A Composite Likelihood View for Multi-Label Classification

Yi Zhang **Jeff Schneider**
School of Computer Science, Carnegie Mellon University

Abstract

Given limited training samples, learning to classify multiple labels is challenging. Problem decomposition [24] is widely used in this case, where the original problem is decomposed into a set of easier-to-learn subproblems, and predictions from subproblems are combined to make the final decision.

In this paper we show the connection between composite likelihoods [17] and many multi-label decomposition methods, e.g., one-vs-all, one-vs-one, calibrated label ranking, probabilistic classifier chain. This connection holds promise for improving problem decomposition in both the choice of subproblems and the combination of subproblem decisions.

As an attempt to exploit this connection, we design a composite marginal method that improves pairwise decomposition. Pairwise label comparisons, which seem to be a natural choice for subproblems, are replaced by bivariate label densities, which are more informative and natural components in a composite likelihood. For combining subproblem decisions, we propose a new mean-field approximation that minimizes the notion of composite divergence and is potentially more robust to inaccurate estimations in subproblems.

Empirical studies on five data sets show that, given limited training samples, the proposed method outperforms many alternatives.

1 Introduction

Multi-label classification has received considerable attention from the machine learning community [24]. In such problems, it is desirable to capture the (conditional) dependency of labels while keeping the algorithm both computationally and statistically efficient.

Many researchers focus on problem decomposition techniques, which usually decompose the multi-label classification problem into a set of simpler and easier-to-learn subproblems, estimate prediction models for the subproblems, and then combine the predictions from subproblems to make the final classification. Popular choices of subproblems include one-vs-all decomposition (i.e., the relevance of each individual label) [23], one-vs-one decomposition (i.e., the pairwise comparison between any two labels) [8, 12], a combination of both one-vs-all and one-vs-one decompositions [9], and conditional relevance of one label given other labels (e.g., classifier chain methods) [22, 6].

Despite having some empirical success, the choice of subproblems in many decomposition methods seems arbitrary and relies on intuition, and more importantly, the combination of subproblem predictions in the final decision making is usually based on heuristics (e.g., voting). To address these issues, we first show the connection between multi-label decomposition algorithms and composite likelihood [26, 16], a technique based on partial specification of the likelihood as the product of simple component likelihoods to efficiently model complex dependencies. We believe that this connection holds great promise for improving the design of multi-label decomposition methods, especially in the choice of subproblems and the combination of subproblem predictions in decision making.

As an attempt to exploit this connection, we design a composite marginal method that improves the popular pairwise decomposition approaches. For choosing subproblems, although pairwise label comparison is widely used and seems to be a natural choice, bivariate densities provide more information about pairwise label relation and are more natural as part of a composite likelihood. For combining subproblem predictions, we propose a new mean-field approximation procedure based on composite likelihood, which minimizes the notion of composite divergence and is potentially more robust to inaccurate estimations in subproblems.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

On five real-world data sets, we compare our method with several multi-label decomposition methods and a joint learning approach that captures label dependency using a graphical model. Empirical results show that, given limited training samples, the proposed method outperforms alternatives and provides superior predictions under a variety of evaluation criteria.

In the rest of this paper, we will first introduce composite likelihoods in Section 2, and establish the connection between multi-label decomposition methods and composite likelihood in Section 3. To exploit this connection, we propose a composite marginal method for multi-label classification in Section 4, and present empirical results in Section 5. We conclude in Section 6.

2 Composite Likelihood Methods

In this section we first give an overview on commonly used forms of composite likelihoods, and then we briefly motivate and discuss parameter estimation via composite likelihoods.

2.1 Overview

Composite likelihood is a partial specification of the full likelihood function by multiplying a set of simple component likelihoods, where each component likelihood is in the form of either a marginal or a conditional density. Composite likelihood can be viewed as an oversimplified form of the full likelihood, but such an approximation can provide certain benefits in parameter estimation, notably in computation, statistical efficiency (with limited samples), and robustness (to model specification). Research on composite likelihood can be dated back to Besag's pseudolikelihood [1] and Cox's partial likelihood [3]. Lindsay [17] formalizes the term composite likelihood for "product of likelihoods". Some excellent overviews and discussion on this subject have been recently published [26, 16].

Formally, consider q random variables $\mathbf{Y} = (Y_1, \ldots, Y_q)$, the parameter vector $\theta \in \Theta$ and the full likelihood function $L(\theta; \mathbf{y})$ with one observation $\mathbf{y}$. Following the notation in [17, 26], we denote by $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ a collection of K marginal or conditional events, with corresponding component likelihoods $\{L_k(\theta; \mathbf{y}) \propto f(\mathbf{y} \in \mathcal{A}_k; \theta)\}_{k=1}^K$. A composite likelihood approximates the likelihood function $L(\theta; \mathbf{y})$ by:

$$L_C(\theta; \mathbf{y}) = \prod_{k=1}^{K} L_k(\theta; \mathbf{y})^{w_k} \tag{1}$$

where $\{L_k(\theta; \mathbf{y})\}_{k=1}^K$ are marginal or conditional densities, and $\{w_k\}_{k=1}^K$ are (optional) nonnegative weights on components. In this paper we mainly consider $w_k = 1, k = 1, 2, \ldots, K$ for simplicity.

