24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Learning classifier chains using matrix regularization: application to multimorbidity prediction

Paweł Teisseyre [1]

Abstract. We study a problem of learning classifier chains in multi-label classification with a special focus on feature selection. It turns out that standard classifier chains tend to select too many features, when feature selection method is embedded in base learner, which is due to the fact that selection is performed separately for each of the models in the chain. This can be a serious limitation in domains where the acquisition of feature values is costly or where including too many features (e.g. diagnostic tests) is associated with negative effects. We propose a novel method parCC (parsimonious classifier chains) that controls the total number of features without significant deterioration in the quality of the prediction. In the proposed method we jointly learn all models in the chain by combining $\ell_{2,1}$ regularization to select features shared across the models and ℓ_1 regularization to select relevant labels in each model. In theoretical analysis we provide a bound on generalization error for the algorithm using Rademacher complexity. We apply our method to predict multimorbidity (co-occurrence of multiple diseases in one patient) using various medical diagnostic tests. The experiments carried out on a large clinical database (MIMIC III) show that parCC achieves higher accuracy than related methods when the number of features is limited. We also demonstrate the efficacy of the proposed method on a set of standard benchmark datasets.

1 Introduction

Multi-label classification (MLC) is one of the most intensively studied problems in machine learning. This is motivated by a large number of new applications, including medicine, genetics, text categorization, image and video annotation among others, see e.g. [9]. In MLC each object of interest (e.g. patient, text, image) is described by a vector of features and a vector of target variables (labels). The main objective is to learn a model which predicts labels for previously unseen instance using the feature vector. In this work we focus on interesting application of multi-label learning, that is predicting multiple diseases in one patient based on various features, such as results of diagnostic tests or administrative information.

Classifier chains (CC) introduced by [25] are among the most popular and successful approaches in MLC. The basic idea of CC is to transform the multi-label learning problem into a chain of single-label problems, where subsequent models in the chain are built using the predictions of preceding ones. Classifier chains have several advantages. First, unlike in Binary Relevance, the dependencies between labels are exploited, which improves classification performance [5]. Secondly, the inference (prediction for new instances) is straightforward and does not require significant computational effort. This feature distinguishes Classifier Chains from many other approaches, e.g. Conditional Dependency Networks [11], in which prediction is computationally expensive (usually Gibbs Sampling is used for inference in CDN). Finally, they are generic in a sense that every single-label classifier can be used as a base learner. Moreover it is possible to use different models for different tasks in the chain and even mix regression and classification problems. Such general scenario is considered in this work.

Despite many advantages, Classifier Chains also have limitations. In this paper we deal with a significant limitation of CC, that is the lack of control of the number of features used in the model. It turns out that, when some feature selection procedure is embedded in the base learner, CC-based methods tend to select too many features. This is due to the fact that the models in the chain are trained independently and thus feature selection is performed separately for each of the models in the chain. Correlations between features which are relevant in different models in the chain are not taken into account. This may be a serious problem when one wants to fit the model under a constraint on the number of features. Such constraints are important in domains where the acquisition of the feature values is costly, e.g. in medical diagnosis where each diagnostic test is associated with its cost. In such cases it may be better to have a model with an acceptable classification performance, but a much lower cost. The issue is important, as a considerable proportion of healthcare costs in hospitals is spent on such tests and thus it is important to reduce the number of unnecessary expensive diagnostic tests in order to reduce the total cost of medical care [29]. Moreover, unnecessary diagnostic tests or treatments may cause negative effects and even increase risk of death; examples include treatments under general anesthesia [19] or diagnostic X-rays [13].

To overcome the above drawback of CC-based algorithms we propose a novel method, called parCC (parsimonious classifier chains), of learning classifier chains that allows to significantly reduce the total number of features without deterioration in the quality of the prediction. In the method we jointly learn all models in the chain. We combine $\ell_{2,1}$ regularization to select features shared across the models and ℓ_1 regularization to select relevant labels in each model. So the proposed approach aims to integrate the strengths of common prediction structure based methods and classifier chains. We provide a bound on generalization error for the algorithm using Rademacher complexity. We compare the proposed method with the related ones, e.g. Adaptive Classifier Chains (ACC) proposed recently in [28].

A natural application of the proposed method is predicting multimorbidity, which is a co-occurrence of two or more diseases in one patient. The goal is to predict occurrences of diseases (labels) using various features, mainly results of diagnostic tests (see Section

[1] Institute of Computer Science, Polish Academy of Sciences, Poland, e-mail: teisseyrep@ipipan.waw.pl

'Experiments' for examples). Many diagnostic tests may be redundant to predict the diseases or are associated with costs and risk of negative effects. Therefore, the total admissible number of features should be limited and one should learn the model under the constraint on the number of features. We perform experiments on a large clinical database MIMIC III, containing more than 30000 patients and more than 300 features. They show that parCC achieves higher accuracy than the related methods when the total number of features is limited. This is confirmed by the experiments on standard benchmark datasets.

The paper is organized as follows. We first introduce the related work and present a general framework for learning classifier chains. Next, the proposed learning method is presented as well as the most similar methods. We then report the experimental results and conclude the paper. Supplement (`https://home.ipipan.waw.pl/p.teisseyre/PUBLICATIONS/parcc/parcc_supplement.pdf`) contains results of additional experiments and information about datasets.

2 Related work

Multi-label classification has received increasing attention in recent years, due to its wide applications in practice. There is a rich body of work on multi-label learning in the literature, see comprehensive reviews [9, 30]. We provide a review to the most related methods in this section. Classifier Chains (CC) [25] are among the most promising MLC methods which use the problem transformation strategy. CC-based methods originates from Binary Relevance (BR), which simply decomposes a multi-label problem into a set of binary classification problems, ignoring dependencies between labels. CC overcomes the label independence assumption of BR, by augmenting the feature space with labels from the previous models. There are some challenges associated with CC-based methods that can be divided into three groups: (1) designing optimal inference strategy [4, 18], (2) optimizing order of the chain [24, 20] and (3) reducing dimensionality of the feature/label space. In recent years, several modifications of the Classifier Chains have been proposed that address the above problems.

In this work, we focus on issue (3), i.e. dimensionality reduction of the feature/label space. Several approaches have been proposed to select relevant features in MLC (see the comprehensive reviews [23, 17]), but they are usually not associated directly with classifier chains. CC-specific feature selection method was proposed in [21], who use information-theoretic measures to select features in CC and in [27], who uses ℓ_1 regularization to select features in subsequent models in the chain. However in the latter approach feature selection is performed separately, for each of the models in the chain. As a result, the total number of selected features is too large. The most related method to our approach is ACC (Adaptive Classifier Chains), proposed recently by [28] in the context of cost-sensitive learning. ACC is inspired by the adaptive lasso [31] designed for single-label classification, where adaptive weights are used for penalizing different coefficients in the ℓ_1 penalty. In adaptive lasso the penalties are based on some preliminary calculation of the coefficients in linear or logistic regression (e.g. using ordinary least squares). In contrast to adaptive lasso, in ACC penalties are calculated based on the previous models in the chain (see Section 5 for description).

