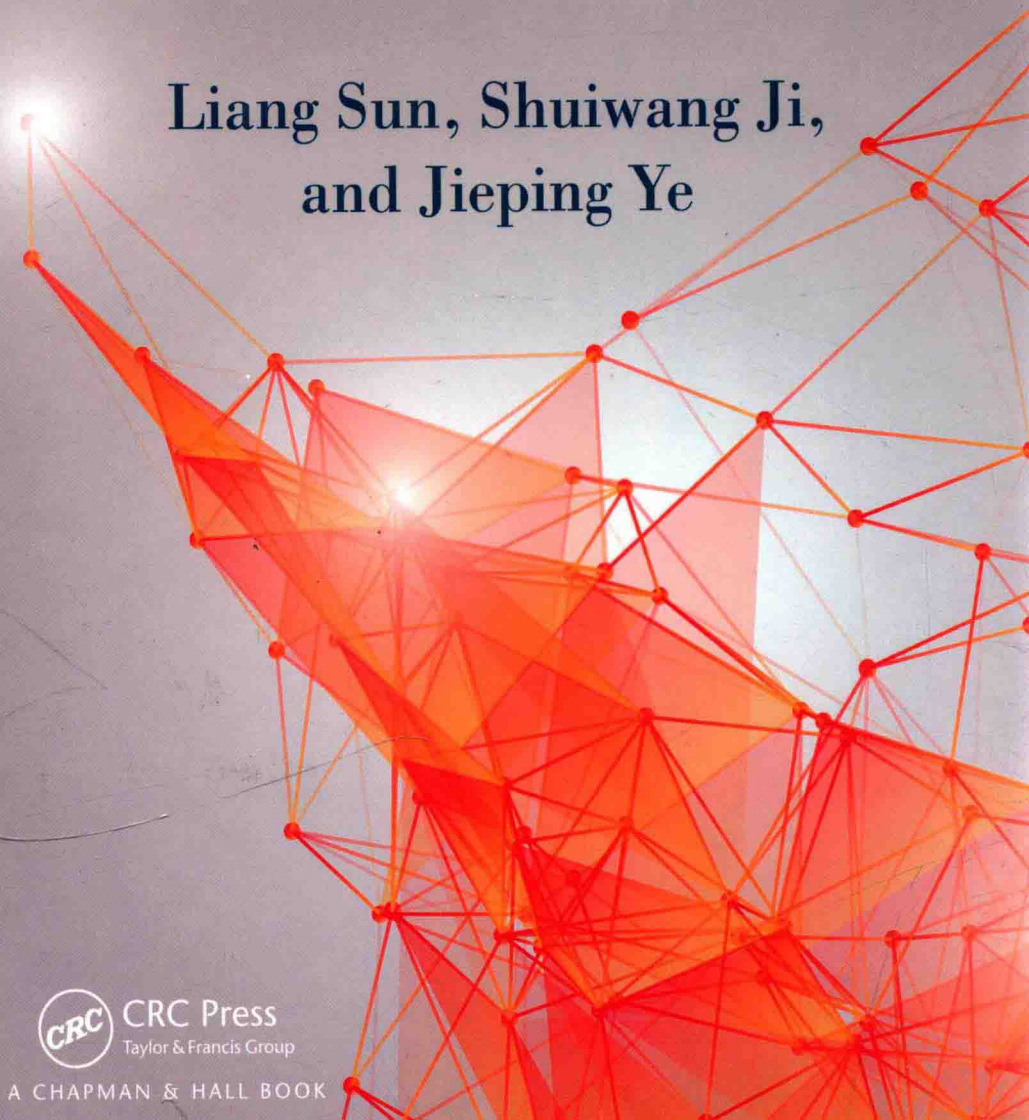
Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series

# Multi-Label Dimensionality Reduction

Liang Sun, Shuiwang Ji,
and Jieping Ye



CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series

# Multi-Label Dimensionality Reduction



Liang Sun, Shuiwang Ji, and Jieping Ye

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742



Printed on acid-free paper
Version Date: 20131009

International Standard Book Number-13: 978-1-4398-0615-9 (Hardback)

