Pattern Recognition 90 (2019) 257–270

Contents lists available at ScienceDirect

Pattern Recognition

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

Noisy multi-label semi-supervised dimensionality reduction

Karl Øyvind Mikalsen[a,b,*], Cristina Soguero-Ruiz[b,c], Filippo Maria Bianchi[d,b], Robert Jenssen[d,b]

[a] Department of Mathematics and Statistics, UiT The Arctic University of Norway, Tromsø, Norway
[b] UiT Machine Learning Group, Norway
[c] Department of Signal Theory and Comm., Telematics and Computing, Universidad Rey Juan Carlos, Fuenlabrada, Spain
[d] Dept. of Physics and Technology, UiT, Tromsø, Norway

ARTICLE INFO

Article history:
Received 18 September 2018
Revised 21 December 2018
Accepted 25 January 2019
Available online 29 January 2019

Keywords:
Label noise
Multi-label learning
Semi-supervised learning
Dimensionality reduction

ABSTRACT

Noisy labeled data represent a rich source of information that often are easily accessible and cheap to obtain, but label noise might also have many negative consequences if not accounted for. How to fully utilize noisy labels has been studied extensively within the framework of standard supervised machine learning over a period of several decades. However, very little research has been conducted on solving the challenge posed by noisy labels in non-standard settings. This includes situations where only a fraction of the samples are labeled (semi-supervised) and each high-dimensional sample is associated with multiple labels. In this work, we present a novel semi-supervised and multi-label dimensionality reduction method that effectively utilizes information from both noisy multi-labels and unlabeled data. With the proposed *Noisy multi-label semi-supervised dimensionality reduction (NMLSDR)* method, the noisy multi-labels are denoised and unlabeled data are labeled simultaneously via a specially designed label propagation algorithm. NMLSDR then learns a projection matrix for reducing the dimensionality by maximizing the dependence between the enlarged and denoised multi-label space and the features in the projected space. Extensive experiments on synthetic data, benchmark datasets, as well as a real-world case study, demonstrate the effectiveness of the proposed algorithm and show that it outperforms state-of-the-art multi-label feature extraction algorithms.

1. Introduction

Supervised machine learning crucially relies on the accuracy of the *observed labels* associated with the training samples [1–10]. Observed labels may be corrupted and, therefore, they do not necessarily coincide with the true class of the samples. Such inaccurate labels are also referred to as *noisy* [2,4,11]. Label noise can occur for various reasons in real-world data, e.g. because of imperfect evidence, insufficient information, label-subjectivity or fatigue on the part of the labeler. In other cases, noisy labels may result from the use of frameworks such as anchor learning [12,13] or silver standard learning [14], which have received interest for instance in healthcare analytics [15,16]. A review of various sources of label noise can be found in [2].

In standard supervised machine learning settings, the challenge posed by noisy labels has been studied extensively. For example, many noise-tolerant versions of well-known classifiers have been proposed, including discriminant analysis [8,17], logistic regression [18], the k-nearest neighbor classifier [19], boosting algorithms [20,21], perceptrons [22,23], support vector machines [24], deep neural networks [7,25,26]. Others have proposed more general classification frameworks that are not restricted to particular classifiers [4,11].

However, very little research has been conducted on solving the challenge posed by noisy labels in non-standard settings, where the magnitude of the noisy label problem is increased considerably. Some examples of such a non-standard setting occur for instance within image analysis [27], document analysis [28], named entity recognition [29], crowdsourcing [30], or in the healthcare domain, used here as an illustrative case-study. Non-standard settings include (i) *Semi-supervised learning* [31,32], referring to a situation where only a few (noisy) labeled data points are available, making the impact of noise in those few labels more prevalent, and where information must also jointly be inferred from unlabeled data points. In healthcare, it may be realistic to obtain some labels through a (imperfect) manual labeling process, but the vast amount of data remains unlabeled; (ii) *Multi-label*

[*] Corresponding author at: Department of Mathematics and Statistics, Faculty of Science and Technology, UiT - The Arctic University of Norway, Tromsø N-9037, Norway.

E-mail address: karl.o.mikalsen@uit.no (K.Ø. Mikalsen).

https://doi.org/10.1016/j.patcog.2019.01.033
0031-3203/© 2019 Elsevier Ltd. All rights reserved.

learning [33–41], wherein objects may not belong exclusively to one category. This situation occurs frequently in a number of domains, including healthcare, where for instance a patient could suffer from multiple chronic diseases; (iii) High-dimensional data, where the abundance of features and the limited (noisy) labeled data, lead to a curse of dimensionality problem. In such situations, *dimensionality reduction* (DR) [42] is useful, either as a preprocessing step, or as an integral part of the learning procedure. This is a well-known challenge in health, where the number of patients in the populations under study frequently is small, but heterogeneous potential sources of data features from electronic health records for each patient may be enormous [43–46].