Composite marginal likelihoods is a large class of composite likelihoods that use low-dimensional marginal densities as component likelihoods. The simplest example of composite marginal likelihood is the so-called *independence likelihood*. For q variables $\mathbf{y} = (y_1, \ldots, y_q)$, we have:

$$L_{ind}(\theta; \mathbf{y}) = \prod_{i=1}^{q} f(y_i; \theta) \tag{2}$$

Naturally this specification does not capture interactions between different variables. Another well studied class of likelihoods, which considers small blocks of variables, is the *pairwise likelihood* [15, 5, 26, 16]:

$$L_{pair}(\theta; \mathbf{y}) = \prod_{i=1}^{q-1} \prod_{j=i+1}^{q} f(y_i, y_j; \theta) \tag{3}$$

which includes interactions between pairs of variables into the likelihood function.

Composite conditional likelihoods ia another popular class of composite likelihoods based on conditional densities, which date back to pseudolikelihood [1]. Its general form is the following:

$$L_{cond}(\theta; \mathbf{y}) = \prod_{k=1}^{K} f(y_{B_k} | y_{\mathcal{N}(B_k)}; \theta) \tag{4}$$

where each B_k indexes a block of variables, and $\mathcal{N}(B_k)$ indexes neighbors of the variables in B_k. Two widely used composite conditional likelihoods in both longitudinal studies [20] and bioinformatics [18] are *pairwise conditional likelihood*:

$$L_{pcl}(\theta; \mathbf{y}) \quad = \quad \prod_{i=1}^{q} \prod_{j=1}^{q} f(y_i | y_j; \theta) \tag{5}$$

and *full conditional likelihood*:

$$L_{fcl}(\theta; \mathbf{y}) \quad = \quad \prod_{i=1}^{q} f(y_i | y_{(-i)}; \theta) \tag{6}$$

where $y_{(-i)}$ denotes all variables but y_i.

2.2 Computation, robustness, and statistical efficiency of estimation

Composite likelihood provides an approximation to the full likelihood function that has been widely used in parameter estimation. Research on composite likelihood estimation has been focused on computational efficiency, robustness to model misspecification, and statistical efficiency. Parameter estimation via composite likelihoods is computationally more efficient than optimizing the joint likelihood function, especially when

parameters in component likelihoods are decoupled and estimated separately, e.g., as in [7, 20]. Also, it has been argued by many researchers that composite likelihood estimation is more robust to model misspecification since only model assumptions on lower dimensional conditional or marginal densities, instead of the detailed specification of full joint density, are required [26, 16]. Statistical efficiency of the estimation, as a result of reduced interactions among parameters, has also been intensively studied [19, 11, 26, 16].

3 A Composite Likelihood View of Multi-Label Problem Decomposition

In this section, we show the connection between composite likelihoods and many recent multi-label decomposition methods.

3.1 The Full Multi-Label Model

One way to capture the full dependency of q labels is to treat each of the 2^q label combinations as a separate class. This is usually called the *label powerset method* (LP) [24]. In the statistics literature, this method is linked to directly modeling the joint probability of all 2^q entries in a q-dimensional contingency table [4]. The drawback of this approach is obvious. Large amounts of observations are needed to estimate this full model well, and the computational complexity of the training algorithm usually scales exponentially in q. This is what motivates composite likelihoods: the full likelihood is expensive to estimate and perform inference.

3.2 Multi-Label Problem Decomposition and Composite Likelihoods

Multi-label decomposition methods transform the original problem into a set of subproblems, learn each subproblem, and combine subproblem predictions to make the final classification. In general, the choice of subproblems in a decomposition method corresponds to a certain instantiation of the composite likelihood:

$$L_C(\theta; \mathbf{y}|\mathbf{x}) = \prod_{k=1}^{K} L_k(\theta; \mathbf{y}|\mathbf{x}) \tag{7}$$

where $\mathbf{y} = (y_1, \ldots, y_q) \in \{0, 1\}^q$ denotes the label vector and $\mathbf{x}$ denotes the feature vector. Learning the subproblems corresponds to assuming the parameters inside different $L_k(\theta; \mathbf{y}|\mathbf{x})$ are independent and estimating them separately. The combination of subproblem predictions is usually based on heuristics or exhaustive search, without explicitly leveraging properties of the composite likelihood.

The simplest way to decompose multi-label classification is *one-vs-all decomposition*, or *binary relevance method* (BR), which has been empirically justified in the context of multi-class classification [23]. For q labels, we consider q subproblems, each to model the relevance of one label independently. This method corresponds to a composite likelihood of the form:

$$L_{BR}(\theta; \mathbf{y}|\mathbf{x}) = \prod_{i=1}^{q} f(y_i|\mathbf{x}; \theta_i) \tag{8}$$

where θ_i denotes the parameters used to model label i. For prediction, each subproblem provides the conditional probability of one label, and prediction can be done by thresholding (e.g., at 0.5) and there is no need to combine subproblem decisions. Since this method ignores the dependency among labels, suboptimal performance will be obtained if the evaluation criterion calls for capturing the label dependency [6]. For example, this method might predict each label reasonably well, but rarely get all the labels classified correctly.

