
FEATURE SELECTION FOR KNOWLEDGE DISCOVERY AND DATA MINING

**THE KLUWER INTERNATIONAL SERIES
IN ENGINEERING AND COMPUTER SCIENCE**

Library of Congress Cataloging-in-Publication Data

Liu, Huan.

Feature selection for knowledge discovery and data mining / by
Huan Liu and Hiroshi Motoda.

p. cm. -- (Kluwer international series in engineering and
computer science ; 454)

Includes bibliographical references and index.

ISBN 978-1-4613-7604-0 ISBN 978-1-4615-5689-3 (eBook)

DOI 10.1007/978-1-4615-5689-3

1. Database management. 2. Data mining. I. Motoda, Hiroshi.

II. Title. III. Series : Kluwer international series in engineering
and computer science ; SECS 454.

QA76.9.D3L595 1998

006.3--dc21

98-25204

CIP

Copyright © 1998 by Springer Science+Business Media New York

Originally published by Kluwer Academic Publishers, New York in 1998. Second
Printing 2000.

Softcover reprint of the hardcover 1st edition 1998

All rights reserved. No part of this publication may be reproduced, stored in a
retrieval system or transmitted in any form or by any means, mechanical, photo-
copying, recording, or otherwise, without the prior written permission of the publisher,
Springer Science+Business Media, LLC.

Printed on acid-free paper.

FEATURE SELECTION FOR KNOWLEDGE DISCOVERY AND DATA MINING

by

Huan Liu

*National University of Singapore
SINGAPORE*

and

Hiroshi Motoda

*Osaka University
Osaka, JAPAN*



SPRINGER SCIENCE+BUSINESS MEDIA, LLC

Contents

List of Figures	xi
List of Tables	xv
Preface	xix
Acknowledgments	xxiii
1. DATA PROCESSING AND KDD	1
1.1 Inductive Learning from Observation	2
1.1.1 Features	2
1.1.2 Feature-based classification	5
1.2 Knowledge Discovery and Data Mining	6
1.2.1 Data mining - a new challenge	7
1.3 Feature Selection and Its Roles in KDD	9
1.4 Summary	12
References	13
2. PERSPECTIVES OF FEATURE SELECTION	17
2.1 Feature Selection for Classification	18
2.2 A Search Problem	20
2.2.1 Search directions	20
2.2.2 Search strategies	22
2.3 Selection Criteria	24
2.3.1 The ideal classifier	25
2.3.2 Information measures	26
2.3.3 Distance measures	27
2.3.4 Dependence measures	27
2.3.5 Consistency measures	28
2.3.6 Accuracy measures	28
2.3.7 Discussion	28

2.4	Univariate vs. Multivariate Feature Selection	30
2.4.1	A statistical approach	30
2.4.2	A decision tree approach	31
2.4.3	A neural network approach	32
2.4.4	Summary	33
2.5	Filter vs. Wrapper Models	33
2.5.1	A machine learning approach	34
2.5.2	A data mining approach	35
2.6	A Unified View	36
2.7	Conclusion	38
	References	38
3.	ASPECTS OF FEATURE SELECTION	43
3.1	Overview	43
3.2	Basic Feature Generation Schemes	46
3.2.1	Sequential forward generation	46
3.2.2	Sequential backward generation	48
3.2.3	Bidirectional generation	49
3.2.4	Random generation	49
3.3	Search Strategies	50
3.3.1	Complete search	50
3.3.2	Heuristic search	55
3.3.3	Nondeterministic search	59
3.4	Evaluation Measures With Examples	62
3.4.1	Classic measures	62
3.4.2	Consistency measures	66
3.4.3	Accuracy measures	68
3.5	Conclusion	69
	References	70
4.	FEATURE SELECTION METHODS	73
4.1	Representative Feature Selection Algorithms	74
4.1.1	Complete methods	75
4.1.2	Heuristic methods	78
4.1.3	Stochastic methods	80
4.1.4	Feature weighting methods	84
4.1.5	Hybrid methods	85
4.1.6	Incremental approach	87
4.2	Employing Feature Selection Methods	89
4.3	Conclusion	90
	References	90
5.	EVALUATION AND APPLICATION	97