In this paper, and to the best of our knowledge, we propose the first noisy label, semi-supervised and multi-label DR machine learning method, which we call the *Noisy multi-label semi-supervised dimensionality reduction (NMLSDR)* method. Towards that end, we propose a label propagation method that can deal with noisy multi-label data. Label propagation [47–54], wherein one propagates the labels to the unlabeled data in order to obtain a fully labeled dataset, is one of the most successful and fundamental frameworks within semi-supervised learning. However, in contrast to many of these methods that clamp the labeled data, in our multi-label propagation method we allow the labeled part of the data to change labels during the propagation to account for noisy labels. In the second part of our algorithm we aim at learning a lower dimensional representation of the data by maximizing the feature-label dependence. Towards that end, similarly to other DR methods [55,56], we employ the Hilbert-Schmidt independence criterion (HSIC) [57], which is a non-parametric measure of dependence.

The NMLSDR method is a DR method, which is general and can be used in many different settings, e.g. for visualization or as a preprocessing step before doing classification. However, in order to test the quality of the NMLSDR embeddings, we (preferably) have to use some quantitative measures. For this purpose, a common baseline classifier such as the multi-label k-nearest neighbor (ML-kNN) classifier [58] has been applied to the low-dimensional representations of the data [59,60]. Even though this is a valid way to measure the quality of the embeddings, to apply a supervised classifier in a semi-supervised learning setting is not a realistic setup since one suddenly assumes that all labels are known (and correct). Therefore, as an additional contribution, we introduce a novel framework for semi-supervised classification of noisy multi-label data.

In our experiments, we compare NMLSDR to baseline methods on synthetic data, benchmark datasets, as well as a real-world case study, where we use it to identify the health status of patients suffering from potentially multiple chronic diseases. The experiments demonstrate that for partially and noisy labeled multi-label data, NMLSDR is superior to existing DR methods according to seven different multi-label evaluation metrics and the Wilcoxon statistical test.

In summary, the contributions of the paper are as follows.

- A new label noise-tolerant semi-supervised multi-label dimensionality reduction method based on dependence maximization.
- A novel framework for semi-supervised classification of noisy multi-label data.
- A comprehensive experimental section that illustrate the effectiveness of the NMLSDR on synthetic data, benchmark datasets and on a real-world case study.

The remainder of the paper is organized as follows. Related work is reviewed in Section 2. In Section 3, we describe our proposed NMLSDR method and the novel framework for semi-supervised classification of noisy multi-label data. Section 4 describes experiments on synthetic and benchmark datasets, whereas Section 5 is devoted to the case study where we study chronically ill patients. We conclude the paper in Section 6.

2. Related work

In this section we review related unsupervised, semi-supervised and supervised DR methods.[1]

Unsupervised DR methods do not exploit label information and can therefore straightforwardly be applied to multi-label data by simply ignoring the labels. For example, principal component analysis (PCA) aims to find the projection such that the variance of the input space is maximally preserved [62]. Other methods aim to find a lower dimensional embedding that preserves the manifold structure of the data, and examples of these include Locally linear embedding [63], Laplacian eigenmaps [64] and ISOMAP [65].

One of the most well-known supervised DR methods is linear discriminative analysis (LDA) [66], which aims at finding the linear projection that maximizes the within-class similarity and at the same time minimizes the between-class similarity. LDA has been extended to multi-label LDA (MLDA) in several different ways [67–71]. The difference between these methods basically consists in the way the labels are weighted in the algorithm. Following the notation in [71], wMLDAb [67] uses binary weights, wMLDAe [68] uses entropy-based weights, wMLDAc [69] uses correlation-based weights, wMLDAf [70] uses fuzzy-based weights, whereas wMLDAd [71] uses dependence-based weights.

Canonical correlation analysis (CCA) [72] is a method that maximizes the linear correlation between two sets of variables, which in the case of DR are the set of labels and the set of features derived from the projected space. CCA can be directly applied also for multi-labels without any modifications. Multi-label informed latent semantic indexing (MLSI) [73] is a DR method that aims at both preserving the information of inputs and capturing the correlations between the labels. In the Multi-label least square (ML-LS) method one extracts a common subspace that is assumed to be shared among multiple labels by solving a generalized eigenvalue decomposition problem [74].