Here we propose a novel method that finds a more parsimonious feature representation. The method combines $\ell_{2,1}$ and ℓ_1 regularization terms. Using $\ell_{2,1}$ norm ensures that the selected features will be shared across the models in the chain. Such regularization has been effectively exploited in simultaneous multi-task learning problems [1] and also in multi-label classification [14, 22]. Different from these works which consider only common input features and ignore label dependencies, in our approach we additionally consider regularization term corresponding to label vectors. Various forms of joint sparsity regularizations are considered for Conditional Dependency Network (CDN), see e.g. [12]. Note however than in CDN, the inference for new instance is much more difficult than for CC method, which is a drawback of this method in the case of large-scale problems.

Applying multi-label methods for multimorbidity prediction has been investigated in [32] and [28]. In both studies, the previous version of MIMIC database (MIMIC II) has been used. Zufferey et al. [32] compare several multi-label algorithms (including CC-based methods), but they did not perform any feature selection, which is of the particular interest in the present paper. In the present work, we use an updated version of the database-MIMIC III [15].

3 General framework for learning classifier chains

We consider a general framework for learning classifier chains allowing for different loss functions. To reduce over-fitting and select relevant features, we add regularization terms to the loss function. First we introduce some notations. Let $||\cdot||_1$, $||\cdot||_2$ denote ℓ_1 and ℓ_2 norms, respectively. Moreover, let $\langle\cdot,\cdot\rangle$ denote a scalar product. In multi-label classification each object is described by feature vector $\mathbf{x} \in R^p$ and label vector $\mathbf{y} \in \{0,1\}^K$. For simplicity we assume that all labels are binary, but actually all results presented below remain valid for any types of labels. The goal is to use training sample $S = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) : i = 1, \ldots, N\}$ to learn a model to predict $\mathbf{y}$ using $\mathbf{x}$. Both $\mathbf{x}$ and $\mathbf{y}$ are assumed to be random. Let $\mathbf{A} \in R^{K\times p}$ be a matrix of parameters, whose k-th row $\mathbf{a}_k$ is a parameter vector corresponding to the model for k-th label. Define $\mathbf{y}_{1:k-1} = (y_1, \ldots, y_{k-1}, \mathbf{0}_{K-k})$, for $k \geq 2$ (for convenience we augment first $k-1$ labels by zeros and we set $\mathbf{y}_{1:0} = \mathbf{0}_{K-1}$). Let $\mathbf{b}_k \in R^{K-1}$ be the corresponding parameter vectors and $\mathbf{B} \in R^{K\times(K-1)}$ be matrix of parameters, whose k-th row is $\mathbf{b}_k$. Moreover let $\mathbf{W} = [\mathbf{A}, \mathbf{B}] \in R^{K\times(p+K-1)}$ be matrix containing all parameters and $\mathbf{w}_k$ its k-th row. In classifier chains the linear predictor for k-th label is $\langle \mathbf{x}, \mathbf{a}_k\rangle + \langle \mathbf{y}_{1:k-1}, \mathbf{b}_k\rangle$. The quality of the prediction for the k-th label is assessed by a loss function $l_k(s, y)$. The most popular are: logistic loss $l(s, y) = \log(1 + \exp(-ys))$ or hinge-loss $l(s, y) = \max\{0, 1 - sy\}$, corresponding to SVM method. In the case of quantitative labels, the natural choice is squared loss $l(s, y) = (s - y)^2$. To keep generality we allow that the loss functions may differ across the models. The empirical risk based on sample $S = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) : i = 1, \ldots, N\}$ is defined as

$$L_S(\mathbf{W}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} l_k\left(\langle \mathbf{x}^{(i)}, \mathbf{a}_k\rangle + \langle \mathbf{y}^{(i)}_{1:k-1}, \mathbf{b}_k\rangle, y_k^{(i)}\right). \quad (1)$$

We also define theoretical risk

$$L(\mathbf{W}) := E_{\mathbf{x},\mathbf{y}} \sum_{k=1}^{K} l_k\left(\langle \mathbf{x}, \mathbf{a}_k\rangle + \langle \mathbf{y}_{1:k-1}, \mathbf{b}_k\rangle, y_k\right) \quad (2)$$

which will be used in our theoretical results. Matrix $\mathbf{W}$ can be learnt by minimizing regularized empirical risk, i.e. solving $\underset{\mathbf{W}\in\mathcal{W}_0}{\operatorname{argmin}}\{L_S(\mathbf{W})\}$, where $\mathcal{W}_0 = \{\mathbf{W} = [\mathbf{A}, \mathbf{B}] : \sum_{k=1}^{K} ||\mathbf{a}_k||_1 + \sum_{k=1}^{K} ||\mathbf{b}_k||_1 \leq \Lambda\}$ and Λ is a parameter. The first term induces

sparsity in a parameter vector $\mathbf{a}_k$ which in turn allows to select relevant features. The second term induces sparsity in vector $\mathbf{b}_k$, which allows to select labels influencing the current target label. The above problem is equivalent to solving dual problem

$$\operatorname*{argmin}_{\mathbf{W}} \left\{ L_S(\mathbf{W}) + \lambda \sum_{k=1}^{K} ||\mathbf{a}_k||_1 + \lambda \sum_{k=1}^{K} ||\mathbf{b}_k||_1 \right\}, \quad (3)$$

which reduces to solving K independent problems with ℓ_1 regularization each. Parameter λ in (3) corresponds to Λ.

