classification tables and probabilistic decision tables acquired from data. In particular, the notion of attribute reduct, in the framework of probabilistic approach, is of interest in the chapter and it includes two efficient reduct computation algorithms.
Chapter 7 provides an introduction to a rough set approach to attribute reduction [1], treated as removing condition attributes with preserving some part of the lower/upper approximations of the decision classes, because the approximations summarize the classification ability of the condition attributes [42]. Several types of reducts according to structures of the approximations are presented, called "structure-based" reducts. Definitions and theoretical results for structures-based attribute reduction are given [33, 36].

Part III Rule Discovery and Evaluation

Chapter 8 compares a strategy of rule induction based on feature selection [32], exemplified by the LEM1 algorithm, with another strategy, not using feature selection, exemplified by the LEM2 algorithm [15, 16]. The LEM2 algorithm uses all possible attribute-value pairs as the search space. It is shown that LEM2 significantly outperforms LEM1, a strategy based on feature selection in terms of an error rate. The LEM2 algorithm induces smaller rule sets with the smaller total number of conditions as well. The time complexity for both algorithms is the same [31].
Chapter 9 addresses action rules extraction. Action rules present users with a set of actionable tasks to follow to achieve a desired result. The rules are evaluated using their supporting patterns occurrence and their confidence [41]. These measures fail to measure the feature values transition correlation and applicability, hence meta-actions are used in evaluating action rules, which is presented in terms of likelihood and execution confidence [14]. Also an evaluation model of the application of meta-actions based on cost and satisfaction is given.
Chapter 10 explores the use of a feature subset selection measure, along with a number of common statistical interestingness measures, via structure-preserving flat representation for tree-structured data [34, 35]. A feature subset selection is used prior to association rule generation. Once the initial set of rules is obtained, irrelevant rules are determined as those that are comprised of attributes not determined to be statistically significant for the classification task [22].

Part IV Data- and Domain-Oriented Methodologies

Chapter 11 gives a survey of hubness-aware classification methods and instance selection. The presence of hubs, the instances similar to exceptionally large number

of other instances, has been shown to be one of the crucial properties of time-series data sets [4, 7]. There are proposed some selected instances for feature construction, detailed description of the algorithms provided, and experimental results on large number of publicly available real-world time-series data sets shown.
Chapter 12 presents an analysis of descriptors that utilize various aspects of image data: colour, texture, gradient, and statistical moments, and this list is extended with local features [2]. The goal of the analysis is to find descriptors that are best suited for particular task, i.e. re-identification of objects in a multi-camera environment. For descriptor evaluation, scatter and clustering measures [12] are supplemented with a new measure derived from calculating direct dissimilarities between pairs of images [5, 6].
Chapter 13 deals with the selection of the most appropriate moment features used to recognise known patterns [13]. For this purpose, some popular moment families are presented and their properties are discussed. Two algorithms, a simple Genetic Algorithm (GA) and the Relief algorithm are applied to select the moment features that better discriminate human faces and facial expressions, under several pose and illumination conditions [9].
Chapter 14 contains considerations on grouped features. When features are grouped, it is desirable to perform feature selection groupwise in addition to selecting individual features. It is typically the case in data obtained by modern high-throughput genomic profiling technologies such as exon microarrays. To handle grouped features, feature selection methods are discussed with the focus on a popular shrinkage method, lasso, and its variants, that are based on regularized regression with generalized linear models [6].

1.3 Concluding Remarks

In this book some advances and research dedicated to feature selection for data and pattern recognition are presented. Even though it has been the subject of interest for some time, feature selection remains one of actively pursued avenues of investigations due to its importance and bearing upon other problems and tasks. It can be studied within a domain from which features are extracted, independently of it, taking into account specific properties of involved algorithms and techniques, with feedback from applications, or without it. Observations from executed experiments can bring local and global conclusions, with theoretical and practical significance.