Another popular strategy is *one-vs-one decomposition* [8], which is used in *pairwise label ranking* (PLR) [12]. For q labels, this method captures label dependency by formulating $\frac{q(q-1)}{2}$ subproblems, each learning a classifier to compare two labels. In this sense, the choice of subproblems in pairwise label ranking indicates the following composite likelihood:

$$L_{PLR}(\theta; \mathbf{y}|\mathbf{x}) = \prod_{i=1}^{q-1} \prod_{j=i+1}^{q} f(y_i \geq y_j|\mathbf{x}; \theta_{ij}) \tag{9}$$

where θ_{ij} is the vector of parameters for label pair (i, j), and the component form $f(y_i \geq y_j|\mathbf{x}; \theta_i)$ comes from the fact that in one-vs-one decomposition, subproblems are *pairwise comparisons*. For prediction, votes are collected from all pairwise classifiers, and labels are ranked by the number of wining votes they receive, e.g., the label receiving $q-1$ wining votes will be ranked first, as it wins all $q-1$ comparisons with other labels. Note that this method is only a label ranking method: a rank order does not give a classification, and one has to find a threshold for classification.

Calibrated Label Ranking (CLR) is a strategy that combines both one-vs-one and one-vs-all decompositions for multi-label classification [9]. The key idea is to introduce a virtual label that is always ranked between the relevant set and the irrelevant set of labels. As a result, all we need for classifying q labels is to obtain a rank order of $q + 1$ labels (q labels plus one virtual label) by pairwise label ranking. The difficulty, however, is how to learn the pairwise classifier between each actual label and the virtual label. In [9], this is solved by noticing that the pairwise comparison between label i and the virtual label is semantically

equivalent to the one-vs-all decision for label i. As CLR uses both one-vs-one and one-vs-all subproblems, it corresponds to the composite likelihood:

$$L_{CLR}(\theta;\mathbf{y}|\mathbf{x}) = [\prod_{i=1}^{q} f(y_i|\mathbf{x};\theta_i)]\cdot[\prod_{i=1}^{q-1}\prod_{j=i+1}^{q} f(y_i \geq y_j|\mathbf{x};\theta_{ij})] \quad (10)$$

Classifier chains (CC) [22] and probabilistic classifier chains (PCC) [6] use subproblems that describe the conditional relation of labels. Both methods start with a randomly chosen label order. CC learns a function to predict each label, conditional on the input features and antecedent labels along the chosen order. In PCC, each prediction function is probabilistic, and all functions together define the joint label probability as in a Bayesian network. In this sense, given a label order $((1),(2),\ldots,(q))$, CC and PCC are linked to:

$$L_{CC}(\theta;\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{q} f(y_{(i)}|\mathbf{x}, y_{(1)},\ldots,y_{(i-1)};\theta_{(i)}) \quad (11)$$

For classification using CC, labels are predicted in the chosen order, based on the input features and predicted antecedent labels. For classification using PCC, however, efficient inference is not provided and exhaustive search is used to find label assignments [6]. Indeed, if we consider (11) as a Bayesian network, the largest conditional table contains all the labels.

The label powerset method considers all the 2^q label combinations as 2^q classes. This corresponds to the full likelihood approach with 2^q configurations. Along this direction, the pruned labelset method considers only label combinations that are sufficiently frequent in the data [21], and the random k-labelsets focuses on label combinations that only involve k labels [25]. These methods generate multi-class classification problems with less than 2^q classes, and it corresponds to specifying a single component likelihood, but with less entries than in the full likelihood.

3.3 Other Multi-Label Classification Methods

All the decomposition methods discussed in Section 3.2 can be enhanced using randomization and ensemble learning. A committee of experts can be constructed by launching a given method multiple times with certain randomization, such as sampling training examples, label orders or label subsets. Ensemble-based extensions that have been specifically studied include ensemble classifier chains [22], ensemble probabilistic classifier chains [6], ensemble pruned labelsets [21] and random K-labelsets [25]. It is generally accepted that an ensemble version of an algorithm will outperform its original non-ensemble version, usually at the cost of more intensive computation.

Instead of decomposing and learning subproblems individually, another direction is to jointly model the dependency among all the labels and features via a single graphical model. Research along this direction includes specifying a conditionally trained undirected graphical model (i.e., CRF) [10] or learning the structure of Bayesian networks from data [29, 28] to capture the joint label relation (conditioned on features). Joint modeling via a single graphical model provides an elegant method of multi-label learning, but potential challenges include: 1) learning the graphical model is usually intractable (e.g., training an undirected graphical model, or learning the structure of a Bayesian network); 2) joint modeling is likely to demand more training samples than learning simpler subproblems.

Aa an alternative to decomposition methods which are model-independent, researchers have also been adapting specific models (such as decision trees, SVMs, KNNs, neural networks, boosting) internally to produce multi-label predictions.