4 Parsimonious classifier chains (parCC)

4.1 Description of the method

In this Section we describe the proposed method and provide theoretical analysis. The proposed approach aims to integrate the strengths of classifier chains and joint feature selection, derived from multi-task learning. It allows to exploit label dependencies and at the same time reduce the number of features. Learning under a constraint on the number of features can be expressed as solving $\operatorname{argmin}_{\mathbf{W}\in\mathcal{W}_C}\{L_S(\mathbf{W})\}$, where $\mathcal{W}_C = \{\mathbf{W} = [\mathbf{A}, \mathbf{B}] : \sum_{j=1}^{p} \mathbf{1}\{\exists_{1\leq k\leq K} |a_{k,j}| \neq 0\} \leq T\}$, $a_{k,j}$ is j-th coordinate of $\mathbf{a}_k$ and T is maximal admissible number of features. The discrete and non-convex nature of the constraint makes its optimization challenging. The possible strategy for tackling this problem is to approximate the constraint with continuous and convex function. We approximate the constraint in $\mathcal{W}_C$ using $\ell_{2,1}$ norm (sum of the ℓ_2 norms of the columns) and in addition we use l_1 norm to ensure sparsity in label vectors. In the proposed method we solve $\operatorname{argmin}_{\mathbf{W}\in\mathcal{W}}\{L_S(\mathbf{W})\}$, where

$$\mathcal{W} = \{\mathbf{W} = [\mathbf{A}, \mathbf{B}] : \sum_{j=1}^{p} ||\mathbf{A}_j||_2 + \sum_{k=1}^{K} ||\mathbf{b}_k||_1 \leq \Lambda\}, \quad (4)$$

$\mathbf{A}_j$ is j-th column of matrix $\mathbf{A}$ and Λ is a regularization parameter. The first term in (4) ensures common sparsity pattern in $\mathbf{A}$, whereas the second term induces sparsity in label vectors. The second term is introduced to discard labels that are conditionally independent from the label being a current target, given the feature vector. Problem (4) is equivalent to

$$\operatorname*{argmin}_{\mathbf{W}} \left\{ L_S(\mathbf{W}) + \lambda \sum_{j=1}^{p} ||\mathbf{A}_j||_2 + \lambda \sum_{k=1}^{K} ||\mathbf{b}_k||_1 \right\}, \quad (5)$$

where λ corresponds to Λ. From the practical point of view it is important to establish a connection between regularization parameter λ and T, the maximal number of features that can be used in the model. It is relatively easy to obtain the solutions for many different values of λ (using technique of so-called 'warm starts' [8]), which in turn allows to select the value of λ best corresponding to the desired number of features. We use the following strategy. Assume that the maximal admissible number of features is T_0 and we would like to select λ corresponding to T_0. For a decreasing sequence of regularization parameters $\lambda_1 > \lambda_2 > \ldots > \lambda_L$ we have corresponding subsets of selected features of sizes $T_1 < T_2 < \ldots < T_L$. If $T_0 = T_i$, for some $i = 1, \ldots, L$, then we select λ_i as the value corresponding to T_0. Otherwise we select λ_i, such that $T_0 \in (T_i, T_{i+1})$, i.e. a value of λ that best matches T_0. In order to obtain more accurate solution, one can take denser grid of λ.

Finally, let us mention that parCC can be modified when some other constraints are of interest, e.g. when total cost associated with the features is limited, e.g. $\sum_{j=1}^{p} c_j \mathbf{1}\{\exists_{1\leq k\leq K} |a_{k,j}| \neq 0\} \leq T$, where c_j is a cost for feature j. Then, instead of (5) we would consider $\operatorname{argmin}_{\mathbf{W}} \left\{ L_S(\mathbf{W}) + \lambda \sum_{j=1}^{p} c_j ||\mathbf{A}_j||_2 + \lambda \sum_{k=1}^{K} ||\mathbf{b}_k||_1 \right\}$.

4.2 Theoretical analysis

Let $\widehat{\mathbf{W}}$ be a solution of (5). In the following we will bound the generalization error of $\widehat{\mathbf{W}}$ using Rademacher complexity bounds. The multi-label Rademacher Complexity (RC) [16] is defined as

$$R_S(\mathcal{W}) := E_\epsilon \left[\frac{1}{N} \sup_{\mathbf{W}\in\mathcal{W}} \sum_{i=1}^{N} \sum_{k=1}^{K} \epsilon_i^k \left(\langle \mathbf{x}^{(i)}, \mathbf{a}_k \rangle + \langle \mathbf{y}_{1:k-1}^{(i)}, \mathbf{b}_k \rangle \right) \right].$$

Here, the expectation is over ϵ_i^k, which are i.i.d. Rademacher random variables, i.e. $P(\epsilon_i^k = 1) = P(\epsilon_i^k = -1) = 0.5$. Bounds on Rademacher complexity of a class yield generalization error bounds for classifiers picked from that class (assuming the loss function is Lipschitz), see Theorems 2 and 3 below. The following Theorem, giving upper bound of RC for class $\mathcal{W}$, is our main tool in bounding the generalization error.

Theorem 1 *Assume that* $\max_i ||\mathbf{x}^{(i)}||_\infty < B$ *and let* $\mathcal{W}$ *be a class defined in (4). Then*

$$R_S(\mathcal{W}) \leq O\left(\Lambda B \sqrt{\frac{K \log(p)}{N}} \right) + O\left(\Lambda \sqrt{\frac{\log(K(K-1))}{N}} \right).$$

Proof. Define class $\mathcal{W}_1 := \{\mathbf{W} = [\mathbf{A}, \mathbf{B}] : ||\mathbf{A}||_{2,1} \leq \Lambda \text{ and } ||\mathbf{B}||_{1,1} \leq \Lambda\}$. Define also classes $\mathcal{A} := \{\mathbf{A} \in R^{K\times p} : ||\mathbf{A}||_{2,1} \leq \Lambda\}$ and $\mathcal{B} := \{\mathbf{B} \in R^{K\times(K-1)} : ||\mathbf{B}||_{1,1} \leq \Lambda\}$. Since $\mathcal{W} \subseteq \mathcal{W}_1$ we have

$$\begin{aligned} &R_S(\mathcal{W}) \leq R_S(\mathcal{W}_1) = \\ &E_\epsilon \Big[\sup_{\mathbf{W}\in\mathcal{W}_1} \Big(\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \epsilon_i^k \langle \mathbf{x}^{(i)}, \mathbf{a}_k \rangle + \\ &\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \epsilon_i^k \langle \mathbf{y}_{1:k-1}^{(i)}, \mathbf{b}_k \rangle \Big) \Big] \leq \\ &E_\epsilon \left[\frac{1}{N} \sup_{\mathbf{W}\in\mathcal{W}_1} \sum_{i=1}^{N} \sum_{k=1}^{K} \epsilon_i^k \langle \mathbf{x}^{(i)}, \mathbf{a}_k \rangle \right] + \\ &E_\epsilon \left[\frac{1}{N} \sup_{\mathbf{W}\in\mathcal{W}_1} \sum_{i=1}^{N} \sum_{k=1}^{K} \epsilon_i^k \langle \mathbf{y}_{1:k-1}^{(i)}, \mathbf{b}_k \rangle \right] = \\ &R_S(\mathcal{A}) + R_S(\mathcal{B}), \end{aligned}$$

where the inequality above follows from the fact that $\sup_{a\in A}[f(a) + g(a)] \leq \sup_{a\in A} f(a) + \sup_{a\in A} g(a)$. It follows from [16] that

$$R_S(\mathcal{A}) \leq O\left(\Lambda B \sqrt{\frac{K \log(p)}{N}} \right) \quad \text{and}$$
$$R_S(\mathcal{B}) \leq O\left(\Lambda \sqrt{\frac{\log(K(K-1))}{N}} \right),$$

which yields the assertion.