5.1	Performance Assessment	97
5.2	Evaluation Methods for Classification	100
5.2.1	Basic statistics	100
5.2.2	Error rate and how to measure it	106
5.2.3	Commonly used evaluation methods	110
5.3	Evaluation of Selected Features	111
5.3.1	Direct evaluation of features	112
5.3.2	Indirect evaluation of features	112
5.3.3	Which type of evaluation should we use?	112
5.4	Evaluation: Some Examples	113
5.4.1	Experiment purposes and design	113
5.4.2	Experiments and results	117
5.4.3	Issue of scalability	130
5.4.4	Remarks	131
5.5	Balance between Different Performance Criteria	131
5.6	Applying Feature Selection Methods	137
5.6.1	Prior knowledge	137
5.6.2	Processing speed	138
5.6.3	Data characteristics	139
5.6.4	How to choose a feature selection method?	140
5.7	Conclusions	145
	References	146
6.	FEATURE TRANSFORMATION AND DIMENSIONALITY REDUCTION	151
6.1	Feature Extraction	152
6.2	Feature Construction	157
6.3	Feature Discretization	163
6.4	Beyond the Classification Model	170
6.4.1	Unsupervised feature selection: clustering	170
6.4.2	Unsupervised feature selection: entropy	177
6.5	Conclusions	182
	References	183
7.	LESS IS MORE	189
7.1	A Look Back	190
7.2	A Glance Ahead	191
	References	194
	Appendices	196
A-	Data Mining and Knowledge Discovery Sources	197
A.1	Web Site Links	197
A.2	Electronic Newsletters, Pages and Journals	200

x	FEATURE SELECTION FOR KDD	
	A.3 Some Publically Available Tools	202
B-	Data Sets and Software Used in This Book	205
	B.1 Data Sets	205
	B.2 Software	206
	References	207
Index		211

List of Figures

1.1	The hierarchy of feature types.	3
1.2	A general model of knowledge discovery and data mining (KDD).	7
2.1	Feature selection as a search problem: a lattice and three features.	20
2.2	The relations between the five types of measures.	29
2.3	A simple neural network: Perceptron.	32
2.4	A wrapper model of feature selection.	34
2.5	A filter model of feature selection.	36
2.6	A unified model of feature selection.	37
3.1	Three principal dimensions of feature selection: search strategy, evaluation measure, and feature generation scheme.	44
3.2a	Depth-first search illustration with three features a, b, c .	51
3.2b	Breadth-first search illustration with three features a, b, c .	51
3.3	Branch & Bound search illustration with three features a, b, c . Numbers beside nodes are their costs.	54
3.4	Best-first search illustration with three features a, b, c .	57
3.5	Beam search illustration with three features $a, b, c, \eta = 2$.	58
3.6	Approximate Branch & Bound search illustration with three features a, b, c .	60
3.7	Information gain calculation example: Data D is partitioned by feature X into data subsets $D_i, i = 1, 2, \dots, p$.	64
4.1	A simple example with four features. The first two are relevant.	78
4.2	A typical trend of the decreasing number of valid features versus the number of runs performed. The Mushroom data is used in the experiment.	86
5.1	A frequency histogram for feature Hair of the sunburned data.	101

- 5.2 A learning curve: observing the change of accuracy with increasing instances. One chunk of data is roughly equal to N/m instances, and $m = 11$. 110
- 5.3 (a) Error rates of C4.5 on the Parity5+5 data using the features ordered by NBC and C4.5. (b) Error rates of NBC on the Parity5+5 data using the features ordered by NBC and C4.5. 124
- 5.4 (a) Features of the Corral data are ordered by WSFG using C4.5; (b) Features are ordered by WSBG using C4.5. Test results (two curves) are obtained by C4.5 and NBC respectively. 125
- 5.5 The effect of feature selection on the data size required for the learning to converge. The learning curves with and without feature selection show average error rates of 10-fold cross validation for C4.5 on the Parity5+5 data with various chunks of data. 126
- 5.6 (a) Error rates of C4.5 on the Iris data using features ordered (sequential forward) by C4.5; (b) Error rates of C4.5 on Iris using features ordered (sequential backward) by C4.5. 129
- 5.7 Average CPU time (seconds). The result for 100% data is used as the reference. The difference between any two samples (e.g., 10% vs. 100%, or 20% vs. 100%) is the most significant (light grey), significant (dark grey), or insignificant (black). 132
- 5.8 Average number of selected features. The result for 100% data is used as the reference. The difference between any two samples (e.g., 10% vs. 100%, or 20% vs. 100%) is most significant (light grey), significant (dark grey), or insignificant (black). 133
- 5.9 Comparing end results: with and without feature selection. 134
- 5.10 Choosing a proper feature selection method based on knowledge and time available. 141
- 5.11 An influence map shows how many factors have influence over evaluation measures, search strategies, and selection methods. 144
- 5.12 Choosing a method: bold italic words are the major factors with their values in normal fonts; search strategy is also in bold italic; feature selection methods are in bold type. 145
- 6.1 The Iris data is represented by two features extracted by Principal Component Analysis. The original data is 4-dimensional. 154
- 6.2 The Iris data is represented by two features extracted from a neural network. The original data is of 4-dimensional. 158
- 6.3 The Iris data is represented by two original features: Petal Length and Petal Width. 159