# Multi-Label Dimensionality Reduction

Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series

SERIES EDITORS

**Ralf Herbrich**
Amazon Development Center
Berlin, Germany

**Thore Graepel**
Microsoft Research Ltd.
Cambridge, UK

**AIMS AND SCOPE**

This series reflects the latest advances and applications in machine learning and pattern recognition through the publication of a broad range of reference works, textbooks, and handbooks. The inclusion of concrete examples, applications, and methods is highly encouraged. The scope of the series includes, but is not limited to, titles in the areas of machine learning, pattern recognition, computational intelligence, robotics, computational/statistical learning theory, natural language processing, computer vision, game AI, game theory, neural networks, computational neuroscience, and other relevant topics, such as machine learning applied to bioinformatics or cognitive science, which might be proposed by potential contributors.

**PUBLISHED TITLES**

MACHINE LEARNING: An Algorithmic Perspective
*Stephen Marsland*

HANDBOOK OF NATURAL LANGUAGE PROCESSING,
Second Edition
*Nitin Indurkhya and Fred J. Damerau*

UTILITY-BASED LEARNING FROM DATA
*Craig Friedman and Sven Sandow*

A FIRST COURSE IN MACHINE LEARNING
*Simon Rogers and Mark Girolami*

COST-SENSITIVE MACHINE LEARNING
*Balaji Krishnapuram, Shipeng Yu, and Bharat Rao*

ENSEMBLE METHODS: FOUNDATIONS AND ALGORITHMS
*Zhi-Hua Zhou*

MULTI-LABEL DIMENSIONALITY REDUCTION
*Liang Sun, Shuiwang Ji, and Jieping Ye*

# Preface

Multi-label learning concerns supervised learning problems in which each instance may be associated with multiple labels simultaneously. A key difference between multi-label learning and traditional binary or multi-class learning is that the labels in multi-label learning are not mutually exclusive. Multi-label learning arises in many real-world applications. For example, in web page categorization, a web page may contain multiple topics. In gene and protein function prediction, multiple functional labels may be associated with each gene and protein, since an individual gene or protein usually performs multiple functions. In automated newswire categorization, multiple labels can be associated with a newswire story indicating its subject categories and the regional categories of reported events. Motivated by the increasing number of applications, multi-label learning has attracted significant attention in data mining and machine learning recently.

In comparison with traditional binary and multi-class classification, multi-label classification is more general and is thus more challenging to solve. One significant challenge in multi-label learning is how to effectively exploit the label structure to improve classification performance. Since the labels in multi-label learning are often correlated, how to measure and capture the correlations in the label space for improved prediction is crucial. This problem becomes particularly important when more sophisticated relations, such as hierarchical structures, exist among labels. Another challenge of multi-label learning lies in the class imbalance problem. When each label is modeled independently, the number of instances related to a specific label is much less than the number of instances that are not related to this label. In this case, it is difficult to build a highly accurate classifier for these labels without considering label correlations. The third challenge is concerned with the effectiveness and efficiency of multi-label learning for large-scale problems, especially when both the data dimensionality and the number of labels are large.

Similar to other data mining and machine learning tasks, multi-label learning also suffers from the so-called curse of dimensionality. Dimensionality reduction, which extracts a small number of features by removing irrelevant, redundant, and noisy information, is an effective way to mitigate the curse of dimensionality. Although dimensionality reduction has been well studied in the literature, we lack a unified treatment of multi-label dimensionality reduction that includes both algorithmic developments and applications. In this monograph, we give a selective treatment of dimensionality reduction for multi-label learning with emphasis on our own work. We cover a wide variety of topics, ranging from methodological developments to theoretical properties, computational algorithms, and applications. Specifically, this book

focuses on the following fundamental research questions posed by multi-label dimensionality reduction: How to fully exploit label correlations for effective dimensionality reduction; How to scale dimensionality reduction algorithms to large-scale problems; How to effectively combine dimensionality reduction with classification; How to derive sparse dimensionality reduction algorithms to enhance model interpretability; How to perform multi-label dimensionality reduction effectively in practical applications. To expedite the applications of these algorithms, a MATLAB® software package that implements many popular dimensionality reduction algorithms is provided online at http://www.public.asu.edu/~jye02/Software/MLDR/. We hope that this book will appeal to both researchers and practitioners in diverse areas working on multi-label learning.