Note that if $2\log(K) \leq K \log(p)$ ($k^2 \leq p^K$), the bound in Theorem 1 can be simplified to $O\left(\Lambda B \sqrt{\frac{K \log(p)}{N}} \right)$.

The following Lemma is an important tool to prove the generalization error bounds.

Lemma 1 *Let $\mathcal{W}$ be a class defined in (4) and assume that l_k is ρ-Lipschitz loss function, for $k = 1, \ldots, K$. The following inequality holds*

$$E_S \sup_{\mathbf{W} \in \mathcal{W}} [L(\mathbf{W}) - L_S(\mathbf{W})] \leq 2\rho E_S R_S(\mathcal{W}).$$

The above Lemma can be proved using contraction Lemma (see e.g. Lemma 26.9 in [26]) and the straightforward adaptation of the proof of Lemma 26.2 there to the multi-label case.

The following Theorem shows that, in expectation, $\widehat{\mathbf{W}}$ is close to the optimal matrix in class $\mathcal{W}$.

Theorem 2 *Assume that $\max_i ||\mathbf{x}^{(i)}||_\infty < B$ and let l_k be ρ-Lipschitz loss function. Define $\mathbf{W}^* := \arg\min_{\mathbf{W} \in \mathcal{W}} L(\mathbf{W})$. Then we have*

$$E_S[L(\widehat{\mathbf{W}}) - L(\mathbf{W}^*)] \leq O\left(\rho \Lambda B \sqrt{\frac{K \log(p)}{N}}\right) + O\left(\rho \Lambda \sqrt{\frac{\log(K(K-1))}{N}}\right).$$

Proof. Note that

$$E_S[L(\widehat{\mathbf{W}}) - L(\mathbf{W}^*)] \leq E_S[L(\widehat{\mathbf{W}}) - L_S(\widehat{\mathbf{W}})] \leq 2\rho E_S R_S(\mathcal{W}),$$

where the first inequality follows from $L(\mathbf{W}^*) = E_S L_S(\mathbf{W}^*) \geq E_S L_S(\widehat{\mathbf{W}})$ and the second inequality follows directly from Lemma 1. Using Theorem 1 yields the assertion.

The next Theorem gives standard generalization error bound based on Rademacher complexity.

Theorem 3 *Assume that $\max_i ||\mathbf{x}^{(i)}||_\infty < B$ let l_k be ρ-Lipschitz loss function. Moreover, assume that for all s, y we have that $|l(s, y)| \leq c$. Define $\mathbf{W}^* := \arg\min_{\mathbf{W} \in \mathcal{W}} L(W)$. Then, with probability of at least $1 - \delta$,*

$$L(\widehat{\mathbf{W}}) - L(\mathbf{W}^*) \leq O\left(\rho \Lambda B \sqrt{\frac{K \log(p)}{N}}\right) + O\left(\rho \Lambda \sqrt{\frac{\log(K(K-1))}{N}}\right) + 5c\sqrt{\frac{2\log(8/\delta)}{N}}.$$

Proof. Proceeding in the same way as for single-label classification (see Theorem 26.5 in [26]) it can be shown that $L(\widehat{\mathbf{W}}) - L(\mathbf{W}^*) \leq 2\rho R_S(\mathcal{W}) + 5c\sqrt{2\log(8/\delta)/N}$. The proof is based on McDiarmid's inequality and Lemma 1. Combining this with Theorem 1 yields the assertion.

5 Related methods

5.1 Adaptive classifier chains

In order to reduce the total number of features one should generally prefer features which are relevant in all models. In Adaptive Classifier Chains (ACC), penalty factors corresponding to the considered features are adaptively changed. The key idea is as follows. If a given feature has been selected as relevant in one of the models in a chain, then it is more likely to be selected in the next model. Adaptive classifier chains was proposed in our recent paper [28], in a simplified form, only for logistic loss and one particular choice of penalty function h, in the context of cost-sensitive feature selection. Here we consider more general form, allowing arbitrary loss functions. We also propose long-memory penalty functions. The method works as follows. For $k = 1$ we minimize w.r.t $\mathbf{a}_k$

$$\frac{1}{N}\sum_{i=1}^{N} l_k\left(\langle \mathbf{x}^{(i)}, \mathbf{a}_k \rangle, y_k^{(i)}\right) + \lambda \sum_{j=1}^{p} |a_{k,j}|,$$

as in standard classifier chains. The next steps in ACC and CC are different, namely in ACC, for $k \geq 2$ we minimize with respect to $\mathbf{a}_k$ and $\mathbf{b}_k$

$$\frac{1}{N}\sum_{i=1}^{N} l_k\left(\langle \mathbf{x}^{(i)}, \mathbf{a}_k \rangle + \langle \mathbf{y}_{1:k-1}^{(i)}, \mathbf{b}_k \rangle, y_k^{(i)}\right) + \lambda \sum_{j=1}^{p} h(|\hat{a}_{1,j}|, \ldots, |\hat{a}_{k-1,j}|)|a_{k,j}| + \lambda ||\mathbf{b}_k||_1, \quad (6)$$

where $a_{k,j}$ denotes j-th coordinate of $\mathbf{a}_k$ and $\lambda > 0$ is a regularization parameter. Function h plays an important role in the above formula. It can be seen as a penalty for j-th feature. The possible choices for $h(|\hat{a}_{1,j}|, \ldots, |\hat{a}_{k-1,j}|)$ are: $(1 + \sum_{s=1}^{k-1} |\hat{a}_{s,j}|)^{-1}$, $(1 + \max_{1 \leq s \leq k-1} |\hat{a}_{s,j}|)^{-1}$ (long-memory penalties) or $(1+|\hat{a}_{k-1,j}|)^{-1}$ (short-memory penalty), the last option was used in [28]. If feature j was selected as relevant in the previous model(s) then $h(|\hat{a}_{1,j}|, \ldots, |\hat{a}_{k-1,j}|) < 1$ and we reduce the penalty term in model k. The more relevant is the feature in the previous models, the larger the reduction in model k. The method is similar to adaptive lasso method [31] designed for linear and logistic regression, where adaptive weights are used for penalizing different coefficients in the ℓ_1 penalty. In experiments, as penalty functions, we used: sum (this variant is denoted as ACC-SUM), max (ACC-MAX) and short-memory penalty (ACC-ABS). Since usually ACC-SUM outperforms ACC-MAX, we only present the results for ACC-SUM and ACC-ABS.