In [55], a supervised method for doing DR based on dependence maximization [57] called Multi-label dimensionality reduction via dependence maximization (MDDM) was introduced. MDDM attempts to maximize the feature-label dependence using the Hilbert–Schmidt independence criterion and was originally formulated in two different ways. MDDMp is based on orthonormal projection directions, whereas MDDMf makes the projected features orthonormal. Yu et al. showed that MDDMp can be formulated using least squares and added a PCA term to the cost function in a new method called Multi-label feature extraction via maximizing feature variance and feature-label dependence simultaneously (MVMD) [56].

The most closely related existing DR methods to NMLSDR are the semi-supervised multi-label methods. The Semi-supervised dimension reduction for multi-label classification method (SSDR-MC) [75], Coupled dimensionality reduction and classification for supervised and semi-supervised multilabel learning [76], and Semisupervised multilabel learning with joint dimensionality reduction [77] are semi-supervised multi-label methods that simultaneously learn a classifier and a low dimensional embedding.

Other semi-supervised multi-label DR methods are semi-supervised formulations of the corresponding supervised multi-label DR method. Blascho et al. introduced semi-supervised CCA based on Laplacian regularization [78]. Several different semi-supervised formulations of MLDA have also been

[1] DR may be obtained both by feature extraction, i.e. by a data transformation, and by feature selection [61]. Here, we refer to DR in the sense of feature extraction.

proposed. Multi-label dimensionality reduction based on semi-supervised discriminant analysis (MSDA) adds two regularization terms computed from an adjacency matrix and a similarity correlation matrix, respectively, to the MLDA objective function [79]. In the Semi-supervised multi-label dimensionality reduction (SSMLDR) [59] method one does label propagation to obtain soft labels for the unlabeled data. Thereafter the soft labels of all data are used to compute the MLDA scatter matrices. An other extension of MLDA is Semi-supervised multi-label linear discriminant analysis (SMLDA) [80], which later was modified and renamed Semi-supervised multi-label dimensionality reduction based on dependence maximization (SMDRdm) [60]. In SMDRdm the scatter matrices are computed based on only labeled data. However, a HSIC term is also added to the familiar Rayleigh quotient containing the two scatter matrices, which is computed based on soft labels for both labeled and unlabeled data obtained in a similar way as in SSMLDR.

Common to all these methods is that none of them explictly assume that the labels can be noisy. In SSMLDR and SMDRdm, the labeled data are clamped during the label propagation and hence cannot change. Moreover, these two methods are both based on LDA, which is known heavily affected by outliers, and consequently also wrongly labeled data [81–83].

3. The NMLSDR method

We start this section by introducing notation and the setting for noisy multi-label semi-supervised linear feature extraction, and thereafter elaborate on our proposed NMLSDR method.

3.1. Problem statement

Let $\{x_i\}_{i=1}^n$ be a set of n D-dimensional data points, $x_i \in \mathbb{R}^D$. Assume that the data are ordered such that the l first of the data points are labeled and u are unlabeled, $l + u = n$. Let X be a $n \times D$ matrix with the data points as row vectors.

Assume that the number of classes is C and let $Y_i^L \in \{0, 1\}^C$ be the label-vector of data point x_i, $i = 1, \dots, l$. The elements are given by $Y_{ic}^L = 1,\ c = 1, \dots, C$ if data point x_i belongs to the cth class and $Y_{ic}^L = 0$ otherwise. Define the label matrix $Y^L \in \{0,\ 1\}^{l\times C}$ as the matrix with the known label-vectors $Y_i^L,\ i = 1, \dots, l$ as row vectors and let $Y^U \in \{0,\ 1\}^{u\times C}$ be the corresponding label matrix of the unknown labels.

The objective of linear feature extraction is to learn a projection matrix $P \in \mathbb{R}^{D\times d}$ that maps a data point in the original feature space $x \in \mathbb{R}^D$ to a lower dimensional representation $z \in \mathbb{R}^d$,

$$z = P^T x, \tag{1}$$

where $d < D$ and P^T denotes the transpose of the matrix P.

In our setting, we assume that the label matrix Y^L is potentially noisy and that Y^U is unknown. The first part of our proposed NMLSDR method consists of doing label propagation in order to learn the labels Y^U and update the estimate of Y^L. We do this by introducing soft labels $F \in \mathbb{R}^{n\times C}$ for the label matrix $Y = \binom{Y^L}{Y^U}$, where F_{ic} represents the probability that data point x_i belong to the cth class. We obtain F with label propagation and thereafter use F to learn the projection matrix P. However, we start by explaining our label propagation method.

3.2. Label propagation using a neighborhood graph

The underlying idea of label propagation is that similar data points should have similar labels. Typically, the labels are propagated using a neighborhood graph [47]. Here, inspired by [84], we formulate a label propagation method for multi-labels that is robust to noise. The method is as follows.