An recent overview on a variety of multi-label learning methods can be found in [24].

4 A Composite Marginal Model for Multi-Label Classification

Exploiting the composite likelihood view of problem decomposition, we propose a composite marginal method for multi-label classification. We consider the subproblems of estimating the univariate and bivariate label densities as components in a composite likelihood, and based on this composite likelihood, we develop a new mean-field approximation procedure that minimizes the notion of composite divergence for inference and combining subproblem decisions, which is more robust to certain common estimation and prediction errors in multi-label classification.

4.1 Implications of Composite Likelihood to Multi-Label Problem Decomposition

Subproblem design. A primary implication of the composite likelihood perspective to multi-label classification is on the design of subproblems. For example, consider the composite marginal likelihood of pairwise label ranking in (9) and that of calibrated label ranking in (10). Both composite likelihoods contain *pairwise comparison* events $\{f(y_i \geq y_j|\mathbf{x};\theta_{ij})\}$ as component likelihoods. However, is the component likelihood in the form of $f(y_i \geq y_j|\mathbf{x};\theta_{ij})$ the most informative description of the pairwise label relationship? This issue will be addressed in Section 4.2.

Combination of subproblem decisions. Combining subproblem predictions to make a final classification is critical to multi-label decomposition methods. However, most existing decomposition methods combine subproblem decisions using heuristics such as voting. The composite likelihood function approximates the joint label probability in a composite form of subproblem predictions, and this form enables the design of efficient inference procedures that are particularly suitable for multi-label classification. In Section 4.3, we develop a new robust mean-field approximation by exploiting the composite form of the likelihood.

4.2 Composite Marginal Modeling: Specification and Estimation

As raised in Section 4.1, one issue in (10) is the use of likelihood components of the form $\{f(y_i \geq y_j|\mathbf{x};\theta_{ij})\}$. Although pairwise comparison between two labels is very natural subproblem to consider from the algorithm perspective, it is not the most informative component in a likelihood function. Knowing "$y_i \geq y_j$" only rules out one scenario $(y_i = 0, y_j = 1)$ and still leaves three possibilities: $(y_i = 0, y_j = 0)$, $(y_i = 1, y_j = 0)$, or $(y_i = 1, y_j = 1)$. In this sense, the bivariate marginal density fully describes the pairwise relation of two labels and is thus provides more information to the composite likelihood. Also, bivariate densities of the form $f(y_i, y_j|\mathbf{x};\theta_{ij})$ is also more natural than $\{f(y_i \geq y_j|\mathbf{x};\theta_{ij})\}$ when considered as part of a likelihood function. Therefore, we consider univariate and bivariate density estimation as subproblems, which give the following composite likelihood:

$$L_{CMM}(\theta;\mathbf{y}|\mathbf{x}) = [\prod_{i=1}^{q} f(y_i|\mathbf{x};\theta_i)]\cdot[\prod_{i=1}^{q-1}\prod_{j=i+1}^{q} f(y_i,y_j|\mathbf{x};\theta_{ij})]^{\lambda} \tag{12}$$

where $\{f(y_i|\mathbf{x};\theta_i)\}$ and $\{f(y_i,y_j|\mathbf{x};\theta_{ij})\}$ are univariate and bivariate marginal densities of labels conditional on the feature vector $\mathbf{x}$, $\{\theta_i\}$ and $\{\theta_{ij}\}$ are two sets of parameter vectors for univariate and bivariate densities, and λ is the nonnegative weight of bivariate densities and we set $\lambda = 1$ in this paper.

Given a set of N training examples $\mathbf{D} = \{\mathbf{y}^n, \mathbf{x}^n\}_{n=1}^N$, the overall composite likelihood function is:

$$L_{CMM}(\theta;\mathbf{D}) = \prod_{n=1}^{N} L_{CMM}(\theta;\mathbf{y}^n|\mathbf{x}^n) \tag{13}$$

Since we assume parameters in different components are independent, maximum composite likelihood estimation (MCLE) can be performed separately:

$$\hat{\theta}_i = \operatorname*{argmax}_{\theta_i} \prod_{n=1}^{N} f(y_i^n|\mathbf{x}^n;\theta_i), \quad i = 1,2,\ldots,q \tag{14}$$

$$\hat{\theta}_{ij} = \operatorname*{argmax}_{\theta_{ij}} \prod_{n=1}^{N} f(y_i^n, y_j^n|\mathbf{x}^n;\theta_{ij}), \quad i = 1,2,\ldots,q-1 \text{ and } j = i+1,\ldots,q \tag{15}$$

Note that problem (14) is essentially to estimate a probabilistic binary classifier, and problem (15) is to estimate a probabilistic 4-class classifier (for the four possible configurations of two labels).