5.2 Binary relevance with $\ell_{2,1}$ penalty

The natural way to enforce a common sparsity pattern is to learn models using $\ell_{2,1}$ regularization [22] i.e. to solve

$$\operatorname*{argmin}_{\mathbf{A}} \left\{ \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} l_k\left(\langle \mathbf{x}^{(i)}, \mathbf{a}_k \rangle\right) + \lambda \sum_{j=1}^{p} ||\mathbf{A}_j||_2 \right\}. \quad (7)$$

We call this method binary relevance with $\ell_{2,1}$ regularization (*BR+l21*). The method aims to select features that are relevant for different labels simultaneously. Note however that the method does not utilize label dependencies directly, which results in lower accuracy of prediction when compare to parCC (see Section 'Experiments').

6 Experiments

In the experiments we compared the proposed method parCC with the related ones, describe above: BR+l21, ACC-ABS and ACC-SUM. As baseline we use standard classifier chains with ℓ_1 regularization separately in each model in the chain (denoted as CC+l1) [25] and standard BR with ℓ_1 regularization (denoted as BR+l1). To make a comparison fair, in all methods we used logistic loss. We also experimented with hinge loss, for which the conclusions are similar. We study how the number of selected features affect classification performance; in particular we are interested in improving the performance for limited number of features. We consider 4 popular evaluation measures: subset accuracy, example-based F measure, Hamming measure (1-Hamming loss) and ranking measure (1-ranking loss) [9], averaged over 10 CV folds (repeated 5 times). Formal definitions of the measures are included in the Supplement.

6.1 Application to mutlimorbidity prediction

Multimorbidity (co-occurence of two or more diseases in one patient) is one of major challenges facing health care systems worldwide. Multimorbidity is associated with significant reductions in functional status, quality of life and increased risk of death [7, 10]. Successful prediction of multimorbidity enhances the decision support of medical doctors, helps to develop preventive strategies and identify individuals at risk. We apply the methods discussed above to large clinical database MIMIC III, which can be used to multimorbidity prediction. The database contains information on 33166 patients of various intensive care units who are diagnosed according to the coding scheme ICD-9. As labels we consider indicators of 10 families of diseases, which were already used in previous studies [32, 3]: hypertension, kidney, fluid, hypotension, lipoid, liver, diabetes, thyroid, copd and thrombosis. Table 1 contains summary statistics and distributions of labels. Among considered diseases, Hypertensive disease is the most common (66% of patients) whereas Thrombosis is the rarest disease (6% of patients). Label density (average fraction of active labels among all labels per observation) is 0.27. We observe significant dependencies between diseases (see Figure 1), among which (fluid, kidney) is the pair with the highest correlation 0.31 (non-significant dependencies are marked with 'X', i.e. those with p-value of the corresponding test larger than sign. level $\alpha = 0.05$). We use 308 numerical features, most of which correspond to some laboratory events. All features are listed in Tables 1-3 in Supplement. The first, relatively small, group of features is obtained from medical history or are administrative information (e.g. sex, age, marital status). The second group are medical scores used to track a person's status during the stay in an intensive care unit (e.g. Braden score used to assess a risk of developing a pressure ulcer). The largest group of features are blood and diagnostic tests (e.g. Glucose, Sodium, etc.). Obviously some diagnostic tests may be irrelevant in prediction of a particular disease. Methods proposed in this paper are useful to reduce the number of unnecessary diagnostic tests, which is important to control healthcare costs as well as to reduce the possible negative effects of some diagnostic procedures.

Table 1. Summary statistics of MIMIC III database.

number of observations (patients)	33166
number of features	308
number of labels (diseases)	10
label density (LD)	0.27
% of patients with hypertension disease	66%
% of patients with kidney disease	38%
% of patients with fluid disease	37%
% of patients with hypotension disease	35%
% of patients with lipoid disease	31%
% of patients with liver disease	23%
% of patients with diabetes disease	14%
% of patients with thyroid disease	14%
% of patients with copd disease	7%
% of patients with thrombosis disease	6%

Figure 3 shows how the considered performance measures depend on the number of features. First, the proposed method parCC outperforms other methods with respect to Subset Accuracy, F measure and ranking measure. For Hamming measure, parCC works on par with BR+l21, which is concordant with well-known results stating that utilizing label dependencies does not improve Hamming measure [5]. Observe that BR+l21 is second best, followed by CC+l1. The ranking of the methods 'parCC'$\succ$'BR+l21'$\succ$'CC+l1'$\succ$'BR+l1' suggests that both chaining and simultaneous learning improve the

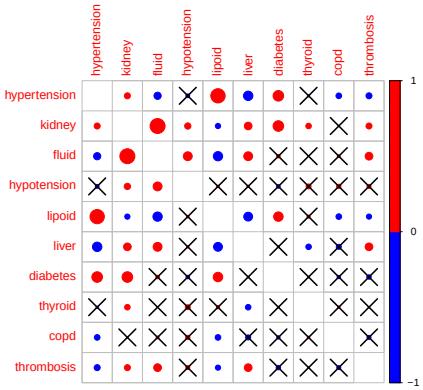

Figure 1. MIMIC III database. Label-label dependencies.

results. Surprisingly, adaptive classifier chains (ACC) work better than CC+l1 and BR+l1 only for small number of features (below 40), whereas for larger number of features the performance of ACC deteriorates. The curves show that parCC allows to choose the most parsimonious model. For example assume that we would like to achieve subset accuracy equal to 10%, then we need to use 60 features for parCC and as many as 100 features for the second best method BR+l21. Alternatively we can assume that our budget is limited to 60 features. Then for parCC, Subset Accuracy is 10%, whereas for BR+l21 is 9%. The experiment indicates that the proposed method can be successfully applied to real-world multi-label medical datasets for which the admissible number of features should be limited. We also analysed which features are selected as the most relevant ones by the proposed methods. Note that parCC selects features which are correlated with many labels simultaneously. To verify this, we performed a post-hoc analysis, namely we analysed dependencies between top features selected by parCC and labels. The results are shown in Figure 2. The non-significant dependencies are marked with 'X'. The larger the circle, the larger the dependency. BUN (blood urea nitrogen), which is a medical test that measures the amount of urea nitrogen found in blood, is a top-ranked feature, accorindg to parCC. It is correlated with 5 diseases simultaneously and the largest correlation is with kidney disease. The importance of BUN in predicting mortality in ICU patients has been proved in other studies, e.g. [2] report that BUN is associated with mortality in critically ill patients admitted to ICU and might constitute an important parameter for risk stratification in the critically ill.