Step 1. First, a neighbourhood graph is constructed. The graph is described by its adjacency matrix W, which can be designed e.g. by setting the entries to

$$W_{ij} = \exp(-\sigma^{-2}\|x_i - x_j\|^2), \tag{2}$$

where $\|x_i - x_j\|$ is the Euclidean distance between the datapoints x_i and x_j, and σ is a hyperparameter. Alternatively, one can use the Euclidian distance to compute a k-nearest neighbors (kNN) graph where the entries of W are given by

$$W_{ij} = \begin{cases} 1, & \text{if } x_i \text{ among } x_j\text{'s } k\text{NN or } x_j \text{ among } x_i\text{'s } k\text{NN} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Step 2. Symmetrically normalize the adjacency matrix W by letting

$$\tilde{W} = D^{-1/2} W D^{-1/2}, \tag{4}$$

where D is a diagonal matrix with entries given by $d_{ii} = \sum_{k=1}^n W_{ik}$.

Step 3. Calculate the stochastic matrix

$$T = \tilde{D}^{-1}\tilde{W}, \tag{5}$$

where $\tilde{d_{ii}} = \sum_{k=1}^n \tilde{W}_{ik}$. The entry T_{ij} can now be considered as the probability of a transition from node i to node j along the edge between them.

Step 4. Compute soft labels $F \in \mathbb{R}^{n\times C}$ by iteratively using the following update rule

$$F(t+1) = I_\alpha T F(t) + (I - I_\alpha)Y, \tag{6}$$

where I_α is a $n \times n$ diagonal matrix with the hyperparameters α_i, $0 \le \alpha_i < 1$, on the diagonal. To initialize F, we let $F(0) = Y$, where the unlabeled data are set to $Y_{ic}^U = 0,\ c = 1, \dots, C$.

3.2.1. Discussion

Setting $\alpha_i = 0$ for the labeled part of the data corresponds to clamping of the labels. However, this is not what we aim for in the presence of noisy labels. Therefore, a crucial property of the proposed framework is to set $\alpha_i > 0$ such that the labeled data can change labels during the propagation.

Moreover, we note that our extension of label propagation to multi-labels is very similar to the single-label variant introduced in [84], with the exception that we do not add the outlier class, which is not needed in our case. In other extensions to multi-label propagation [59,60], the label matrix Y is normalized such that the rows sum to 1, which ensures that the output of the algorithm F also has rows that sum to 1. In the single-label case this makes sense in order to maintain the interpretability of probabilities. However, in the multi-label case the data points do not necessarily exclusively belong to a single class. Hence, the requirement $\sum_c F_{ic} = 1$ does not make sense since then x_i can maximally belong to one class if one think of F as a probability and require the probability to be 0.5 or higher in order to belong to a class.

On the other hand, in our case, a simple calculation shows that $0 \le F_{ic}(t+1) \le 1$:

$$\begin{aligned} F_{ic}(t+1) &= \alpha_i \sum_{m=1}^n T_{im} F_{mc}(t) + (1-\alpha_i)Y_{ic} \\ &\le \alpha_i \sum_{m=1}^n T_{im} + (1-\alpha_i) = \alpha_i + (1-\alpha_i) = 1, \end{aligned} \tag{7}$$

since $F_{ic}(t) \le 1$ and $Y_{ic} \le 1$. However, we do not necessarily have that $\sum_c F_{ic} = 1$.

From matrix theory it is known that, given that $I - I_\alpha T$ is nonsingular, the solution of the linear iterative process (6) converges to the solution of

$$(I - I_\alpha T)F = (I - I_\alpha)Y, \tag{8}$$

for any initialization $F(0)$ if and only if $I_\alpha T$ is a *convergent matrix* [85] (spectral radius $\rho(I_\alpha T) < 1$). $I_\alpha T$ is obviously convergent if $0 \leq \alpha_i < 1$ $\forall i$. Hence, we can find the soft labels F by solving the linear system given by Eq. (8).

Moreover, F_{ic} can be interpreted as the probability that datapoint x_i belongs to class c, and therefore, if one is interested in hard label assignments, $\tilde{Y}$, these can be found by letting $\tilde{Y}_{ic} = 1$ if $F_{ic} > 0.5$ and $\tilde{Y}_{ic} = 0$ otherwise.

3.3. Dimensionality reduction via dependence maximization

In this section we explain how we use the labels obtained using label propagation to learn the projection matrix P.

The motivation behind dependence maximization is that there should be a relation between the features and the label of an object. This should be the case also in the projected space. Hence, one should try to maximize the dependence between the feature similarity in the projected space and the label similarity. A common measure of such dependence is the Hilbert–Schmidt independence criterion (HSIC) [57], defined by

$$HSIC(X, Y) = \frac{1}{(n-1)^2} tr(KHLH), \tag{9}$$

where tr denotes the trace of a matrix. $H \in \mathbb{R}^{n \times n}$ is given by $H_{ij} = \delta_{ij} - n^{-1}$, where $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ otherwise. K is a kernel matrix over the feature space, whereas L is a kernel computed over the label space.