4.3 Composite Marginal Modeling: Inference via Robust Mean-Field Approximation

Given $(\{\hat{\theta}_i\},\{\hat{\theta}_{ij}\})$, the composite likelihood provides a joint conditional probability of labels:

$$\hat{P}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\cdot[\prod_{i=1}^{q} f(y_i|\mathbf{x};\hat{\theta}_i)]\cdot[\prod_{i=1}^{q-1}\prod_{j=i+1}^{q} f(y_i,y_j|\mathbf{x};\hat{\theta}_{ij})]^{\lambda} \tag{16}$$

where $f(y_i|\mathbf{x};\hat{\theta}_i)$ and $f(y_i,y_j|\mathbf{x};\hat{\theta}_{ij})$ are discrete potentials, and Z is the partition function.

For a testing example $\mathbf{x}$, exact inference on $\mathbf{y} \in \{0,1\}^q$ using $\hat{P}(\mathbf{y}|\mathbf{x})$ has a time complexity exponential in q. For efficient classification, we may consider the classical mean-field approximation [13, 14]:

$$Q(\mathbf{y}) = \prod_{i=1}^{q} Q_i(y_i) \tag{17}$$

which is the fully factorized distribution on $\mathbf{y}$, with each $Q_i(y_i)$ a Bernoulli distribution on label y_i. Traditionally, we will use fixed point equations to minimize the following KL divergence [14]:

$$\hat{Q} = \operatorname*{argmin}_{Q} KL(Q||\hat{P}) = \operatorname*{argmin}_{Q} \sum_{\mathbf{y}\in\{0,1\}^q} Q(\mathbf{y}) \log \frac{Q(\mathbf{y})}{\hat{P}(\mathbf{y}|\mathbf{x})} \tag{18}$$

However, $KL(Q||\hat{P})$ is **highly sensitive** to a specific type of estimation error that happens **frequently** in multi-label classification: underestimating the probability of rare label combinations. When label combination $\mathbf{y}$ are rare, maximum likelihood estimation as in (14) and (15) will produce model parameters that further underestimate $\hat{P}(\mathbf{y}|\mathbf{x})$, thus pushing $\hat{P}(\mathbf{y}|\mathbf{x})$ to zero. This is known as the imbalance class problem [2]. Unfortunately, due to the use of Q as the base distribution, $KL(Q||\hat{P})$ is sensitive to regions where Q is non-zero but $\hat{P}$ is close to zero, which is also evident in (18) as $\hat{P}$ appears in the denominator. Thus, minimizing $KL(Q||\hat{P})$ will react dramatically to underestimation errors in $\hat{P}$, i.e., Q is forced to be zero wherever $\hat{P}$ approaches zero.

As a result, we need a new divergence measure $D(Q||\hat{P})$ which, as $KL(Q||\hat{P})$, can be optimized efficiently by fixed point equations (to reach a stationary point), and more importantly, $D(Q||\hat{P})$ needs to be robust to underestimation errors in $\hat{P}$. A convenient way to define a divergence measure for a composite likelihood is to use *composite divergence* [26], which is the linear weighted combination of divergences to all the component distributions. Since $\hat{P}$ is the normalized L_{CMM} in (12), we propose the following divergence:

$$D(Q||\hat{P}) = \sum_{i=1}^{q} D(Q||f(\cdot|\mathbf{x};\hat{\theta}_i)) + \lambda \sum_{i=1}^{q-1}\sum_{j=i+1}^{q} D(Q||f(\cdot,\cdot|\mathbf{x};\hat{\theta}_{ij})) \quad (19)$$

in which the divergence between Q and each univariate discrete distribution $f(\cdot|\mathbf{x};\hat{\theta}_i)$ is defined as:

$$D(Q||f(\cdot|\mathbf{x};\hat{\theta}_i)) = \sum_{y_i=0}^{1} (Q_i(y_i) - f(y_i|\mathbf{x};\hat{\theta}_i))^2 \quad (20)$$

and the divergence between Q and each bivariate discrete distribution $f(\cdot,\cdot|\mathbf{x};\hat{\theta}_{ij})$ is:

$$D(Q||f(\cdot,\cdot|\mathbf{x};\hat{\theta}_{ij})) = \sum_{y_i=0}^{1}\sum_{y_j=0}^{1} (Q_i(y_i)Q_j(y_j) - f(y_i,y_j|\mathbf{x};\hat{\theta}_{ij}))^2 \quad (21)$$

which are the sum of squared differences of the probability mass over possible events in $f(\cdot|\mathbf{x};\hat{\theta}_i)$ and $f(\cdot,\cdot|\mathbf{x};\hat{\theta}_{ij})$, respectively. Note that Q reduces to Q_i and $Q_i \cdot Q_j$ in the two equations as Q is fully factorized. Compared to the KL divergence in (18), the new divergences in (20) and (21) are more robust to the underestimation cases $f(y_i|\mathbf{x};\hat{\theta}_i) \to 0$ and $f(y_i,y_j|\mathbf{x};\hat{\theta}_{ij}) \to 0$, and thus so is the linear combination $D(Q||\hat{P})$ in (19).

Minimizing $D(Q||\hat{P})$ in (19) w.r.t. a single Bernoulli Q_i with all other $\{Q_j\}_{j\neq i}$ fixed can be solved in closed form. Thus, we can iteratively apply fixed point equations and converge quickly to a stationary point.

