Pattern Recognition 74 (2018) 503–517

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

# Optimization of classifier chains via conditional likelihood maximization

Lu Sun*, Mineichi Kudo

*Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan*

ARTICLE INFO

*Article history:*
Received 12 June 2016
Revised 20 September 2017
Accepted 23 September 2017
Available online 25 September 2017

*Keywords:*
Multi-label classification
Classifier chains
Conditional likelihood maximization
*k*-dependence Bayesian network
Multi-label feature selection

ABSTRACT

Multi-label classification associates an unseen instance with multiple relevant labels. In recent years, a variety of methods have been proposed to handle the multi-label problems. Classifier chains is one of the most popular multi-label methods because of its efficiency and simplicity. In this paper, we consider to optimize classifier chains from the viewpoint of conditional likelihood maximization. In the proposed unified framework, classifier chains can be optimized in either or both of two aspects: label correlation modeling and multi-label feature selection. In this paper we show that previous classifier chains algorithms are specified in the unified framework. In addition, previous information theoretic multi-label feature selection algorithms are specified with different assumptions on the feature and label spaces. Based on these analyses, we propose a novel multi-label method, *k*-dependence classifier chains with label-specific features, and demonstrate the effectiveness of the method.



## 1. Introduction

Unlike traditional single-label classification where each instance is associated with only one label, Multi-Label Classification (MLC) refers to the problems assigning multiple labels to a single test instance. MLC can be seen in a wide range of real-world applications such as text categorization, semantic image classification, bioinformatics analysis and video annotation. In fact, MLC is ubiquitous in real-world problems. For example, a news article is possibly relevant to multiple topics, like “science”, “technology”, “economics”, “politics”, etc; a single image is probably associated with a set of semantic concepts, like “sky”, “sea”, “field”, “building”, etc.

To tackle such multi-label problems, various MLC methods have been proposed. The existing MLC methods fall into two broad categories: problem transformation and algorithm adaptation [1]. As a convenient and straightforward way for MLC, problem transformation strategy transforms an MLC problem into one or a set of single-label classification problems, and learns one or a family of classifiers for modeling the single-label memberships. Most of popular baseline MLC methods, such as Binary Relevance (BR) [2], Calibrated Label Ranking (CLR) [3], and Label Powerset (LP) [4], belong to this strategy. Algorithm adaptation strategy induces conventional machine learning algorithms in the multi-label settings. Various MLC methods adopting one of the above two strategies have been developed and succeeded in dealing with multi-label problems.

Classifier Chains (CC) [5] is one of the most promising MLC methods which follow the problem transformation strategy. Originated from BR which simply ignores label correlations, CC constructs a chain structure on labels and determines the presence/absence of the current label under the condition of previously determined labels. CC succeeds in modeling label correlations and achieves higher classification accuracy at similar computational expense with BR. Although CC-based methods have achieved much success in various applications [5–7], further improvement in classification accuracy is still required. Here we seek the possibility to improve CC in terms of two aspects: label correlation modeling and multi-label feature selection. The intuition behind this idea is that all of the previously determined labels are not always necessary for decision on the current label (necessity of limiting label correlations), and irrelevant and redundant features are usually harmful for the performance of CC (necessity of feature selection). In this paper, we propose a unified framework comprising of both label correlation modeling and multi-label feature selection via conditional likelihood maximization of MLC.

The contributions of this work are cast into three-folds. First, we propose a general framework taking label correlation modeling and multi-label feature selection into account via conditional likelihood maximization. Second, the *k*-dependence classifier chains method is proposed based on greedy iterative optimization of a sub-problem of likelihood maximization. Third, a general infor-

* Corresponding author.
*E-mail addresses:* sunlu@main.ist.hokudai.ac.jp (L. Sun), mine@main.ist.hokudai.ac.jp (M. Kudo).

https://doi.org/10.1016/j.patcog.2017.09.034
0031-3203/

mation theoretic feature selection method is proposed for MLC, where three terms on relevancy, redundancy and label correlations are considered for feature subset selection.

The rest of this paper is organized as follows. Section 2 discusses the related works, focusing mainly on CC-based methods and information theoretic feature selection. Section 3 illustrates the unified framework for MLC by conditional likelihood maximization, and induces two sub-problems: model selection and multi-label feature selection. Sections 4 and 5 present the solutions on model selection and multi-label feature selection, respectively. Section 6 summarizes several theoretical findings during the development of the proposed method. Section 7 discusses the implementation issues. Section 8 introduces the experiments, and reports the results. Finally, Section 9 concludes this paper and discusses the further research.

## 2. Related works

Previous efforts have been paid on MLC in terms of various viewpoints, such as label correlations modeling [4,5], loss function analysis [6,8,9], large-scale learning [10,11] and dimension reduction [12–14]. In this paper we concentrate mainly on two aspects: label correlations modeling and dimension reduction. It has been shown in a number of researches [3–5] that modeling label correlations is very crucial to perform accurate classification. On the other hand, various dimension reduction algorithms, including feature selection [15,16] and feature extraction [12,17], have been employed in MLC, in order to simplify the learning phase and overcome the curse of dimensionality.