We would like to thank many people who have supported, encouraged, and inspired us during the preparation of this book. We are deeply indebted to Prof. Sudhir Kumar and his FlyExpress team, who provided support in the exploration of gene expression pattern image annotation. We would like to thank all former and current members of the machine learning research laboratory at Arizona State University, including Betul Ceran, Rita Chattopadhyay, Jianhui Chen, Pinghua Gong, Jun Liu, Yashu Liu, Zhi Nie, Qian Sun, Jie Wang, Zhen Wang, Shuo Xiang, Sen Yang, Lei Yuan, Zheng Zhao, Jiayu Zhou, and Chao Zhang. Each and every member of this dynamic group has helped us in various ways during numerous discussions and interactions. Many of our colleagues provided thoughtful reviews. We thank Jun Li, Shan Yang, Hang Zhang, and Hou Zhou for their feedback. We would like to thank anonymous reviewers for their constructive comments. We are grateful to the National Science Foundation for supporting our research on multi-label dimensionality reduction. Last but not least, we would like to thank our families for their love, understanding, and support.

Liang Sun — San Diego, California
Shuiwang Ji — Norfolk, Virginia
Jieping Ye — Tempe, Arizona

## Symbol Description

| | |
|---|---|
| $x$ | Variable. |
| $\mathbf{x}$ | Vector. |
| $\mathbf{A}$ | Matrix. |
| $\mathbb{R}$ | The set of real numbers. |
| $\mathbb{R}^n$ | The set of real $n$-vectors ($n \times 1$ matrices). |
| $\mathbb{R}_+$ | The set of nonnegative real numbers. |
| $\mathbb{R}_{++}$ | The set of positive real numbers. |
| $\mathbb{S}^n$ | The set of symmetric $n \times n$ matrices. |
| $\mathbb{S}^n_+$ | The set of symmetric and positive semidefinite $n \times n$ matrices. |
| $\mathbb{S}^n_{++}$ | The set of symmetric and positive definite $n \times n$ matrices. |
| $\mathbf{1}$ | The vector of ones. |
| $\mathbf{0}$ | The matrix or vector of all zeros. |
| $\mathbf{I}$ | The identity matrix. |
| $\mathbf{A}^T$ | The transpose of matrix $\mathbf{A}$. |
| $\mathrm{Tr}(\mathbf{A})$ | The trace of matrix $\mathbf{A}$. |
| $\mathbf{A}^\dagger$ | The Moore-Penrose pseudoinverse of matrix $\mathbf{A}$. |
| $\mathrm{rank}(\mathbf{A})$ | The rank of matrix $\mathbf{A}$. |
| $\mathcal{R}(\mathbf{A})$ | The range space of matrix $\mathbf{A}$. |
| $\mathcal{N}(\mathbf{A})$ | The null space of matrix $\mathbf{A}$. |
| $\|\mathbf{v}\|$ | A norm of vector $\mathbf{v}$. |
| $\|\mathbf{v}\|_1$ | 1-norm of vector $\mathbf{v}$. |
| $\|\mathbf{v}\|_2$ | 2-norm of vector $\mathbf{v}$. |
| $\|\mathbf{v}\|_\infty$ | $\infty$-norm of vector $\mathbf{v}$. |
| $\|\mathbf{A}\|$ | A norm of matrix $\mathbf{A}$. |
| $\|\mathbf{A}\|_1$ | 1-norm of matrix $\mathbf{A}$. |
| $\|\mathbf{A}\|_2$ | 2-norm of matrix $\mathbf{A}$. |
| $\|\mathbf{A}\|_\infty$ | $\infty$-norm of matrix $\mathbf{A}$. |
| $\|\mathbf{A}\|_F$ | Frobenius norm of matrix $\mathbf{A}$. |
| $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ | Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. |
| $\mathrm{diag}\,(\mathbf{v})$ | Diagonal matrix with diagonal entries from $\mathbf{v}$. |
| $d$ | The data dimensionality. |
| $n$ | The number of samples. |
| $k$ | The number of labels. |
| $\mathcal{X}$ | Input instance space. |
| $\mathcal{Y}$ | The label space $\{0, 1\}^k$. |
| $\mathfrak{L}$ | The label set $\mathfrak{L} = \{C_1, \cdots, C_k\}$. |
| $\mathcal{L}$ | The normalized Laplacian matrix. |
| $\mathbf{X}$ | The data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the $i$th sample. |
| $\mathbf{Y}$ | The label matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \{0, 1\}^{k \times n}$, where $\mathbf{y}_i \in \{0, 1\}^k$ is the label information for the $i$th sample. |
| $\mathbf{K}$ | The kernel matrix. |
| $\mathbf{A}_i$ | The $i$th column of matrix $\mathbf{A}$. |
| $\mathbf{A}^i$ | The $i$th row of matrix $\mathbf{A}$. |
| $\mathbb{E}[x]$ | The expectation of variable $x$. |
| $\mathbb{P}[e]$ | The probability of event $e$. |
| $\mathrm{std}(x)$ | The standard deviation of variable $x$. |
| $\mathrm{cov}(x, y)$ | The covariance between variables $x$ and $y$. |
| $\mathrm{sgn}(x)$ | The sign function. |
| $\|S\|$ | The cardinality of set $S$. |