6.4a	Illustration of the replication problem in a decision tree building for target concept $x_1x_2 + \bar{x}_3x_4$.	161
6.4b	The removal of replication in a decision tree building using constructed features x_1x_2 and $\bar{x}_3x_4$.	161
6.5	The effect of discretization in decision tree building.	163
6.6a	The Iris data in 3-dimensional and 2-dimensional spaces.	178
6.6b	The Iris data in 3-dimensional and 2-dimensional spaces.	179
6.7	The effect on error rates: including features from the most important to the least important x_3, x_4, x_1, x_2 .	182
7.1	A data flow diagram: relations between data, feature manipulation techniques and knowledge discovery.	192

List of Tables

1.1	An example of feature-based data.	4
2.1	A 2-d table about search strategies and search directions.	24
3.1	A univariate feature ranking algorithm.	45
3.2	A minimum subset algorithm. $\#$ - a cardinality operator returns the number of features in a set.	46
3.3	Sequential forward feature set generation - SFG .	47
3.4	Sequential backward feature generation - SBG .	48
3.5	Bidirectional feature set generation - FBG .	49
3.5a	Exhaustive search: depth first - DEP with explicit use of a stack.	52
3.5b	Exhaustive search: depth first - DEP without explicit use of a stack.	53
3.6	Exhaustive search: breadth first - BRD .	53
3.7	Complete search: Branch & Bound - BAB .	55
3.8	Heuristic search: best-first - BEST .	56
3.9	Heuristic search: beam search - BEAM .	58
3.10	Heuristic search: approximate branch & bound - ABAB .	59
3.11	Random search - RAND .	61
3.12	An example of feature-based data - revisited.	63
3.13	Priors and class conditional probabilities for the sunburn data.	64
3.14	Ordering features according to their information gains.	65
3.15	Ordering features according to accuracy.	69
4.1	A 3-d structure for feature selection algorithms categorization.	75
4.2	Focus - an exhaustive forward search algorithm.	76
4.3	ABB - an automated branch and bound (backward search) algorithm.	77
4.4	A heuristic feature selection algorithm - inducing a classifier.	79

4.5	A stochastic filter feature selection algorithm - LVF.	82
4.6	A stochastic wrapper feature selection algorithm - LVW.	83
4.7	A multivariate feature ranking algorithm.	85
4.8	A hybrid algorithm of feature selection.	87
4.9	LVI - an incremental multivariate feature selection algorithm.	89
5.1	A hypothesis testing procedure about the sample mean.	103
5.2	A confusion matrix for three classes (0, 1, and 2).	107
5.3	A cost matrix for various types of errors.	108
5.4	Summary of data sets used in evaluation of feature selection methods. Notations: C - the number of distinct classes, N - the number of features, S - the size of the data set, S_d - the size of the training data, S_t - the size of the test data. Training and test data sets are split randomly if not specified.	116
5.5	A summary of feature selection methods used in experimentation. Comp - Complete, Heur - Heuristic, Wrap - Wrapper, Nond - Nondeterministic, Hybr - Hybrid, Incr - Incremental, Fwd - Forward, Bwd - Backward, Rnd - Random.	118
5.6	Selected features using training data. K.R.F. means known relevant features. $\bar{d}$ means feature d is redundant.	119
5.7	Ordered features using training data. K.R.F. means known relevant features. $\bar{d}$ means feature d is redundant.	120
5.8	Average error rates (E-Rate) and average tree size (Tr-Size) of C4.5 Before and After feature selection and their p -values. The known relevant features are used as selected ones.	122
5.9	Average error rates (E-Rate) and average table size (Ta-Size) of NBC Before and After feature selection and their p -values. The known relevant features are used as selected ones.	122
5.10	A run time comparison between LVF and LVW.	127
5.11	Error rates of three classifiers on Corral and Parity5+5 data before and after feature selection. Features are selected by Focus.	128
5.12	Feature selection methods and their capabilities of handling data of different characteristics. Comp - Complete, Heur - Heuristic, Nond - Nondeterministic; Cont/Disc - Continuous or Discrete, Rednt - Redundant, RndCorr - Randomly Correlated Noise; Dpnd means it depends.	142
5.13	Output types of some feature selection methods and their models (filter or wrapper).	144
6.1	ChiMerge: a χ^2 statistic based discretization algorithm.	166
6.2	Chi2: a χ^2 statistic based, aggressive discretization algorithm.	168
6.3	Nearest Neighbor algorithm (NearN).	171

6.4	An agglomerative algorithm.	173
6.5	An algorithm for concept formation (ConForm).	175

Preface

As computer power grows and data collection technologies advance, a plethora of data is generated in almost every field where computers are used. The computer generated data should be analyzed by computers; without the aid of computing technologies, it is certain that huge amounts of data collected will not ever be examined, let alone be used to our advantages. Even with today's advanced computer technologies (e.g., machine learning and data mining systems), discovering knowledge from data can still be fiendishly hard due to the characteristics of the computer generated data. Taking its simplest form, raw data are represented in feature-values. The size of a dataset can be measured in two dimensions, number of features (N) and number of instances (P). Both N and P can be enormously large. This enormity may cause serious problems to many data mining systems.