6.2 Benchmark datasets

The experiments are conducted on 11 benchmark datasets `http://mulan.sourceforge.net/datasets.html`. The effectiveness of multi-label methods exploiting label dependencies can most likely be demonstrated on datasets whose label density (LD) is reasonably large. We thus chose ten most popular labels, from each of the datasets, aiming to produce multi-label datasets with reasonable LD. For mediamill and nuswide, we take random samples of 10000 observations. Table 2 shows: number of observations (n), features (p), labels (K) and LD for the datasets. We consider three constraints on the number of features: $T = 10$, $T = 20$ and $T = 50$.

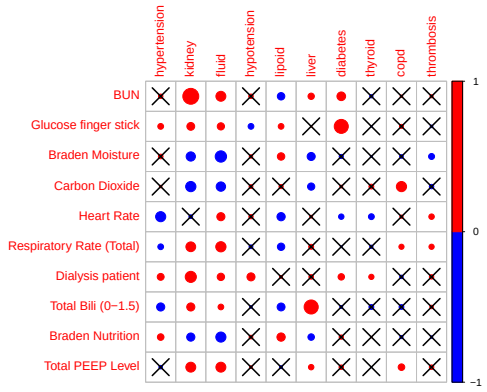

Figure 2. MIMIC III database. Dependencies between labels and top-ranked features according to parCC.

Table 2. Summary statistics of benchmark datasets.

dataset	n	p	K	LD
cal500	502	68	10	0.606
nuswide	10000	128	10	0.123
medical	978	1449	10	0.096
emotions	592	71	6	0.312
yeast	2417	103	10	0.393
scene	2407	294	6	0.179
mediamill	10000	120	10	0.301
flags	194	19	7	0.485
Science	6428	1859	10	0.166
Reference	8027	1983	10	0.183
Health	9205	1530	10	0.199

Tables 3-6 show subset accuracy, F measure, Hamming measure and ranking measure, for $T = 50$, for 11 benchmark datasets. Results for top 10 and top 20 features are presented in Supplemental Material. The best performing method is highlighted in boldface. The last row contains the averaged ranks, the larger the ranks the better. Note that parCC is the best according to the first two evaluation measures. For Hamming measure, BR+l21 works slightly better than parCC, which is second best. This is consistent with theoretical results indicating that BR is an optimal strategy when Hamming loss is of interest (we observed the similar effect for MIMIC III database). ACC-ABS is the second best in terms of subset accuracy. As expected, BR+l1 works poorly, particularly wrt subset measure and F measure, which is due to ignoring dependencies between labels and not utilizing simultaneous learning via $\ell_{2,1}$ regularization.

Our findings confirm that simultaneous learning improves classification accuracy when the number of features is limited. Indeed, $\ell_{2,1}$ regularization helps for both BR and CC approaches when compare to standard ℓ_1 regularization. It seems that both components (simultaneous learning via $\ell_{2,1}$ regularization as well as exploiting label dependencies via chaining) play important role. Observe that for most evaluation measures CC+l1$\succ$BR+l1, which supports usefulness of chaining and parCC$\succ$CC+l1, which confirms advantages of joint learning. Although the averaged ranks suggest that parCC is an overall winner, note that ACC works better for selected dataset. For example, ACC-SUM has the highest subset accuracy for scene dataset. Regarding ACC, it follows from the experiments that both penalties work similarly.

To analyse the results in detail, we followed the two-step statistical procedure recommended by [6]. In the first step we use the Friedman test (based on averaged ranks) to assess the null hypothesis that all methods have equal performance. When null hypothesis is rejected a Nemenyi post-hoc test is used to compare methods in a pairwise way. Figure 4 shows the results for $T = 50$ (the results for $T = 10, 20$ are shown in Supplement). In all cases, the null hypothesis of the Friedman test is rejected, when standard significance level 0.05 is assumed. The blue line denotes the Nemenyi critical region. When the averaged rank for method 'A' is within a critical region of method 'B', then we conclude that there is no significant difference in performances between 'A' and 'B'. The critical region for parCC is highlighted. In all cases, parCC is in a group of best-performing methods. Although the differences between the two first methods are not significant, note that in almost all cases there is a significant difference between the best performing method and the two least efficient methods (BR+l1 and CC+l1).

7 Conclusions

In this paper we proposed a novel method parCC of learning classifier chains. By employing an individual sparsity inducing ℓ_1 norm and a group sparsity inducing $\ell_{2,1}$-norm together, the proposed model can induce both a sparse dependency structure over labels, and a common set of predictive features across the multiple models in a chain. In theoretical analysis we bounded the Rademacher Complexity, which allowed to determine the generalization error bounds using well-known techniques. We demonstrated that the method can be successfully applied to multimorbidity prediction, the task where limiting the number of features is particularly important due to costs and risk of negative effects of diagnostic tests. Our empirical results on a number of benchmark multi-label data sets show that the proposed approach can outperform related methods.

Table 3. Subset accuracy for benchmark datasets ($T = 50$).

	BR+l21	parCC	CC+l1	BR+l1	ACC-ABS	ACC-SUM
cal500	0.012	0.014	0.012	**0.014**	0.008	0.012
nuswide	0.324	**0.331**	0.308	0.308	0.312	0.311
medical	0.748	**0.754**	0.723	0.713	0.631	0.530
emotions	0.263	0.275	0.270	0.248	0.282	**0.284**
yeast	0.136	**0.185**	0.148	0.050	0.171	0.173
scene	0.182	0.206	0.032	0.032	0.421	**0.472**
mediamill	0.121	**0.160**	0.158	0.129	0.140	0.121
flags	0.144	0.211	0.191	0.144	**0.257**	0.252
Science	0.203	0.209	0.185	0.179	0.219	**0.221**
Reference	0.309	**0.323**	0.231	0.226	0.293	0.292
Health	0.332	0.349	0.321	0.324	**0.354**	0.344
avg. rank	3.1	**5.1**	2.7	2.0	4.2	3.9

Table 4. F measure for benchmark datasets ($T = 50$).

