

Multilabel Classification

Francisco Herrera · Francisco Charte
Antonio J. Rivera · María J. del Jesus

Multilabel Classification

Problem Analysis, Metrics and Techniques

Francisco Herrera
University of Granada
Granada
Spain

Antonio J. Rivera
University of Jaén
Jaén
Spain

Francisco Charte
University of Granada
Granada
Spain

María J. del Jesus
University of Jaén
Jaén
Spain

ISBN 978-3-319-41110-1

ISBN 978-3-319-41111-8 (eBook)

DOI 10.1007/978-3-319-41111-8

Library of Congress Control Number: 2016943388

© Springer International Publishing Switzerland 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG Switzerland

To my family

Francisco Herrera

To María Jesús, my beloved life partner

Francisco Charte

To my family

Antonio J. Rivera

To Jorge, my beloved life partner

María J. del Jesus

Preface

The huge growth of information stored everywhere, from mobile phones to datacenter servers, as well as the large user base of many Internet services, such as social networks and online services for publishing music, pictures, and videos, demands automated systems for categorizing and labeling all this information. A common characteristic of texts published in news sites and blogs, videos, images, and pieces of music is that all of them can be assigned to multiple categories at once. Hence, the need to have algorithms able to adequately classify the data assigning it the proper labels.

Multilabel classification is a data mining area that encompasses several tasks specific for this type of data, including custom metrics aimed to characterize multilabel datasets and also to evaluate results, specialized preprocessing methods able to solve the peculiarities of multilabeled data, and also specific classification algorithms qualified for learning from this type of data, among others. Most of these techniques are pretty new and many of them are still in development.

Multilabel classification is a topic which has generated a notable interest in late years. Beside its multiple applications to classify different types of online information, it is also useful in many other areas, such as genomic and biology. Consequently, the demand for multilabel techniques is constantly growing. This book will guide the reader to the discovery of all aspects of multilabel classification.

Based on the experience of the authors after several years focused on multilabel learning techniques, this book reviews the specificities of this kind of classification, including all the custom metrics and techniques designed to deal with it, and provides a comprehensive reference for anyone interested in the field.

After portraying the context that multilabel classification belongs to, in the introduction, a formal definition of this problem along with a broad view on how it has been faced and the fields it has been applied to are provided in the second chapter. The third one is devoted to introducing most of the publicly available multilabel use cases, as well as the metrics defined to characterize and evaluate them. Chapters 4–6 review multilabel classification methods grouping them into three groups, depending on the approach followed to tackle the task, data

transformation, method adaptation, or the use of ensembles. Two of the most relevant obstacles in working with multilabel data, high dimensionality and class imbalance, are discussed in Chaps. 7 and 8. Chapter 9 introduces several software tools and frameworks aimed to ease the work with multilabel data, including obtaining this kind of datasets, performing exploratory analysis and conducting experiments.

Although multilabel learning is still in an early development stage with respect to other data mining techniques, the amount of proposed algorithms, most of them classification methods, is impressive. In the foreseeable future, it predictably will further expand to additional application fields, and the volume of new techniques grows almost every day.

The intended audience of this book are developers and engineers aiming to apply multilabel techniques to solve different kinds of real-world problems, as well as researchers and students needing a comprehensive review on multilabel literature, methods, and tools. In addition to the text itself, the authors supply the readers with a software repository containing data, code, and links, along with two R packages as tools to work with multilabel data.

We wish to thank all our collaborators of the research groups “Soft Computing and Intelligent Information Systems” and “Intelligent Systems and Data Mining.” We are also thankful to our families for their helpful support.

Granada, Spain
Granada, Spain
Jaén, Spain
Jaén, Spain
May 2016

Francisco Herrera
Francisco Charte
Antonio J. Rivera
María J. del Jesus

Contents

1	Introduction	1
1.1	Overview	1
1.2	The Knowledge Discovery in Databases Process	2
1.3	Data Preprocessing	3
1.4	Data Mining	6
1.4.1	DM Methods Attending to Available Data	6
1.4.2	DM Methods Attending to Target Objective	7
1.4.3	DM Methods Attending to Knowledge Representation	8
1.5	Classification	11
1.5.1	Binary Classification	11
1.5.2	Multiclass Classification	12
1.5.3	Multilabel Classification	13
1.5.4	Multidimensional Classification	14
1.5.5	Multiple Instance Learning	14
	References	15
2	Multilabel Classification	17
2.1	Introduction	17
2.2	Problem Formal Definition	18
2.2.1	Definitions	18
2.2.2	Symbols	18
2.2.3	Terminology	19
2.3	Applications of Multilabel Classification	19
2.3.1	Text Categorization	20
2.3.2	Labeling of Multimedia Resources	20
2.3.3	Genetics/Biology	21
2.3.4	Other Application Fields	21
2.3.5	MLDs Repositories	22