Let the projection of x be given by the projection matrix $P \in \mathbb{R}^{D \times d}$ and function $\Phi : \mathbb{R}^D \to \mathbb{R}^d$, $\Phi(x) = P^T x$. We select a linear kernel over the feature space, and therefore the kernel function is given by

$$\mathcal{K}(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle = \langle P^T x_i, P^T x_j \rangle = P^T x_i x_j^T P \tag{10}$$

Hence, given data $\{x_i\}_{i=1}^n$, the kernel matrix can be approximated by $K = XP^T PX^T$.

The kernel over the label space, $\mathcal{L}$, is given via the labels $y_i \in \{0, 1\}^C$. One possible such kernel is the linear kernel

$$\mathcal{L}(y_i, y_j) = \langle y_i, y_j \rangle. \tag{11}$$

However, in our semi-supervised setting, some of the labels are unknown and some are noisy. Hence, the kernel $\mathcal{L}$ cannot be computed. In order to enable DR in our non-standard problem, we propose to estimate the kernel using the labels obtained via our label propagation method. For the part of the data that was labeled from the beginning we use the hard labels, $\tilde{Y}^L$, obtained from the label propagation, whereas for the unlabeled part we use the soft labels, F^U. Hence, the kernel is approximated via $L = \tilde{F}\tilde{F}^T$, where $\tilde{F} = \begin{pmatrix} \tilde{Y}^L \\ F^U \end{pmatrix}$. The reason for using the hard labels obtained from label propagation for the labeled part is that we want some degree of certainty for those labels that change during the propagation (if the soft label F_{ic}^L changes with less than 0.5 from its initial value 0 or 1 during the propagation, the hard label Y_{ic}^L does not change).

The constant term, $(n-1)^{-2}$, in Eq. (9) is irrelevant in an optimization setting. Hence, by inserting the estimates of the kernels into Eq. (9), the following objective function is obtained,

$$\Psi(P) = tr(HXP^T PX^T H\tilde{F}\tilde{F}^T) = tr(P^T X^T H\tilde{F}\tilde{F}^T HXP). \tag{12}$$

Note that the matrix $X^T H\tilde{F}\tilde{F}^T HX$ is symmetric. Hence, by requiring that the projection directions are orthogonal and that the new dimensionality is d, the following optimization problem is obtained

$$\arg\max_P \Psi(P) = \arg\max_P tr(P^T (X^T H\tilde{F}\tilde{F}^T HX)P), \tag{13}$$
$$s.t.\, P \in \mathbb{R}^{D \times d},\ PP^T = I.$$

As a consequence of the Courant-Fisher characterization [86], it follows that the maximum is achieved when P is an orthonormal basis corresponding to the d largest eigenvalues. Hence, P can be found by solving the eigenvalue problem

$$X^T H\tilde{F}\tilde{F}^T HXP = \Lambda P. \tag{14}$$

The dimensionality of the projected space, d, is upper bounded by the rank of $\tilde{F}\tilde{F}^T$, which in turn is upper bounded by the number of classes C. Hence, d cannot be set larger than C. The pseudo-code of the NMLSDR method is shown in Algorithm 1.

Algorithm 1 Pseudo-code for NMLSDR.

```
Require: $X$ : $n \times D$ feature matrix, $Y$ : $n \times C$ label matrix, hyper-parameters $k$, $I_\alpha$ and $d$.
  Initialize $F$ by letting $F(0) = Y$, where the unlabeled data are set to $Y_{ic}^U = 0$, $c = 1, \ldots, C$.
  Construct a neighbourhood graph by calculating the adjacency matrix $W$ using Eqs. (2) or (3).
  Symmetrically normalize the adjacency matrix $W$ by letting $\tilde{W} = D^{-1/2} W D^{-1/2}$,
  Calculate the stochastic matrix $T = \tilde{D}^{-1}\tilde{W}$, where $\tilde{d}_{ii} = \sum_{k=1}^n \tilde{W}_{ik}$.
  Solve the linear system $(I - I_\alpha T)F = (I - I_\alpha)Y$.
  Compute $\tilde{F}$.
  Construct the matrix $X^T H\tilde{F}\tilde{F}^T HX$.
  Eigendecompose $X^T H\tilde{F}\tilde{F}^T HX$ and construct projection matrix $P \in \mathbb{R}^{D \times d}$.
Ensure: Projection $P$ : $\mathbb{R}^D \to \mathbb{R}^d$
```