	BR+l21	parCC	CC+l1	BR+l1	ACC-ABS	ACC-SUM
cal500	0.712	0.701	0.713	**0.719**	0.691	0.696
nuswide	**0.092**	0.074	0.010	0.021	0.025	0.023
medical	0.654	**0.655**	0.634	0.638	0.531	0.456
emotions	0.579	**0.599**	0.573	0.522	0.577	0.581
yeast	**0.605**	0.593	0.569	0.545	0.580	0.589
scene	0.206	0.228	0.034	0.034	0.456	**0.510**
mediamill	0.595	0.590	0.592	**0.596**	0.566	0.544
flags	0.678	0.688	0.687	0.683	**0.699**	0.698
Science	0.527	0.531	0.519	0.516	**0.533**	0.532
Reference	0.682	**0.688**	0.653	0.650	0.675	0.675
Health	0.536	0.552	0.536	0.529	**0.566**	0.549
avg. rank	4.1	**4.6**	2.5	2.4	3.8	3.5

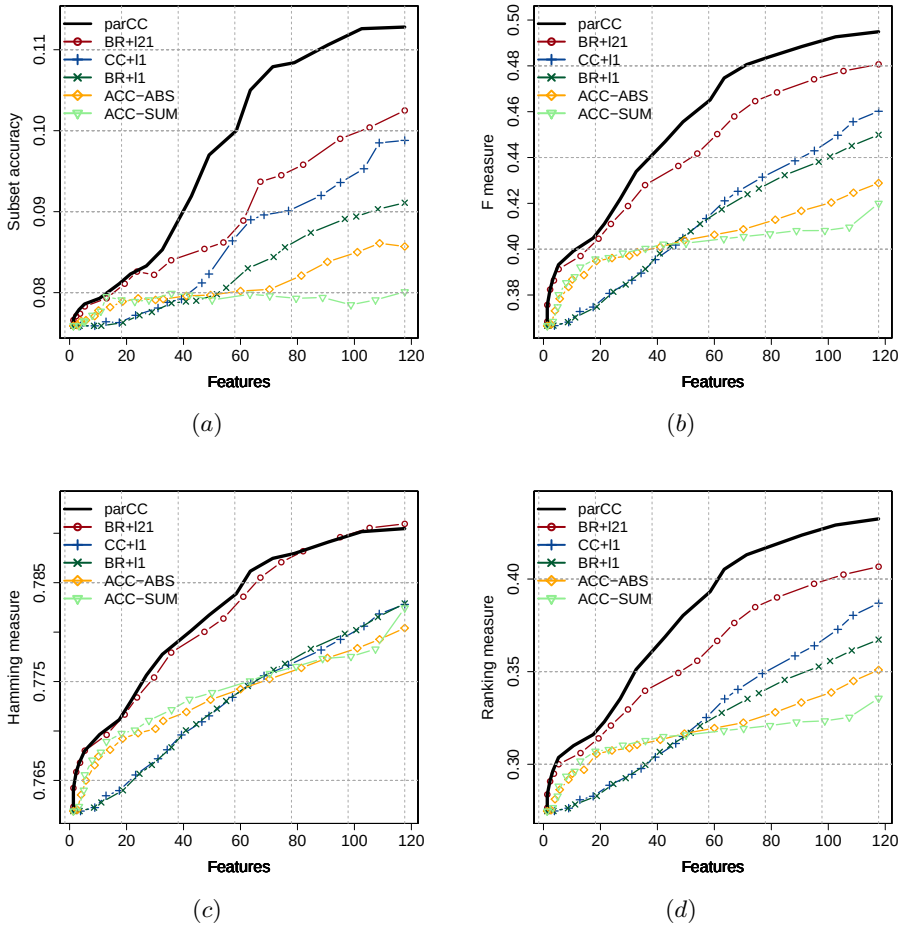

Figure 3. MIMIC3 dataset. (a) Subset accuracy, (b) F measure, (c) Hamming measure and (d) ranking measure wrt to number of features.

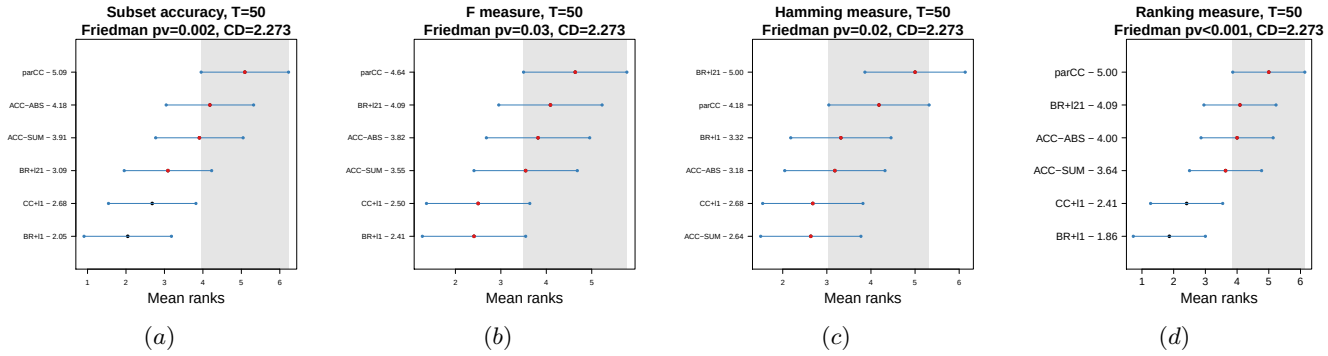

Figure 4. Results of Friedman and pairwise tests ($T = 50$). The blue line denotes the Nemenyi critical region. The critical region for parCC is highlighted.

Table 5. Hamming measure for benchmark data ($T = 50$).

	BR+l21	parCC	CC+l1	BR+l1	ACC-ABS	ACC-SUM
cal500	**0.628**	0.618	0.615	0.625	0.606	0.613
nuswide	**0.889**	0.887	0.879	0.880	0.882	0.882
medical	**0.967**	0.967	0.964	0.964	0.954	0.943
emotions	**0.802**	0.784	0.786	0.792	0.766	0.768
yeast	**0.744**	0.726	0.731	0.728	0.719	0.719
scene	0.852	**0.854**	0.826	0.826	0.836	0.830
mediamill	0.811	0.811	0.812	**0.813**	0.801	0.794
flags	0.733	0.737	0.731	0.735	**0.739**	0.738
Science	0.876	0.876	0.873	0.873	0.877	**0.877**
Reference	0.908	**0.909**	0.900	0.899	0.906	0.906
Health	**0.843**	0.829	0.822	0.834	0.831	0.828
avg. rank	**5.0**	4.2	2.7	3.3	3.2	2.6

Table 6. Ranking measure for benchmark data ($T = 50$).