# Contents

**Preface** xi

**List of Symbols** xiii

**1 Introduction** 1
- 1.1 Introduction to Multi-Label Learning . . . 1
- 1.2 Applications of Multi-Label Learning . . . 2
  - 1.2.1 Scene Classification . . . 2
  - 1.2.2 Text Categorization . . . 3
  - 1.2.3 Functional Genomics Analysis . . . 4
  - 1.2.4 Gene Expression Pattern Image Annotation . . . 6
- 1.3 Challenges of Multi-Label Learning . . . 8
- 1.4 State of the Art . . . 9
  - 1.4.1 Problem Transformation . . . 9
    - 1.4.1.1 Copy Transformation . . . 9
    - 1.4.1.2 Binary Relevance . . . 10
    - 1.4.1.3 Label Power-Set . . . 10
    - 1.4.1.4 Single-Label Classification after Transformation . 12
  - 1.4.2 Algorithm Adaptation . . . 12
    - 1.4.2.1 Decision Tree . . . 12
    - 1.4.2.2 Algorithms Based on Probabilistic Framework . . 12
    - 1.4.2.3 Support Vector Machines . . . 13
    - 1.4.2.4 Artificial Neural Networks . . . 14
    - 1.4.2.5 $k$-Nearest Neighbor . . . 15
    - 1.4.2.6 Ensemble Learning . . . 15
    - 1.4.2.7 Other Algorithms . . . 16
- 1.5 Dimensionality Reduction for Multi-Label Learning . . . 18
  - 1.5.1 Introduction to Dimensionality Reduction . . . 18
  - 1.5.2 Linear and Nonlinear Dimensionality Reduction . . . 20
  - 1.5.3 Multi-Label Dimensionality Reduction . . . 20
  - 1.5.4 Related Work . . . 22
- 1.6 Overview of the Book . . . 23
  - 1.6.1 Design and Analysis of Algorithms . . . 24
  - 1.6.2 Scalable Implementations . . . 25
  - 1.6.3 Applications . . . 26
- 1.7 Notations . . . 26

1.8 Organization . . . . . 27

**2 Partial Least Squares** **29**
2.1 Basic Models of Partial Least Squares . . . . . 29
2.1.1 The NIPALS Algorithm . . . . . 30
2.2 Partial Least Squares Variants . . . . . 31
2.2.1 PLS Mode A . . . . . 32
2.2.2 PLS2 . . . . . 32
2.2.3 PLS1 . . . . . 33
2.2.4 PLS-SB . . . . . 34
2.2.5 SIMPLS . . . . . 34
2.2.6 Orthonormalized PLS . . . . . 34
2.2.7 Relationship between OPLS and Other PLS Models . . . . 36
2.3 PLS Regression . . . . . 37
2.3.1 Basics of PLS Regression . . . . . 37
2.3.2 Shrinkage in Regression . . . . . 38
2.3.3 Principal Component Regression . . . . . 41
2.3.4 Ridge Regression . . . . . 41
2.3.5 Shrinkage Properties of PLS Regression . . . . . 43
2.4 Partial Least Squares Classification . . . . . 44