2.4	Learning from Multilabel Data	22
2.4.1	The Data Transformation Approach.	23
2.4.2	The Method Adaptation Approach.	24
2.4.3	Ensembles of Classifiers.	25
2.4.4	Label Correlation Information.	26
2.4.5	High Dimensionality	26
2.4.6	Label Imbalance	27
2.5	Multilabel Data Tools	28
	References	29
3	Case Studies and Metrics	33
3.1	Overview	33
3.2	Case Studies	34
3.2.1	Text Categorization	34
3.2.2	Labeling of Multimedia Resources	38
3.2.3	Genetics/Biology.	40
3.2.4	Synthetic MLDs	41
3.3	MLD Characteristics	41
3.3.1	Basic Metrics	42
3.3.2	Imbalance Metrics.	43
3.3.3	Other Metrics	44
3.3.4	Summary of Characterization Metrics	45
3.4	Multilabel Classification by Example	50
3.4.1	The ML-kNN Algorithm	50
3.4.2	Experimental Configuration and Results	51
3.5	Assessing Classifiers Performance	54
3.5.1	Example-Based Metrics	55
3.5.2	Label-based Metrics	59
	References	61
4	Transformation-Based Classifiers	65
4.1	Introduction	65
4.2	Multilabel Data Transformation Approaches	66
4.3	Binary Classification Based Methods	67
4.3.1	OVO Versus OVA Approaches.	67
4.3.2	Ensembles of Binary Classifiers	68
4.4	Multiclass Classification-Based Methods.	69
4.4.1	Labelsets and Pruned Labelsets	70
4.4.2	Ensembles of Multiclass Classifiers	71
4.5	Data Transformation Methods in Practice	72
4.5.1	Experimental Configuration	72
4.5.2	Classification Results.	73
4.6	Summarizing Comments.	77
	References	78

5	Adaptation-Based Classifiers	81
5.1	Overview	81
5.2	Tree-Based Methods	82
5.2.1	Multilabel C4.5, ML-C4.5	82
5.2.2	Multilabel Alternate Decision Trees, ADTBoost.MH	82
5.2.3	Other Tree-Based Proposals	83
5.3	Neuronal Network-Based Methods	83
5.3.1	Multilabel Back-Propagation, BP-MLL	83
5.3.2	Multilabel Radial Basis Function Network, ML-RBF	84
5.3.3	Canonical Correlation Analysis and Extreme Learning Machine, CCA-ELM	85
5.4	Vector Support Machine-Based Methods	85
5.4.1	MODEL-x	85
5.4.2	Multilabel SVMs Based on Ranking, Rank-SVM and SCRANK-SVM	86
5.5	Instance-Based Methods	86
5.5.1	Multilabel kNN, ML-kNN	86
5.5.2	Instance-Based and Logistic Regression, IBLR-ML	87
5.5.3	Other Instance-Based Classifiers	87
5.6	Probabilistic Methods	88
5.6.1	Collectible Multilabel Classifiers, CML and CMLF	88
5.6.2	Probabilistic Generic Models, PMM1 and PMM2	88
5.6.3	Probabilistic Classifier Chains, PCC	89
5.6.4	Bayesian and Tree Naïve Bayes Classifier Chains, BCC and TNBCC	89
5.6.5	Conditional Restricted Boltzmann Machines, CRBM	89
5.7	Other MLC Adaptation-Based Methods	90
5.8	Adapted Methods in Practice	91
5.8.1	Experimental Configuration	92
5.8.2	Classification Results	92
5.9	Summarizing Comments	97
	References	98
6	Ensemble-Based Classifiers	101
6.1	Introduction	101
6.2	Ensembles of Binary Classifiers	102
6.2.1	Ensemble of Classifier Chains, ECC	102
6.2.2	Ranking by Pairwise Comparison, RPC	102
6.2.3	Calibrated Label Ranking, CLR	103
6.3	Ensembles of Multiclass Classifiers	103
6.3.1	Ensemble of Pruned Sets, EPS	103
6.3.2	Random k-Labelsets, RAKEL	104
6.3.3	Hierarchy of Multilabel Classifiers, HOMER	104
6.4	Other Ensembles	104