5 Empirical Studies

Data. We perform our experiments on five real-world multi-label data sets: Enron, Medical, Yeast, Scene, and Emotion[1]. Enron and Medical are text data with labels related to email analysis and medical decision. Yeast is a biological data set where genes are labeled by functions. Scene is an image data set where labels denote different scenes. Emotion is a collections of songs labeled by emotions. We select the top ten labels if the data set has more than ten labels.

[1]http://mulan.sourceforge.net/

Methods. We compare our proposed method to several popular multi-label decomposition methods as well as a joint modeling approach that captures the label dependency using graphical models [29]. See Section 3 for a review of decomposition methods.

- *Binary relevance (BR)*: learn to classify each label independently (i.e., one-vs-all decomposition).
- *Pairwise label ranking (PLR)*: perform pairwise label comparison to rank all the labels.
- *Calibrated label ranking (CLR)*: combine both pairwise comparison and one-vs-all classifiers.
- *Classifier chain (CC)*: a chain of classifiers to model conditional label relation given a label order.
- *Random k-labelsets (RK)*: an ensemble algorithm combines classifiers for label sets of size K.
- *Multi-label learning by exploiting label dependency (LEAD)*: a joint modeling approach which learns and incorporates a Bayesian network into multi-label classification to capture label dependency.
- *Composite marginal model (CMM)*: the proposed method described in Section 4, which exploits the composite likelihood view to improve subproblem choice and prediction combination.

All the methods need base learners to solve the subproblems. We use linear SVMs for binary decisions, multi-class SVMs (based on one-vs-one decomposition) for multi-class cases. If probabilistic estimates are required for binary problems, logistic regression is used instead of SVMs, and for multi-class probabilistic estimates, we use the method in [27]. All these learners are available in the LIBSVM[2] and LIBLINEAR packages[3]. Structure learning of Bayesian networks is performed using Bayesian Net Toolbox (BNT)[4] and its structure learning package extension (BNT-SLP)[5].

All problem decomposition methods can be extended to randomized ensemble versions by sampling examples, label orders or label sets. However, in this paper we limit our attention to the methods themselves (except RK, which is already the ensemble version). Random k-labelsets is tested in its ensemble form (of size 30) since the raw version is not a complete multi-label classifier. All methods except CLR need a threshold to assign 0/1 values to labels. We simply use 0.5. One can also search for an optimal threshold in $[0, 1]$. The regularization parameters of the base learners are chosen by cross validation. We use K=2 for RK since most rivals focus on pairwise label relations.

Evaluation. We report on four evaluation measures.

[2]http://www.csie.ntu.edu.tw/ cjlin/libsvm/
[3]http://www.csie.ntu.edu.tw/ cjlin/liblinear/
[4]http://code.google.com/p/bnt/
[5]http://ofrancois.tuxfamily.org/slp.html

Table 1: Results on Medical data: means (and standard errors) over 30 random runs. The best model is marked with *, and all competitive models (by paired t-tests with the best model, $\alpha = 0.05$) are shown in **bold**.

Methods	Hamming Loss	Subset 0/1 Loss	Ranking Loss	One Error
BR	0.0325 (0.0007)	0.2608 (0.0048)	0.3627 (0.0075)	0.0787 (0.0033)
PLR	0.3721 (0.0011)	1.0000 (0.0000)	0.3678 (0.0050)	0.0978 (0.0029)
CLR	0.0338 (0.0006)	0.2708 (0.0043)	0.3687 (0.0054)	0.0888 (0.0025)
CC	0.0358 (0.0007)	0.2707 (0.0061)	0.3888 (0.0018)	0.1606 (0.0047)
RK	0.0329 (0.0007)	0.2610 (0.0055)	0.3863 (0.0016)	0.1109 (0.0030)
LEAD	0.0331 (0.0006)	**0.2563** (0.0046)	**0.3400** (0.0113)	0.0770 (0.0029)
CMM	**0.0313***(0.0005)	**0.2524***(0.0043)	**0.3276***(0.0082)	**0.0706***(0.0027)

Table 2: Results on Yeast data: means (and standard errors) over 30 random runs. The best model is marked with *, and all competitive models (by paired t-tests with the best model, $\alpha = 0.05$) are shown in **bold**.

Methods	Hamming Loss	Subset 0/1 Loss	Ranking Loss	One Error
BR	0.3137 (0.0020)	0.9140 (0.0033)	0.4531 (0.0017)	0.4823 (0.0058)
PLR	0.3050 (0.0011)	0.9588 (0.0026)	**0.3833***(0.0009)	**0.2401***(0.0008)
CLR	0.2963 (0.0014)	0.9147 (0.0042)	**0.3839** (0.0008)	**0.2412** (0.0008)
CC	0.3225 (0.0022)	**0.8598** (0.0034)	0.4290 (0.0013)	0.2452 (0.0013)
RK	0.2664 (0.0008)	0.8631 (0.0053)	0.4097 (0.0031)	0.2711 (0.0075)
LEAD	0.2684 (0.0010)	0.8811 (0.0051)	0.3886 (0.0007)	0.2413 (0.0010)
CMM	**0.2651***(0.0009)	**0.8594***(0.0043)	0.3916 (0.0008)	**0.2418** (0.0012)

Table 3: Results on Emotion data: means (and standard errors) over 30 random runs. The best model is marked with *, and all competitive models (by paired t-tests with the best model, $\alpha = 0.05$) are shown in **bold**.