3.4. Semi-supervised classification for noisy multi-label data

The multi-label k-nearest neighbor (ML-kNN) classifier [58] is a widely adopted classifier for multi-label classification. However, similarly to many other classifiers, its performance can be hampered if the dimensionality of the data is too high. Moreover, the ML-kNN classifier only works in a completely supervised setting. To resolve these problems, as an additional contribution of this work, we introduce a novel framework for semi-supervised classification of noisy multi-label data, consisting of two steps. In the first step, we compute a low dimensional embedding using NMLSDR. The second step consists of applying a semi-supervised ML-kNN classifier. For this classifier we use our label propagation method on the learned embedding to obtain a fully labeled dataset, and thereafter apply the ML-kNN classifier.

4. Experiments

In this paper, we have proposed a method for computing a low-dimensional embedding of noisy, partially labeled multi-label data. However, it is not a straightforward task to measure how well the method works. Even though the method is definitely relevant to real-world problems (illustrated in the case study in Section 5), the framework cannot be directly applied to most multi-label benchmark datasets since most of them are completely labeled, and the labels are assumed to be clean. Moreover, the NMLSDR provides a low dimensional embedding of the data, and we need a way to measure how good the embedding is. If the dimensionality is 2 or 3, this can to some degree be done visually by plotting the embedding. However, in order to quantitatively measure the quality and simultaneously maintain a realistic setup, we will apply our proposed end-to-end framework for semi-supervised classification and dimensionality reduction. In our experiments, this realistic semi-supervised setup will be applied in an illustrative example on synthetic data and in the case study.

A potential disadvantage of using a semi-supervised classifier, is that it does not necessarily isolate effect of the DR method that is used to compute the embedding. For this reason, we will also test our method on some benchmark datasets, but in order to keep everything coherent, except for the method used to compute the embedding, we compute the embedding using NMLSDR and baseline DR methods based on only the noisy and partially labeled multi-label training data. Thereafter, we assume that the true multi-labels are available when we train the ML-kNN classifier on the embeddings.

The remainder of this section is organized as follows. First we describe the performance measures we employed, baseline DR methods, and how we select hyper-parameters. Thereafter we provide an illustrative example on synthetic data, and secondly experiments on the benchmark data. The case study is described in the next section.

4.1. Evaluation metrics

Evaluation of performance is more complicated in a multi-label setting than for traditional single-labels. In this work, we decide use the seven different evaluation criteria that were employed in [55], namely Hamming loss (HL), Macro F1-score (MaF1), Micro F1 (MiF1), Ranking loss (RL), Average precision (AP), One-error (OE) and Coverage (Cov).

HL simply evaluates the number of times there is a mismatch between the predicted label and the true label, i.e.

$$HL = \sum_{i=1}^{n} \frac{\|\hat{y}_i \oplus y_i\|_1}{nC}, \tag{15}$$

where $\hat{y}_i$ denotes the predicted label vector of data point x_i and $\oplus$ is the XOR-operator. MaF1 is obtained by first computing the F1-score for each label, and then averaging over all labels.

$$MaF1 = \frac{1}{C}\sum_{c=1}^{C} \frac{2\sum_{i=1}^{n}\hat{y}_{ic}y_{ic}}{\sum_{i=1}^{n}\hat{y}_{ic} + \sum_{i=1}^{n}y_{ic}}, \tag{16}$$

MiF1 calculates the F1 score on the predictions of different labels as a whole,

$$MiF1 = \frac{2\sum_{i=1}^{n}\sum_{c=1}^{C}\hat{y}_{ic}y_{ic}}{\sum_{i=1}^{n}\sum_{c=1}^{C}\hat{y}_{ic} + \sum_{i=1}^{n}\sum_{c=1}^{C}y_{ic}}, \tag{17}$$

We note that HL, MiF1 and MaF1 are computed based on hard labels assignments, whereas the four other measures are computed based on soft labels. In all of our experiments, we obtain the hard labels by putting a threshold at 0.5.

RL computes the average ratio of reversely ordered label pairs of each data point. AP evaluates the average fraction of relevant labels ranked higher than a particular relevant label. OE gives the ratio of data points where the most confident predicted label is wrong. Cov gives an average of how far one needs to go down on the list of ranked labels to cover all the relevant labels of the data point. For a more detailed description of these measures, we point the interested reader to [87].