**3 Canonical Correlation Analysis** **49**
3.1 Classical Canonical Correlation Analysis . . . . . 49
3.1.1 Linear Correlation Coefficient . . . . . 49
3.1.2 The Maximum Correlation Formulation of CCA . . . . . 50
3.1.3 The Distance Minimization Formulation of CCA . . . . . 54
3.1.4 Regularized CCA . . . . . 55
3.1.5 CCA for Multiple Sets of Variables . . . . . 55
3.2 Sparse CCA . . . . . 56
3.2.1 Sparse CCA via Linear Regression . . . . . 57
3.2.2 Sparse CCA via Iterative Greedy Algorithms . . . . . 57
3.2.3 Sparse CCA via Bayesian Learning . . . . . 58
3.3 Relationship between CCA and Partial Least Squares . . . . . 59
3.3.1 A Unified Framework for PLS and CCA . . . . . 59
3.3.2 The Equivalence without Regularization . . . . . 60
3.3.3 The Equivalence with Regularization . . . . . 61
3.3.3.1 Regularization on $\mathbf{X}$ . . . . . 62
3.3.3.2 Regularization on $\mathbf{Y}$ . . . . . 62
3.3.4 Analysis of Regularization on CCA . . . . . 63
3.4 The Generalized Eigenvalue Problem . . . . . 64
3.4.1 The Generalized Rayleigh Quotient Cost Function . . . . . 64
3.4.2 Properties of the Generalized Eigenvalue Problem . . . . . 65
3.4.3 Algorithms for the Generalized Eigenvalue Problem . . . . 66

**4 Hypergraph Spectral Learning** **69**
4.1 Hypergraph Basics . . . . . 69
4.1.1 Clique Expansion . . . . . 71
4.1.2 Star Expansion . . . . . 72
4.1.3 Hypergraph Laplacian . . . . . 73
4.2 Multi-Label Learning with a Hypergraph . . . . . 74
4.3 A Class of Generalized Eigenvalue Problems . . . . . 75
4.3.1 Canonical Correlation Analysis . . . . . 76
4.3.2 Orthonormalized Partial Least Squares . . . . . 76
4.3.3 Hypergraph Spectral Learning . . . . . 77
4.3.4 Linear Discriminant Analysis . . . . . 77
4.4 The Generalized Eigenvalue Problem versus the Least Squares Problem . . . . . 78
4.4.1 Multivariate Linear Regression and Least Squares . . . . . 78
4.4.2 Matrix Orthonormality Property . . . . . 79
4.4.3 The Equivalence Relationship . . . . . 81
4.4.4 Regularized Least Squares . . . . . 82
4.4.5 Efficient Implementation via LSQR . . . . . 83
4.5 Empirical Evaluation . . . . . 84
4.5.1 Empirical Evaluation Setup . . . . . 84
4.5.2 Performance of Hypergraph Spectral Learning . . . . . 85
4.5.3 Evaluation of the Equivalence Relationship . . . . . 86
4.5.4 Evaluation of Scalability . . . . . 88

**5 A Scalable Two-Stage Approach for Dimensionality Reduction** **91**
5.1 The Two-Stage Approach without Regularization . . . . . 91
5.1.1 The Algorithm . . . . . 92
5.1.2 Time Complexity Analysis . . . . . 92
5.1.3 The Equivalence Relationship . . . . . 93
5.2 The Two-Stage Approach with Regularization . . . . . 95
5.2.1 The Algorithm . . . . . 96
5.2.2 Time Complexity Analysis . . . . . 96
5.2.3 The Equivalence Relationship . . . . . 96
5.3 Empirical Evaluation . . . . . 99
5.3.1 Empirical Evaluation Setup . . . . . 99
5.3.2 Performance Comparison . . . . . 100
5.3.3 Scalability Comparison . . . . . 101