Methods	Hamming Loss	Subset 0/1 Loss	Ranking Loss	One Error
BR	0.2354 (0.0027)	0.7863 (0.0045)	0.4937 (0.0030)	0.3394 (0.0074)
PLR	0.2681 (0.0018)	0.8645 (0.0021)	0.4859 (0.0028)	0.2954 (0.0042)
CLR	0.2294 (0.0024)	0.7810 (0.0057)	0.4877 (0.0019)	**0.2914** (0.0049)
CC	0.2431 (0.0033)	0.7818 (0.0067)	0.5134 (0.0026)	0.4285 (0.0069)
RK	**0.2155***(0.0019)	**0.7465***(0.0051)	0.5238 (0.0029)	0.3109 (0.0066)
LEAD	0.2600 (0.0025)	0.8713 (0.0052)	0.4845 (0.0034)	0.3281 (0.0060)
CMM	**0.2163** (0.0018)	**0.7599** (0.0044)	**0.4787***(0.0018)	**0.2884***(0.0044)

1) Hamming loss: the percentage of misclassified labels; 2) subset 0-1 loss: the percentage of examples that at least one label is misclassified, which tests the ability to capture label dependency; 3) ranking loss: the probability that an irrelevant label is ranked higher than a relevant label, which measures the ability to capture the relative order between labels. 4) One error: the percentage of examples for which the top ranked label turns out to be irrelevant.

Experimental settings. We perform 30 random runs and report means and standard errors of each evaluation measure. Data sets come with more than enough training examples (several of them contain thousands of training samples). As we are interested in performance with limited training data, in each random run we sample 200 training examples.

Results on five data sets are shown in Table 5 to Table 3. We summarize the results as follows:

- PLR and CLR perform well on ranking loss and one error, because their decisions are mainly based on pairwise comparison of labels and focus more on obtaining the correct rank order of labels. PLR has terrible performance on hamming loss and subset 0/1 loss because PLR is designed as a ranking algorithm and is incapable of deciding the relevance/irrelance of labels. CLR is designed to improve PLR by introducing a virtual label (which serves as an adaptive threshold between relevant and irrelevant labels) and thus CLR obtains average performance on hamming loss and subset 0/1 loss.
- CC attains average performance on hamming loss and subset 0/1 loss, but falls behind on ranking loss and one error. The conditional relation of labels captures certain label dependencies, but is not well

Table 4: Results on Scene data: means (and standard errors) over 30 random runs. The best model is marked with *****, and all competitive models (by paired t-tests with the best model, $\alpha = 0.05$) are shown in **bold**.

Methods	Hamming Loss	Subset 0/1 Loss	Ranking Loss	One Error
BR	0.1439 (0.0015)	0.6118 (0.0039)	0.5072 (0.0058)	0.3669 (0.0040)
PLR	0.3356 (0.0009)	0.9968 (0.0005)	**0.4753** (0.0053)	0.3265 (0.0062)
CLR	0.1397 (0.0012)	0.6181 (0.0046)	**0.4733***(0.0057)	0.3237 (0.0052)
CC	0.1491 (0.0015)	0.5410 (0.0053)	0.5030 (0.0030)	0.4162 (0.0039)
RK	0.1223 (0.0010)	0.5724 (0.0063)	0.5057 (0.0038)	0.3339 (0.0039)
LEAD	0.1291 (0.0011)	0.5739 (0.0048)	**0.4749** (0.0064)	0.3494 (0.0040)
CMM	**0.1138***(0.0008)	**0.5250***(0.0037)	0.4862 (0.0057)	**0.2872***(0.0027)

Table 5: Results on Enron data: means (and standard errors) over 30 random runs. The best model is marked with *****, and all competitive models (by paired t-tests with the best model, $\alpha = 0.05$) are shown in **bold**.

Methods	Hamming Loss	Subset 0/1 Loss	Ranking Loss	One Error
BR	0.1958 (0.0016)	0.8653 (0.0067)	0.2911 (0.0025)	0.2753 (0.0043)
PLR	0.3121 (0.0016)	0.9921 (0.0005)	0.3058 (0.0039)	**0.2356** (0.0024)
CLR	0.1897 (0.0015)	0.8505 (0.0053)	0.3026 (0.0033)	**0.2341***(0.0030)
CC	0.1960 (0.0019)	0.8417 (0.0046)	0.3730 (0.0038)	0.2622 (0.0042)
RK	**0.1811** (0.0007)	**0.8269***(0.0022)	0.3554 (0.0031)	0.2945 (0.0076)
LEAD	0.1932 (0.0007)	0.8614 (0.0036)	**0.2887** (0.0028)	0.2898 (0.0038)
CMM	**0.1796***(0.0008)	**0.8269***(0.0026)	**0.2854***(0.0020)	0.2450 (0.0036)

suited for ranking labels. It has been reported that ensemble CC has better performance, and interesting future work is to compare ensemble CC with ensemble versions of other methods.