Feature selection is one of the long existing methods that deal with these problems. Its objective is to select a minimal subset of features according to some reasonable criteria so that the original task can be achieved equally well, if not better. By choosing a minimal subset of features, irrelevant and redundant features are removed according to the criterion. When N is reduced, the data space shrinks and in a sense, the data set is now a better representative of the whole data population. If necessary, the reduction of N can also give rise to the reduction of P by eliminating duplicates. Simpler data can lead to more concise results and their better comprehensibility. Because of these advantages, feature selection has been the focus of interest for quite some time. Much work has been done from 70's to the present. With the creation of huge databases and the consequent requirements for good data mining programs, new problems arise and novel approaches to feature selection are in high demand. This is a perfect time to look back and see what have been done, and to look forward to the challenges ahead.

This book offers an overview of the various methods developed since 70's, provides a general framework in order to examine many methods and categorize them, employs simple examples to show the essence of representative feature selection methods, compares them using data sets with combinations of intrinsic properties according to the objective of feature selection, suggests guidelines how to use different methods under various circumstances, and points out some new challenges. This book consists of seven chapters, two appendices, and bibliographies after each chapter. Chapter 1 is about the background knowledge on data explosion, knowledge discovery from raw data, machine learning, and feature selection. Many fields related to these topics will be mentioned such as pattern recognition, statistics, visualization, database management systems, and so on. Chapter 2 describes perspectives of feature selection and explains how these different perspectives can be unified. Representative perspectives are pattern recognition, statistics, visualization, and machine learning. Chapter 3 derives a general framework after studying perspectives of feature selection. Chapter 4 starts with aspects of feature selection and illustrates representative feature selection methods to facilitate the reader in constructing his own feature selection method. Chapter 5 introduces the ways in which the methods should be evaluated, based on which empirical studies are performed. Experimental results are organized and presented in answering questions about the methods. It studies different characteristics of the data, and offers guidelines of applying feature selection methods under varied circumstances. Chapter 6 covers related topics such as feature extraction, feature construction, and feature discretization. Chapter 7 elaborates on the underlying philosophy throughout the book - *less is more* and concludes the book with prospects of feature selection in data mining. In the two appendices, we list the on-line sources for machine learning, data mining and knowledge discovery and provide descriptions of data sets and software used in the book. The soft copies of data sets and programs with instructions can be accessed from <http://www.iscs.nus.edu.sg/~liuh/Fsbook>.

This book can be considered as the first book for those, who start working on knowledge discovery but have not been exposed to feature selection, to understand the essence of feature selection and its various concepts. The book can be used by researchers in machine learning, data mining, and knowledge discovery as a toolbox in which they can find relevant tools that help in solving large real-world problems. The book can also be served as a reference book for those who are taking up tomorrow's challenges and conducting the research about feature selection.

To our families for their love,
support and understanding.

Acknowledgments

We are grateful to the editorial staff of Kluwer Academic Publishers, especially Scott Delman, Sharon Fletcher, and Robert Holland for their patience, interest, and helpfulness in bringing this project to a successful conclusion.

We have been fortunate to be assisted and encouraged by our colleagues and research associates who have contributed to this project in many ways. We are deeply indebted to them for their careful reading, thorough checking, and constructive suggestions. Although there is no space to thank them all, our special thanks to Manoranjan Dash, Farhad Hussain, Jian Shu, and Jun Yao for their invaluable assistance and extraordinary patience in this seemingly endless project, Kiansing Ng for helping design and complete the annotations in Appendices and maintaining the well constructed web sites, Rudy Setiono, Takashi Washio and Tadashi Horiuchi for fruitful collaboration without which the book would have been incomplete, Kaitat Foong and Kiansing Ng for their innovative cover design for this book.

Our families have been a constant source of encouragement throughout this project. Huan's greatest debt to his parents Baoyang Liu and Lihua Chen, and his family: Lan, Feitong, and Feige. Hiroshi's deepest gratitude to his family: Kay, Lena, Yuka and Eri.