**6 A Shared-Subspace Learning Framework** **105**
6.1 The Framework . . . . . 105
6.1.1 Problem Formulation . . . . . 105
6.1.2 A Trace Ratio Formulation . . . . . 107
6.2 An Efficient Implementation . . . . . 109
6.2.1 Reformulation . . . . . 109
6.2.2 Eigendecomposition . . . . . 110

6.2.3 The Main Algorithm . . . 111
6.3 Related Work . . . 111
6.4 Connections with Existing Formulations . . . 112
6.5 A Feature Space Formulation . . . 113
6.6 Empirical Evaluation . . . 114
6.6.1 Empirical Evaluation Setup . . . 115
6.6.2 Web Page Categorization . . . 116
6.6.2.1 Performance Evaluation . . . 116
6.6.2.2 Scalability Evaluation . . . 118
6.6.2.3 Sensitivity Analysis . . . 121
6.6.3 Discussion . . . 121

**7 Joint Dimensionality Reduction and Classification 123**
7.1 Background . . . 123
7.1.1 Squared Loss . . . 124
7.1.2 Hinge Loss . . . 124
7.2 Joint Dimensionality Reduction and Multi-Label Classification . . . 125
7.2.1 Joint Learning with Squared Loss . . . 125
7.2.2 Joint Learning with Hinge Loss . . . 126
7.2.2.1 A Convex-Concave Formulation . . . 127
7.2.2.2 Solving the Min-Max Problem . . . 128
7.2.2.3 Learning Orthonormal Features . . . 128
7.2.2.4 Joint Learning with Squared Hinge Loss . . . 128
7.2.2.5 Related Work . . . 129
7.3 Dimensionality Reduction with Different Input Data . . . 129
7.4 Empirical Evaluation . . . 130
7.4.1 Evaluation on Multi-Label Data Sets . . . 130
7.4.2 Evaluation on Data with Different Inputs . . . 131

**8 Nonlinear Dimensionality Reduction: Algorithms and Applications 133**
8.1 Background on Kernel Methods . . . 133
8.2 Kernel Centering and Projection . . . 134
8.2.1 Kernel Centering . . . 135
8.2.2 Kernel Projection . . . 135
8.3 Kernel Canonical Correlation Analysis . . . 136
8.4 Kernel Hypergraph Spectral Learning . . . 138
8.5 The Generalized Eigenvalue Problem in the Kernel-Induced Feature Space . . . 139
8.6 Kernel Least Squares Regression . . . 140
8.7 Dimensionality Reduction and Least Squares Regression in the Feature Space . . . 140
8.7.1 Matrix Orthonormality Property . . . 140
8.7.2 The Equivalence Relationship . . . 142
8.8 Gene Expression Pattern Image Annotation . . . 143
8.8.1 Problem Description . . . 143

8.8.2 Feature Generation and Kernel Construction . . . . . 145
8.8.3 Multi-Label Multiple Kernel Learning . . . . . 147
8.8.4 Empirical Evaluation Setup . . . . . 149
8.8.5 Annotation Results . . . . . 150

**Appendix Proofs** **155**
A.1 Proofs for Chapter 2 . . . . . 155
A.2 Proofs for Chapter 3 . . . . . 159
A.3 Proofs for Chapter 4 . . . . . 161
A.4 Proofs for Chapter 6 . . . . . 162
A.5 Proofs for Chapter 8 . . . . . 164

**References** **167**

**Index** **191**

# Chapter 1

# Introduction

## 1.1 Introduction to Multi-Label Learning

Supervised learning is concerned with inferring the relations between input instances and class labels. In traditional classification tasks, each instance is associated with one class label. However, in many real-world scenarios, one instance may be associated with multiple labels. For example, in news categorization, a piece of news regarding Apple's release of a new iPhone is associated with both the label *business* and the label *technology*. In other words, each instance is associated with a set of labels instead of only one label. Multi-label learning is a machine learning field devoted to learning from multi-label data in which each instance is associated with potentially multiple labels. A major difference between multi-label learning and traditional binary or multi-class learning is that the labels in multi-label learning are not mutually exclusive, suggesting that each instance may be relevant to multiple labels. Thus, one of the key challenges of multi-label learning is how to exploit the correlations among different labels effectively.