6.5	Ensemble Methods in Practice.	105
6.5.1	Experimental Configuration	106
6.5.2	Classification Results.	107
6.5.3	Training and Testing Times	110
6.6	Summarizing Comments.	111
	References	112
7	Dimensionality Reduction.	115
7.1	Overview	115
7.1.1	High-Dimensional Input Space	116
7.1.2	High-Dimensional Output Space	117
7.2	Feature Space Reduction	117
7.2.1	Feature Engineering Approaches	118
7.2.2	Multilabel Supervised Feature Selection	119
7.2.3	Experimentation	121
7.3	Label Space Reduction.	124
7.3.1	Sparseness and Dependencies Among Labels	124
7.3.2	Proposals for Reducing Label Space Dimensionality	125
7.3.3	Experimentation	126
7.4	Summarizing Comments.	129
	References	129
8	Imbalance in Multilabel Datasets	133
8.1	Introduction	133
8.2	Imbalanced MLD Specificities.	134
8.2.1	How to Measure the Imbalance Level	135
8.2.2	Concurrence Among Imbalanced Labels.	136
8.3	Facing Imbalanced Multilabel Classification	138
8.3.1	Classifier Adaptation	138
8.3.2	Resampling Techniques	139
8.3.3	The Ensemble Approach	145
8.4	Multilabel Imbalanced Learning in Practice.	146
8.4.1	Experimental Configuration	147
8.4.2	Classification Results.	147
8.5	Summarizing Comments.	150
	References	150
9	Multilabel Software	153
9.1	Overview	153
9.2	Working with Multilabel Data.	154
9.2.1	Multilabel Data File Formats	154
9.2.2	Multilabel Data Repositories.	155
9.2.3	The mldr.datasets Package	157
9.2.4	Generating Synthetic MLDs	162

9.3	Exploratory Analysis of MLDs	162
9.3.1	MEKA.	163
9.3.2	The mldr Package	166
9.4	Conducting Multilabel Experiments	179
9.4.1	MEKA.	179
9.4.2	MULAN	182
9.4.3	The RunMLClassifier Utility.	188
9.5	Summarizing Comments.	189
	References	190
	Glossary	193

Acronyms

ACO	Ant colony optimization
ADT	Alternative decision trees
ANN	Artificial neural network
API	Application programing interface
ARFF	Attribute-Relation File Format
AUC	Area under the ROC curve
BCC	Bayesian classifier chains
BID	Binary datasets
BoW	Bag of words
BR	Binary relevance
CC	Classifier chains
CCA	Canonical correlation analysis
CDE	ChiDep ensemble
CL	Compressed labeling
CLR	Calibrated label ranking
CML	Collectible multilabel
CMLPC	Calibrated pairwise multilabel perceptron
CRAN	Comprehensive R Archive Network
CRF	Conditional random fields
CS	Compressed sensing
CSV	Comma-separated values
CT	Classifier trellis
CV	Cross-validation
CVIR	Coefficient of variation for the average imbalance ratio (<i>MeanIR</i>)
CVM	Core vector machine
DLVM	Dual-layer Voting Method
DM	Data mining
DT	Decision trees
ECC	Ensemble of classifier chains
ELM	Extreme learning machine
EML	Ensemble of multilabel learners

EPS	Ensemble of pruned sets
FN	False negatives
FP	False positives
IBL	Instance-based learning
IR	Imbalance ratio or information retrieval depending on the context
JDK	Java Development Kit
JRE	Java Runtime Environment
KDD	Knowledge discovery in databases
KDE	Kernel dependency estimation
kNN	k-nearest neighbors
LDA	Linear discriminant analysis
LP	Label powerset
LSI	Latent semantic indexing
MAP	Maximum a posteriori probabilities
MCD	Multiclass datasets
MIR	Mean imbalance ratio
MLC	Multilabel classification
MLD	Multilabel dataset
MLP	Multilayer perceptron
OVA	One-vs-all
OVO	One-vs-one
PCA	Principal component analysis
PCC	Probabilistic classifier chains
PCT	Predictive clustering tree
PMM	Probabilistic mixture models
PS	Pruned sets
PSO	Particle swarm optimization
QCLR	QWeighted calibrated label ranking
RAkEL	Random k-labelsets
RBFN	Radial basis function network
RBM	Restricted Boltzmann machine
RF-PCT	Random forest of predictive clustering trees
ROC	Receiver operating characteristic
ROS	Random over-sampling
RPC	Ranking by pairwise comparison
RUS	Random under sampling
SOM	Self-organizing map
SVD	Single-value decomposition
SVM	Support vector machine
SVN	Support vector network
TF/IDF	Term frequency/inverse document frequency
TN	True negatives
TP	True positives