In this work, we modify four of the evaluation metrics such that all of them take values in the interval [0, 1] and “higher always is better”. Hence, we define

$$HL' = 1 - HL, \tag{18}$$

$$RL' = 1 - RL, \tag{19}$$

$$OE' = 1 - OE, \tag{20}$$

and normalized coverage (Cov’) by

$$Cov' = 1 - Cov/(C-1). \tag{21}$$

4.2. Baseline dimensionality reduction methods

In this work, we consider the following other DR methods: CCA, MVMD, MDDMp, MDDMf and four variants of MLDA, namely wMLDAb, wMLDAe, wMLDAc and wMLDAd. These methods are supervised and require labeled data, and are therefore trained only on the labeled part of the training data. In addition, we compare to a semi-supervised method, SSMLDR, which we adapt to noisy multi-labels by using the label propagation algorithm we propose in this paper instead of the label propagation method that was originally proposed in SSMLDR. We note that the computational complexity of NMLSDR and the all the baselines is of the same order as all of them require a step involving eigendecomposition.

4.3. Hyper-parameter selection and implementation settings

For the ML-kNN classifier we set $k = 10$. The effect of varying the number of neighbors will be left for further work. In order to learn the NMLSDR embedding we use a kNN-graph with $k = 10$ and binary weights. Moreover, we set $\alpha_i = 0.6$ for labeled data and $\alpha_i = 0.999$ for unlabeled data. By doing so, one ensures that an unlabeled datapoint is not affected by its initial value, but gets all contribution from the neighbors during the propagation. All experiments are run in Matlab using an Ubuntu 16.04 64-bit system with 16 GB RAM and an Intel Core i7-7500U processor.

4.4. Illustrative example on synthetic toy data

Dataset description. To test the framework in a controlled experiment, a synthetic dataset is created as follows.

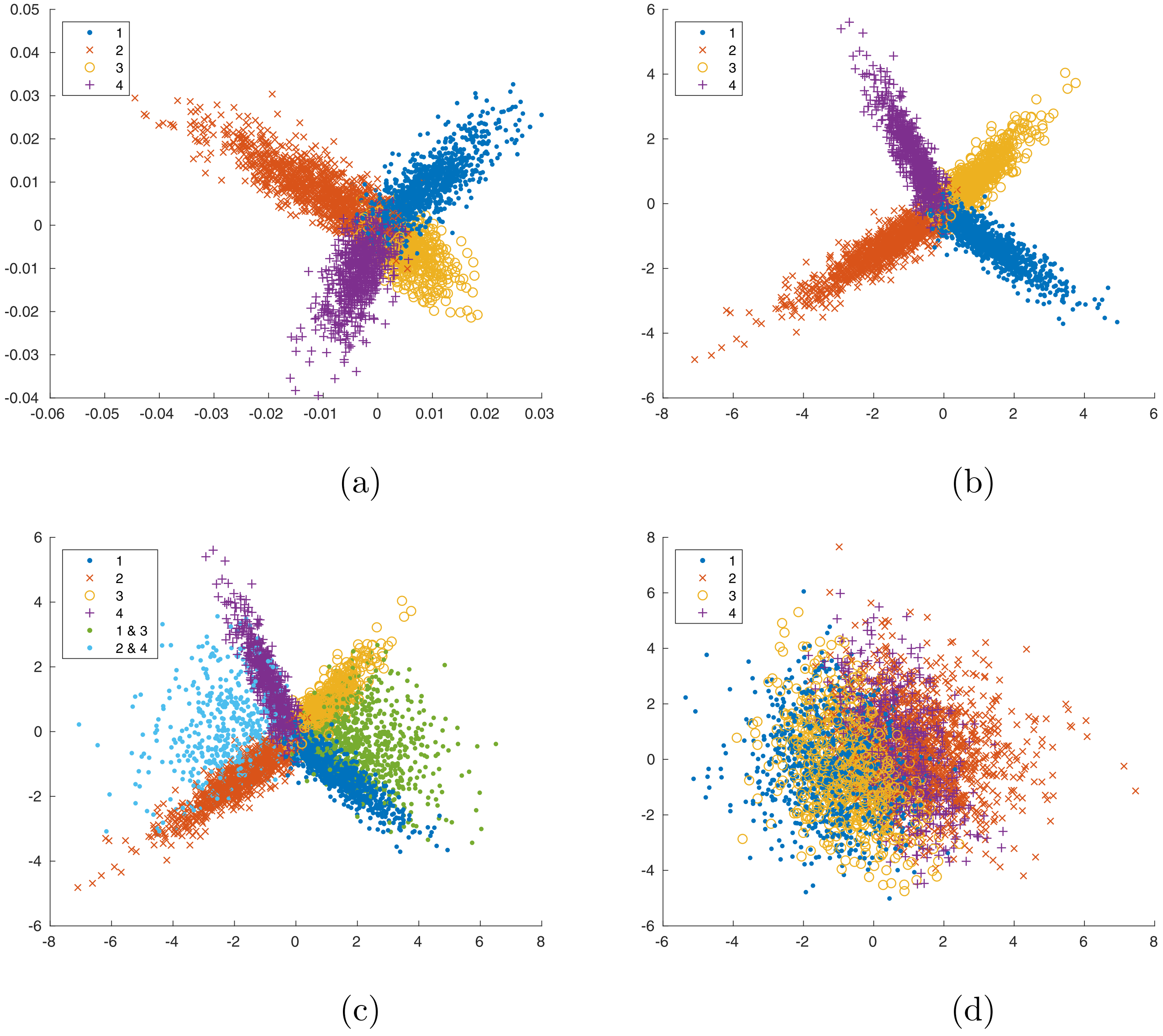
A dataset of size 8000 samples is created, where each of the data points has dimensionality 320. The number of classes is set to 4, and we generate 2000 samples from each class. 30% from class 1 also belong to class 2, and vice versa. 20% from class 2 also belong to class 3 and vice versa, whereas 25% from class 3 also belong to class 4 and vice versa.