In this book, we assume that each instance in the training set is represented as a pair of vectors, one for the input features and the other for the output labels. Multi-label learning concerns the prediction of the labels of unseen instances by building a classifier based on the training data. Formally, let $\mathcal{X}$ and $\mathcal{Y}$ denote the input instance space and the output label space, respectively. In multi-label learning, the label space $\mathcal{Y}$ is defined as $\mathcal{Y} = \{0, 1\}^k$, where $k$ is the number of labels. That is, the $j$th component of the label vector is 1 if the instance is relevant to the $j$th label, and it is 0 otherwise. Similar to traditional classification, given a training data set, the goal of multi-label learning is to learn a classifier $f : \mathcal{X} \to \mathcal{Y}$, which predicts the labels of each instance $\mathbf{x} \in \mathcal{X}$. Specifically, the output of the classifier $f$ for a given instance $\mathbf{x} \in \mathcal{X}$ is

$$f(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_k(\mathbf{x})]^T, \tag{1.1}$$

where $f_j(\mathbf{x})\,(j = 1, \cdots, k)$ is either 1 or 0, indicating the association of $\mathbf{x}$ with the $j$th label. In the following, the set of labels is denoted as $\mathfrak{L} = \{C_1, \cdots, C_k\}$.

Multi-label learning finds applications in many real-world applications, such as text categorization [167, 279], image annotation [34, 126], bioinformatics [28, 279], 3D hand pose estimation [216], and biological literature classification [128]. Motivated by the increasing number of applications, multi-label learning has recently attracted significant attention, and many algorithms have been proposed [190, 235,

237]. These methods are reviewed in Section 1.4 and can be divided into two major categories:

1. *Problem transformation*: This class of methods first transforms the multi-label learning problem into a series of single-label problems, which are then solved using existing single-label learning methods.

2. *Algorithm adaptation*: This class of methods solves the multi-label problems directly by adapting existing methods for single-label learning.

Similar to other machine learning and data mining tasks, multi-label learning also suffers from the so-called *curse of dimensionality* [21]. Although there has been extensive research on dimensionality reduction in the literature, multi-label dimensionality reduction has not been well explored [8, 273, 280]. This book is devoted to the study of *multi-label dimensionality reduction*, which focuses on extracting a small number of features from multi-label data by removing the irrelevant, redundant, and noisy information while exploiting information from the label space such as the correlation among different labels. Specifically, we give a unified treatment of multi-label dimensionality reduction approaches in methodological developments, theoretical properties, computational algorithms, and applications.

In the rest of this chapter, we will briefly introduce multi-label learning and dimensionality reduction, including existing algorithms, applications, and related work. We will also highlight the main challenges of multi-label dimensionality reduction.

## 1.2 Applications of Multi-Label Learning

Multi-label learning has been applied successfully in many real-world applications. In this section, we present several representative examples, including scene classification, text categorization, functional genomics analysis, and gene expression pattern image annotation.

### 1.2.1 Scene Classification

Humans are very proficient at perceiving natural scenes and understanding their contents. In scene classification, the task is to determine the associated semantic labels, such as *mountain*, *lake*, or *party*, for given images. Scene classification finds applications in many areas, including content-based image indexing and content-sensitive image enhancement [34]. For example, many current digital library systems support content-based image retrieval, which allows the user to retrieve images that are similar to a given query image [105]. In this case, knowledge of the semantic labels of the query image can reduce the search space and improve the retrieval

(A) Scene 1

(B) Scene 2

FIGURE 1.1: Examples of multi-label scenes. The first scene (A) is associated with the labels "*lake*" and "*mountain*", and the second scene (B) is associated with the labels "*river*" and "*mountain*".

accuracy. Since a natural scene may contain multiple objects, each image can be associated with multiple labels. Hence, scene classification is naturally a multi-label learning problem. For example, Figure 1.1(A) shows an image associated with the labels "*lake*" and "*mountain*"; Figure 1.1(B) shows an image associated with the labels "*river*" and "*mountain*".