References

1. Abraham, A., Falcón, R., Bello, R. (eds.): Rough Set Theory: A True Landmark in Data Analysis. Studies in Computational Intelligence, vol. 174. Springer, Berlin (2009)
2. Baxes, G.A.: Digital Image Processing: Principles and Applications. Wiley, New York (1994)

3. Blum, A., Langley, P.: Selection of relevant features and examples in machine learning. Artif. Intell. **97**, 245–271 (1997)
4. Botsch, M.: Machine Learning Techniques for Time Series Classification. Cuvillier, San Francisco (2009)
5. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
6. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth & Brooks, Monterey (1984)
7. Carbonell, J.G. (ed.): Machine Learning. Paradigms and Methods. MIT Press, Boston (1990)
8. Chao, L.L.: Introduction to Statistics. Brooks Cole Publishing Co., Monterey (1980)
9. Cipolla, R., Pentland, A.: Computer Vision for Human-Machine Interaction. Cambridge University Press, Cambridge (1998)
10. Dash, M., Liu, H.: Feature selection for classification. Intell. Data Anal. **1**, 131–156 (1997)
11. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, New York (1973)
12. Everitt, B.: Cluster Analysis. Heinemann Educational Books, London (1980)
13. Flusser, J., Zitova, B., Suk, T.: Moments and Moment Invariants in Pattern Recognition. Wiley, New York (2009)
14. Fuernkranz, J., Gamberger, D., Lavrac, N.: Foundations of Rule Learning. Springer, Berlin (2012)
15. Grzymala-Busse, J.W.: Knowledge acquisition under uncertainty—a rough set approach. J. Intell. Robot. Syst. **1**, 3–16 (1988)
16. Grzymala-Busse, J.W.: Data with missing attribute values: generalization of indiscernibility relation and rule induction. Trans. Rough Sets **1**, 78–95 (2004)
17. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
18. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.: Feature Extraction. Foundations and Applications. Springer, Berlin (2006)
19. Hamburg, M.: Statistical Analysis for Decision Making. Harcourt Brace Jovanovich Inc., New York (1983)
20. Hofstadter, D.: Metamagical Themas: Questing for the Essence of Mind and Pattern. Basic Books Inc., New York (1985)
21. Holland, J.H., Holyoak, K.J., Nisbett, R.E.: Induction: Processes of Inference, Learning, and Discovery. MIT Press, Boston (1986)
22. Hollander, M., Wolfe, D.A.: Nonparametric Statistical Methods, 2nd edn. Wiley, New York (1999)
23. Jensen, R., Shen, Q.: Computational Intelligence and Feature Selection. Wiley, Hoboken (2008)
24. Kloesgen, W., Zytkow, J. (eds.): Handbook of Data Mining and Knowledge Discovery. Oxford University Press, New York (2002)
25. Kohavi, R., John, G.: Wrappers for feature selection. Artif. Intell. **97**, 273–324 (1997)
26. Krawiec, K., Stefanowski, J.: Machine Learning and Neural Nets. Poznan University of Technology Press, Poznan (2003)
27. Kuncheva, L.I.: Combining Pattern Classifiers. Methods and Algorithms. Wiley, Hoboken (2004)
28. Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman and Hall/ CRC, Boca Raton (2007)
29. Longman Dictionary of Contemporary English, 6th revised edn. Pearson Longman, London (2014)
30. Maimon, O., Rokach, L. (eds.): Data Mining and Knowledge Discovery Handbook, 2nd edn. Springer, Berlin (2010)
31. Meyers, R.A. (ed.): Encyclopedia of Complexity and Systems Science. Springer, Berlin (2009)
32. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Boston (1991)
33. Polkowski, L., Skowron, A. (eds.): Rough Sets in Knowledge Discovery 1: Methodology and Applications. Physica-Verlag, Heidelberg (1998)

34. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)
35. Roiger, R.J., Geatz, M.W.: Data Mining. A Tutorial-Based Primer. Addison-Wesley, Boston (2003)
36. Slowinski, R.: Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, Boston (1992)
37. Stefanowski, J.: Algorithms of Decision Rule Induction in Data Mining. Poznan University of Technology Press, Poznan (2001)
38. The Merriam-Webster Dictionary. http://www.merriam-webster.com/
39. Wang, J. (ed.): Data Mining: Opportunities and Challenges. Idea Group Publishing, Hershey (2003)
40. Weiss, S.M., Indurkhya, N.: Predictive Data Mining. A Practical Guide. Morgan Kaufmann Publication, San Francisco (1998)
41. Witten, I., Frank, E.: Data Mining. Practical Machine Learning Tools and Techniques. Elsevier, Amsterdam (2005)
42. Zanakis, H., Doukidis, G., Zopounidised, Z. (eds.): Decision Making: Recent Developments and Worldwide Applications. Kluwer Academic Publishers, Boston (2000)
43. Ziarko, W.: Variable precision rough set model. J. Comput. Syst. Sci. **46**(1), 39–59 (1993)
44. Ziarko, W.: Probabilistic approach to rough sets. Int. J. Approx. Reason. **49**, 272–284 (2008)