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

Multilabel Classification

Problem Analysis, Metrics and Techniques

Springer

Multilabel Classification

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

Multilabel Classification

Problem Analysis, Metrics and Techniques

Springer

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

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 — Francisco Herrera
Granada, Spain — Francisco Charte
Jaén, Spain — Antonio J. Rivera
Jaén, Spain — María J. del Jesus
May 2016

Contents

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 (*MeanIR*)
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

Chapter 1
Introduction

Abstract This book is focused on multilabel classification and related topics. Multilabel classification is one specific type of classification, classification being one of the usual tasks in the data mining field. Data mining itself can be seen as a step into a broad process, the discovery of new knowledge from databases. The goal of this first chapter is to introduce all these concepts, aiming to set the working context for the topics covered in the following ones. A global outline to this respect is given in Sect. 1.1. Section 1.2 provides an overview of the whole Knowledge Discovery in Databases process. Section 1.3 introduces the essential preprocessing tasks. Then, the different learning styles in use nowadays are explained in Sect. 1.4, and lastly multilabel classification is introduced in comparison with other traditional types of classification in Sect. 1.5.

1.1 Overview

The technological progress in late years has propelled the availability of huge amounts of data. Storage and communication capabilities have grown exponentially, increasing the needs to automatically process all these data. Due to this fact, machine learning techniques have acquired considerable relevance. In particular, the automatic classification of all kind of digital information, including texts, photos, music and videos, is in growing demand. Multilabel classification is the field where methods to perform this task, labeling resources into several non-exclusive categories, are studied and proposed.

This book presents a review of multilabel classification procedures and related techniques, including the analysis of obstacles specifically tied to this class of methods. Experimentation results from the most relevant proposals are also provided. The goal of this first chapter is to set the context multilabel classification belongs to. It starts from the wide view of the whole Knowledge Discovery in Databases (KDD) process, then narrowing the focus until nonstandard classification methods, where multilabel classification is introduced.

F. Herrera et al., *Multilabel Classification*,
DOI 10.1007/978-3-319-41111-8_1