HOKKAIDO UNIVERSITY

Title	On Improving Multi-Label Classification via Dimension Reduction [an abstract of dissertation and a summary of dissertation review]
Author(s)	孫, 露
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第12850号
Issue Date	2017-09-25
Doc URL	https://hdl.handle.net/2115/67456
Rights(URL)	https://creativecommons.org/licenses/by-nc-sa/2.1/jp/
Type	doctoral thesis
File Information	Lu_Sun_abstract.pdf, 論文内容の要旨

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学　位　論　文　内　容　の　要　旨

博士の専攻分野の名称　　博士（情報科学）　　氏名　孫 露

学　位　論　文　題　名

On Improving Multi-Label Classification via Dimension Reduction
（次元縮約によるマルチラベル識別の改善）

In recent years we have witnessed an explosively increasing of web-related applications in our daily lives, where the scale of data and information has grown dramatically. To deal with such a huge amount of data, machine learning becomes a crucial way to help human beings to be free from the massive tasks, such as classification, pattern recognition and prediction. As one of the fundamental tasks, classification has attracted a lot of attentions from researchers, and been specifically developed in various settings, such as binary classification and multi-class classification, to meet the distinct requirements of real-world applications.

In this thesis, we concentrate our research on Multi-Label Classification (MLC). Different form the traditional single-label classification, where an instance is relevant with one class label, MLC aims to solve the multi-label problems, where an instance probably belongs to multiple labels. Such a generalization greatly increases the difficulty of achieving a desirable classification accuracy at a tractable time cost. As an appealing and challenging supervised learning problem, MLC has a wide range of real-world applications, such as text categorization, semantic image annotation, bioinformatics analysis and music emotion detection.

In general, there are two main concerns on the MLC problems. First, label correlations are strong and ubiquitous in various multi-label datasets. For example, in semantic image annotation, the labels ”lake” and ”reflection” probably concur, and share a strong correlation. Thus, it is important and crucial to capture such label correlations in order to achieve a desirable classification performance. Second, as the rapid increase of web-related applications, more and more datasets emerge in high-dimensionality, whose number of instances, features and labels are far from the regular scale. For example, there are millions of videos in the video-sharing website Youtube, while each one can be tagged by some of millions of candidate categories. Such high-dimensionality of multi-label data significantly increases the time and space complexity in learning, and degrades the classification performance due to the curse of dimensionality. To address the two concerns, various MLC methods have been proposed in recent years, and achieved much success in a number of applications. However, further improvement in terms of time complexity and classification accuracy is recently demanding.

The research objective of this thesis is to improve the performance of MLC by capturing label correlations and reducing dimensionality. According to the objective, the thesis is separated into two major parts: Part I Multi-Label Classification and Part II Multi-Label Dimension Reduction.

In Part I, we focus on solving MLC problems by label correlation modeling and multi-label feature selection. Motivated by the Classifier Chains (CC) method, we propose the Polytree-Augmented Classifier Chains (PACC) in order to save label correlations in one probabilistic graphical model, the

polytree. Benefiting from polytree's flexible structure, the problems of error propagation and poorly ordered chain in CC can be avoided in PACC. To further improve its performance, a two-stage feature selection approach is developed by removing irrelevant and redundant features for each label. In addition, we reconsider both label correlation modeling and feature selection from a unified framework via conditional likelihood maximization. Using this approach, we show that existing CC-based methods and several feature selection approaches are special cases of our generic framework.

In Part II, we aim to improve the classification performance by decreasing the problem size of MLC. To reduce the dimensionality of features, we conduct Feature Space Dimension Reduction (FS-DR) by proposing two ML-DR methods, MLC with Meta-Label-Specific Features (MLSF) and Robust sEmi-supervised multi-lAbel DimEnsion Reduction (READER) via empirical risk minimization. Based on $\ell_{2,1}$-norm loss and regularization, READER performs feature selection in a robust manner through label embedding (label correlation modeling) and manifold learning (semi-supervised learning). To avoid the problem of imperfect label information, we conduct Label Space Dimension Reduction (LS-DR) by extending READER to apply nonlinear Label Embedding (READER-LE) with a linear approximation. Furthermore, in order to utilize parallel computing, for the first time we introduce a novel category for ML-DR, Instance Space Decomposition (ISD), and propose the Clustering-based Local MLC (CLMLC) method to evaluate its efficiency. Different with existing ISD methods, CLMLC conducts the feature-guided ISD in a feature subspace rather than the original feature space, and builds cluster-specific local models.

Based on extensive empirical evidences, our work in this thesis demonstrates proposed MLC methods successfully address the two concerns of MLC, and improve the classification performance compared with the state-of-the-art methods. Therefore, it is hopeful for researchers in the field of MLC to build their MLC systems and develop novel MLC methods on the basis of the research work in this thesis.