- RK: as the only method coupled with randomized ensemble technique, RK achieves decent performance on hamming and subset 0/1 loss but below-average scores on ranking loss and one error. One potential reason for the unsatisfactory ranking performance is the use of simple voting to combine predictions from subproblems, which may lose information that is critical for accurate ranking.
- LEAD offers above-average performance on Medical and Yeast data sets but is not very competitive on other data sets. Jointly modeling the label dependency using a single graphical model is promising, but it also tends to require more training samples than learning simple subproblems in decomposition methods. Also, structure learning of Bayesian nets (or parameter estimation in undirected graphical models) are still computationally intractable.
- CMM delivers the top performance on all four evaluation criteria (in terms of the number of wining data sets). This is the only method competitive for both classification and ranking. The subproblems of estimating univariate and bivariate label densities conveys more information than pairwise label comparisons, and the robust mean-field procedure provides a decent approximation to the composite likelihood without losing critical information for classification and ranking.

6 Conclusion and Future Work

In this paper we show the connection between multi-label decomposition methods and composite likelihoods. This connection holds great promise for improving the design of multi-label problem decomposition in both the choice of subproblems and the combination of subproblem decisions. As an attempt to exploit this connection, we design a composite marginal method that improves pairwise decomposition. Pairwise label comparisons are replaced by bivariate density estimation, which offers more informative and natural components in the composite likelihood. For combining subproblem decisions, we propose a new mean-field approximation that minimizes the notion of composite divergence and is potentially more robust to inaccurate estimations in subproblems. Empirical studies show that the proposed method outperforms many alternatives under a variety of evaluation criteria.

In the future work, we will explore other forms of composite likelihoods in the context of multi-label classification. For example, the composite likelihood for classifier chains in (11) requires a given label order and does not permit efficient inference. Composite conditional likelihoods such as (5), on the other hand, are nice alternatives since they allow for efficient (approximate) inference and do not require a label order.

References

[1] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B*, 36:192–236, 1974.

[2] N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6:1–6, 2004.

[3] D. Cox and N. Reid. Partial likelihood. *Biometrika*, 62:269–276, 1975.

[4] D. R. Cox. The analysis of multivariate binary data. *J. Roy. Statist. Soc. Ser. C*, 21:113–120, 1972.

[5] D. R. Cox. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 92:729–737, 2004.

[6] K. Dembczyński, W. Cheng, and E. Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 279–286. Omnipress, June 2010.

[7] S. Fieuws and G. Verbeke. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2):424–31, 2006.

[8] J. Fürnkranz. Round robin classification. *J. Mach. Learn. Res.*, 2:721–747, March 2002.

[9] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, June 2008.

[10] N. Ghamrawi and A. McCallum. Collective multilabel classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 195–200, New York, NY, USA, 2005. ACM.

[11] N. Hjort and C. Varin. Ml, pl, ql in markov chain models. *Scandinavian Journal of Statistics*, 35:64–82, 2008.

[12] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker. Label ranking by learning pairwise preferences. *Artif. Intell.*, 172(16-17):1897–1916, 2008.

[13] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[14] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[15] A. Kuk. A pairwise likelihood approach to analyzing correlated binary data. *Statistics and Probability Letters*, 47:329–335, 2000.

[16] B. G. Lindsa, G. Y. Yi, and J. Sun. Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21:71–105, 2011.

[17] B. G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:221–239, 1988.

[18] K. V. Mardia, G. Hughes, C. C. Taylor, and H. Singh. A multivariate von mises distribution with applications to bioinformatics. *Canadian Journal of Statistics-revue Canadienne De Statistique*, 36:99–109, 2008.

[19] K. V. Mardia, J. T. Kent, G. Hughes, and C. C. Taylor. Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika*, 96(4):975–982, 2009.

[20] G. Molenberghs and G. Verbeke. *Models for Discrete Longitudinal Data*. Springer, New York, 2005.

[21] J. Read, B. Pfahringer, and G. Holmes. Multi-label classification using ensembles of pruned sets. In *IEEE International Conference on Data Mining*, pages 995–1000, 2008.

[22] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. In *Principles of Data Mining and Knowledge Discovery*, pages 254–269, 2009.

[23] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *J. Mach. Learn. Res.*, 5:101–141, December 2004.

[24] G. Tsoumakas, I. Katakis, and I. P. Vlahavas. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685. 2010.

[25] G. Tsoumakas and I. Vlahavas. Random k-labelsets: an ensemble method for multilabel classification. In *ECML '07: Proceedings of the 18th European conference on Machine Learning*, pages 406–417, Berlin, Heidelberg, 2007.

[26] C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

[27] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, pages 975–1005, 2004.

[28] J. H. Zaragoza, L. E. Sucar, E. F. Morales, C. Bielza, and P. Larrañaga. Bayesian chain classifiers for multidimensional classification. In *IJCAI*, pages 2192–2197, 2011.

[29] M.-L. Zhang and K. Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 999–1008, New York, NY, USA, 2010. ACM.