In order to capture label correlations, Classifier Chains (CC) based methods [5–7] have been proposed at tractable computational complexity. CC-based methods originates from Binary Relevance (BR) [2], which simply decomposes a multi-label problem into a set of binary classification problems, totally ignoring label correlations. In this sense, BR is actually a hamming loss risk minimizer [8]. In CC [5], label correlations are expressed in an ordered chain of labels. In the learning phase, according to a predefined chain order, it builds a set of binary classifiers such that each classifier predicts the correct value of a target label by referring to the correct values of all the preceding labels in addition to the features. In the prediction phase, it predicts in turn the value of the target label using the previously estimated values of its parent labels as extra features. However, the performance of CC is sensitive to distinct chain orders, and it suffers from the problem of error propagation in the prediction phase. Several efforts have been paid to overcome the limitations of CC. Bayesian Classifier Chains (BCC) [7] introduces a directed tree as the probabilistic structure over labels. The directed tree is established by randomly choosing a label as its root and by assigning directions to the remaining edges. It shares the same model with CC, but restricts the number of parent labels no more than 1, which limits its expression ability on label correlations. Probabilistic Classifier Chains (PCC) [6] aims to solve the error propagation problem, providing better estimates than CC at the expense of higher processing time. Although PCC shares the learning model with CC, it chooses the best predictor by searching the *Maximum A Posterior* (MAP) assignment in an exhaustive manner. The exponential cost of PCC in prediction limits its application. To make the prediction tractable for PCC, PCC-beam [18] is proposed by applying beam search to find an approximate MAP assignment of labels to a test instance. MCC [19] utilizes the Monte Carlo scheme to find the sub-optimal chain order and perform efficient inference for the MAP assignment in the learning and prediction phase, respectively. In [20], the dynamic programming technique is used to search the globally optimal chain order of CC. In addition, to speed up the search procedure, a greedy approach is proposed to find locally optimal CC. In a recent work [21], the Classifier Trellis (CT) method is proposed for scalable MLC by extending the 1-dimensional chain of CC to a 2-dimensional trellis structure. CT saves label correlations in the trellis structure, where each label depends only on its adjacent labels. In this way, CT enables to limit the number of parent labels, and thus becomes scalable to the MLC problems with a large number of labels.

In terms of Feature Space Dimensionality Reduction (FS-DR), a variety of traditional supervised dimension reduction approaches have been specifically extended to match the setting of MLC. In [22], a supervised Multi-label Latent Semantic Indexing (MLSI) approach is developed to map the input features into a subspace by preserving the label information. By maximizing the feature-label dependence under the Hilbert-Schmidt independence criterion, Multi-label Dimension reduction via Dependence Maximization (MDDM) [12] derived a closed-form solution to efficiently find the projection into the feature subspace. In addition, several traditional dimension reduction techniques, such as Canonical Correlation Analysis (CCA) and Linear Discriminant Analysis (LDA), are proposed to handle the MLC problem [23,24]. In the field of feature selection, an information theoretic approach for Label Powerset (LP) has been developed in [25]. The method introduces a nearest neighbor estimator for computing mutual information, and applies pruned LP to control the problem size. The multivariate mutual information criterion is used in [26] to select useful features. Due to its computational inefficiency, an approximate solution is proposed to estimate the multivariate mutual information. In [15], the authors extend the feature selection framework in [27] to handle two MLC decomposition methods, BR and LP, which achieves significant improvement compared with several multi-label feature selection methods. In fact, all the methods mentioned above can be regarded as *global FS-DR methods*, since they attempt to find an identical feature subspace globally for all the labels. However, it is more reasonable to think that each label holds a specific supporting feature subset. To overcome the limitations, *local FS-DR methods* [16,17] have been proposed to find label-specific features. In [17], Label-specIfic FeaTures (LIFT) are extracted by conducting cluster analysis on the positive and negative instances of each label. The Learning Label-Specific Features (LLSF) method is proposed in [16]. LLSF selects label-specific features by optimizing the least squares problem with constraints of label correlations and feature sparsity.

## 3. The unified framework for MLC via likelihood maximization

### 3.1. Multi-label classification

In the scenario of MLC, an observation ($\mathbf{x}$, $\mathbf{y}$) consists of a $d$-dimensional feature vector $\mathbf{x}$ and a $q$-dimensional target label vector $\mathbf{y}$, drawn from the underlying random variables $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ and $\mathbf{Y} = (Y_1, \ldots, Y_q) \in \{0, 1\}^q$, respectively. For an observation of labels $\mathbf{y} = (y_1, \ldots, y_q)$, $y_j = 1$ if the $j$th label is relevant to the instance, and $y_j = 0$ otherwise, $j = 1, \ldots, q$.

The task of MLC is to find an optimal classifier $h : \mathbb{R}^d \to \{0, 1\}^q$, which assigns a label vector $\hat{\mathbf{y}} = h(\mathbf{x})$ to each instance $\mathbf{x}$ such that $h$ minimizes a loss function between $\hat{\mathbf{y}}$ and $\mathbf{y}$. For a loss function $L(\mathbf{Y}, h(\mathbf{X}))$, the optimal classifier $h^*$ is

$$h^* = \arg\min_h \mathbb{E}_{\mathbf{xy}} L(\mathbf{Y}, h(\mathbf{X})). \tag{1}$$

Specifically, given the subset 0–1 loss $L_s(\mathbf{y}, \hat{\mathbf{y}}) = \mathbb{1}_{\mathbf{y} \neq \hat{\mathbf{y}}}$, where $\mathbb{1}_{(\cdot)}$ denotes the indicator function, Eq. (1) can be rewritten in a pointwise way,

$$\hat{\mathbf{y}} = h^*(\mathbf{x}) = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}). \tag{2}$$

Here we use $p(\mathbf{Y}|\mathbf{X})$ to represent the conditional probability distribution of label variables $\mathbf{Y}$ given feature variables $\mathbf{X}$. According to