	BR+l21	parCC	CC+l1	BR+l1	ACC-ABS	ACC-SUM
cal500	**0.311**	0.306	0.257	0.272	0.290	0.296
nuswide	**0.385**	0.373	0.313	0.321	0.327	0.325
medical	0.873	**0.874**	0.858	0.866	0.741	0.652
emotions	0.539	**0.559**	0.529	0.483	0.539	0.542
yeast	**0.530**	0.523	0.489	0.450	0.508	0.517
scene	0.205	0.226	0.034	0.034	0.452	**0.504**
mediamill	0.568	0.569	**0.573**	0.573	0.543	0.512
flags	0.496	0.527	0.514	0.495	**0.528**	0.525
Science	0.485	0.490	0.475	0.470	**0.495**	0.493
Reference	0.601	**0.609**	0.557	0.554	0.592	0.591
Health	0.520	0.543	0.527	0.512	**0.560**	0.542
avg. rank	4.1	**5.0**	2.4	1.9	4.0	3.6

REFERENCES

[1] A. Argyriou, T. Evgeniou, and M. Pontil, 'Convex multi-task feature learning', *Machine Learning*, **73**(3), 243–272, (2008).
[2] O. Arihan, B. Wernly, M. Lichtenauer, M. Franz, B. Kabisch, J. Muessig, M. Masyuk, A. Lauten, P. C. Schulze, U. C. Hoppe, M. Kelm, and C. Jung, 'Blood Urea Nitrogen (BUN) is independently associated with mortality in critically ill patients admitted to ICU', *PloS one*, **13**(1), 1–10, (2018).
[3] S. Bromuri, D. Zufferey, J. Hennebert, and M. Schumacher, 'Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms', *Journal of Biomedical Informatics*, **51**, 165–175, (2014).
[4] K. Dembczyński, W. Cheng, and E. Hüllermeier, 'Bayes optimal multilabel classification via probabilistic classifier chains', in *Proceedings of the Twenty-Seventh International Conference on Machine Learning*, volume 22, pp. 109–117, (2010).
[5] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier, 'On label dependence and loss minimization in multi-label classification', *Machine Learning*, **88**, 5–45, (2012).
[6] J. Demšar, 'Statistical comparisons of classifiers over multiple data sets', *J. Mach. Learn. Res.*, **7**, 1–30, (2006).
[7] M. Fortin, L. Lapointe, C. Hudon, A. Vanasse, A. L. Ntetu, and D. Maltais, 'Multimorbidity and quality of life in primary care: a systematic review', *Health and Quality of Life Outcomes*, **2**(1), 1–12, (2004).
[8] J. Friedman, T. Hastie, and R. Tibshirani, 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software*, **22**, (2010).
[9] E. Gibaja and S. Ventura, 'A tutorial on multilabel learning', *ACM Computing Surveys*, **47**(3), 1–38, (2015).
[10] R. Gijsen, N. Hoeymans, F.G. Schellevis, D. Ruwaard, W.A. Satariano, and G.A.M. van den Bos, 'Causes and consequences of comorbidity: A review', *Journal of Clinical Epidemiology*, **54**(7), 661 – 674, (2001).
[11] Y. Guo and S. Gu, 'Multi-label classification using conditional dependency networks', in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, IJCAI'11, pp. 1300–1305, (2011).
[12] Y. Guo and W. Xue, 'Probabilistic multi-label classification with sparse feature learning', in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pp. 1373–1379, (2013).
[13] E. J. Hall and D. J. Brenner, 'Cancer risks from diagnostic radiology', *The British Journal of Radiology*, **81**(965), 362–378, (2008).
[14] S. Ji, L. Tang, S. Yu, and J. Ye, 'A shared-subspace learning framework for multi-label classification', *ACM Trans. Knowl. Discov. Data*, **4**(2), 1–29, (2010).
[15] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, Anthony C. L., and R. G. Mark, 'MIMIC-III, a freely accessible critical care database', *Scientific Data*, **3**, 1–9, (2016).
[16] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari, 'Regularization techniques for learning with matrices', *Journal of Machine Learning Research*, **13**, 1865–1890, (2012).
[17] S. Kashef, H. Nezamabadi-pour, and B. Nikpour, 'Multilabel feature selection: A comprehensive review and guiding experiments', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **8**(2), 1–29, (2018).
[18] A. Kumar, S. Vembu, A. K. Menon, and C. Elkan, 'Beam search algorithms for multilabel learning', *Machine Learning*, **92**(1), (2013).
[19] R. S. Lagasse, 'Anesthesia safety: Model or myth?: A review of the published literature and analysis of current original data', *Anesthesiology: The Journal of the American Society of Anesthesiologists*, **97**(6), 1609–1617, (2002).
[20] W. Liu and I. W. Tsang, 'On the optimality of classifier chain for multi-label classification', in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pp. 712–720, (2015).
[21] S. Lu and K. Mineichi, 'Optimization of classifier chains via conditional likelihood maximization', *Pattern Recognition*, **74**, 503 – 517, (2018).
[22] P. Naula, A. Airola, T. Salakoski, and T. Pahikkala, 'Multi-label learning under feature extraction budgets', *Pattern Recognition Letters*, **40**, 56 – 65, (2014).
[23] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. C. Merschmann, 'Categorizing feature selection methods for multi-label classification', *Artificial Intelligence Review*, **49**(1), 1–22, (2018).
[24] J. Read, L. Martino, and D. Luengo, 'Efficient monte carlo methods for multi-dimensional learning with classifier chains', *Pattern Recognition*, **47**(3), 1535–1546, (2014).
[25] J. Read, B. Pfahringer, G. Holles, and E. Frank, 'Classifier chains for multi-label classification', in *ECML/PKDD*, pp. 254–269, (2009).
[26] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2013.
[27] P. Teisseyre, 'CCnet: Joint multi-label classification and feature selection using classifier chains and elastic net regularization', *Neurocomputing*, **235**, 98 – 111, (2017).
[28] P. Teisseyre, D. Zufferey, and M. Słomka, 'Cost-sensitive classifier chains: Selecting low-cost features in multi-label classification', *Pattern Recognition*, **86**, 290–319, (2019).
[29] I. L. Vegting, M. van Beneden, M. H. H. Kramer, A. Abel Thijs, P. J. Kostense, and P. W. B. Nanayakkara, 'How to save costs by reducing unnecessary testing: Lean thinking in clinical practice', *European Journal of Internal Medicine*, **23**(1), 70 – 75, (2012).
[30] M. Zhang and Z. Zhou, 'A review on multi-label learning algorithms', *IEEE Transactions on Knowledge and Data Engineering*, **26**, 1819 – 1837, (2013).
[31] H. Zou, 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association*, **101**(476), 1418–1429, (2006).
[32] D. Zufferey, T. Hofer, J. Hennebert, M. Schumacher, R. Ingold, and S. Bromuri, 'Performance comparison of multi-label learning algorithms on clinical data for chronic diseases', *Computers in Biology and Medicine*, **65**, 34–43, (2015).