### 1.2.2 Text Categorization

Text Categorization (TC) is the task of classifying text documents into one or more of a set of predefined categories or subject codes [131, 208]. Originally dating back to the early 1960s, the effectiveness of text categorization has been improved significantly in the past decades mainly due to the advances of machine learning methods [131]. Text categorization has been applied in various fields, including web page categorization using hierarchical labels, detection of text genre, text (or hypertext) documents classification given a predefined label set, personalized information delivery, and content filtering [208]. Typically, the predefined labels (or categories) in text categorization are not assumed to be mutually exclusive; thus text categorization can naturally be modeled as a multi-label learning problem. For instance, consider labels *business, technology, entertainment*, and *politics* in news categorization; a news article about Apple's release of a new iPhone may be labeled with both the label *business* and the label *technology*.

When applying multi-label learning to perform text categorization, the first step is to encode documents using a suitable representation, such as the ones based on the vector space model [279] and the binary representation [140]. In the past, many multi-label learning algorithms have been proposed to perform text categorization [87, 136, 167, 206, 240, 279]. One well-known algorithm in text categorization is BoosTexter, which extends the classical boosting algorithm AdaBoost [80] to handle multi-label data. Some other algorithms include the Bayesian approach [167] using the mixture model coupled with the Expectation–Maximization (EM) algorithm, and

the Maximal Figure-of-Merit (MFoM) approach [87]. We will review existing multi-label learning approaches in Section 1.4.

One widely used benchmark data set in multi-label text categorization is the Reuters-21578 data set[1]. This data set was originally collected and labeled by the Carnegie Group, Inc. and Reuters, Ltd. in the course of developing the CONSTRUE text categorization system. Reuters-21578 consists of 21,578 Reuters newswire documents that appeared in 1987. Almost all documents in the Reuters collection come with title, dateline, and text body, and the number of topics (labels) is 135. In particular, three widely used subsets of the Reuters-21578 data set have been extracted [140] by identifying the labels that suggest parent–child relationships, and the labels are organized in a hierarchical structure, as shown in Figure 1.2. Note that the roots of the three category trees are virtual categories.

Another data set that has become very popular for text categorization in recent years is the Reuters Corpus Volume 1 (RCV1) data set[2] [155]. The RCV1 data set consists of over 800,000 manually categorized newswire stories recently made available by Reuters, Ltd. for research purposes. Similar to the Reuters-21578 data set, the labels in the RCV1 data set are organized in a hierarchical structure. The original data set is referred to as RCV1-v1, and a corrected version called RCV1-v2 was generated and has become more popular in text categorization research. More details on this data set can be found in [155].

### 1.2.3 Functional Genomics Analysis

Functional genomics is an important field in bioinformatics. It studies gene and protein functions by conducting large-scale analysis on a vast amount of data collected by genome projects [123, 159]. For example, DNA microarrays allow researchers to simultaneously measure the expression levels of thousands of different genes, and overwhelming amounts of data are produced [161]. Recently, a large body of research has been devoted to automatic analysis of microarray data [159]. In automated gene expression analysis, the task is to predict the functions for genes. Generally, it is based on the assumption that genes with similar functions have similar expression profiles in cells [123]. Note that each gene may be associated with multiple functions in functional genomics. When the functions are considered as labels, the function prediction problem in functional genomics can be modeled as a multi-label learning problem.

A widely used benchmark data set in multi-label learning for functional genomics is the Yeast data set [38, 74]. The Yeast data set consists of microarray expression data and phylogenetic profiles from the budding yeast *Saccharomyces cerevisiae*. It contains 2417 samples, and each sample is represented as a 103-dimensional feature vector[3]. Each sample (gene) is associated with a subset of a total of 190 functional labels. The functional classes (labels) are organized in a tree structure, which is known in the literature [38, 74]. This data set is preprocessed in [74] and only the function

[1] http://www.research.att.com/~lewis/reuters21578.html

[2] http://www.daviddlewis.com/resources/testcollections/rcv1/

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html#yeast