A sample from class i is generated by randomly letting 10% of the features in the interval $\{20(i-1)+1, \ldots, 20i\}$ take a random integer value between 1 and 10. Since there are 4 classes, this means that the first 80 features are directly dependent on the class-membership.

For the remaining 240 features we consider 20 of them at the time. We randomly select 50% of the 8000 samples and randomly let 20% of the 20 features take a random integer value between 1 and 10. We repeat this procedure for the 12 different sets of 20 features $\{20(i-1)+1, \ldots, 20i\}$, $i = 5, 6, \ldots, 16$.

All features that are not given a value using the procedure described above are set to 0. Noise is injected into the labels by randomly flipping a fraction $p = 0.1$ of the labels and we make the data partially labeled by removing 50% of the labels. 2000 of the samples are kept aside as an independent test set. We note that noisy labels are often easier and cheaper to obtain than true labels and it is therefore not unreasonable that the fraction of labeled examples is larger than what it commonly is in traditional semi-supervised learning settings.

Results. We apply the NMLSDR method in combination with the semi-supervised ML-kNN classifier as explained above and compare to SSMLDR. We create two baselines by, for both of these methods, using a different value for the hyperparameter α_i for the labeled part of the data, namely 0, which corresponds to clamping. We denote these two baselines by SSMLDR* and NMLSDR*. In addition, we compare to baselines that only utilize the labeled part of the data, namely the supervised DR methods explained above in combination with a ML-kNN classifier. The data is standardized to



(a)

(b)

(c)

(d)

Fig. 1. 3 dimensional embedding of the synthetic dataset obtained using (a) SSMLDR; (b) NMLSDR; (c) NMLSDR with multi-classes included; and (d) PCA.

0 mean and 1 in standard deviation and we let the dimensionality of the embedding be 3.

Fig. 1a and b show the embeddings obtained obtained using SSMLDR and NMLSDR, respectively. For ivisualization purposes, we have only plotted those datapoints that exclusively belong to one class. In Fig. 1c, we have added two of the multi-classes for the NMLSDR embedding. For comparison, we also added the embedding obtained using PCA in Fig. 1d. As we can see, in the PCA embedding the classes are not separated from each other, whereas in the NMLSDR and SSMLDR embeddings the classes are aligned along different axes. It can be seen that the classes are better separated and more compact in the NMLSDR embedding than the SSMLDR embedding. Fig. 1c shows that the data points that belong to multiple classes are placed where they naturally belong, namely between the axes corresponding to both of the classes they are member of.

Table 1 shows the results obtained using the different methods on the synthetic dataset. As we can see, our proposed method gives the best performance for all metrics. Moreover, NMLSDR with $\alpha_i^L = 0$, which corresponds to clamping of the labeled data during label propagation gives the second best results but cannot compete with our proposed method, in which the labels are allowed to change during the propagation to account for noisy labels. We also note that, even though the SSMLDR improves the MLDA approaches that are based on only the labeled part of the data, it gives results that are considerably worse than NMLSDR.

Table 1
Results in terms of 7 metrics for the synthetic dataset.

Method	HL′	RL′	AP	OE′	Cov′	MaF1	MiF1
CCA	0.863	0.884	0.898	0.852	0.816	0.787	0.785
MVMD	0.906	0.912	0.924	0.897	0.836	0.850	0.849
MDDMp	0.906	0.911	0.924	0.897	0.836	0.851	0.850
MDDMf	0.859	0.888	0.900	0.855	0.819	0.785	0.783
wMLDAb	0.844	0.871	0.885	0.831	0.807	0.754	0.750
wMLDAe	0.864	0.885	0.899	0.855	0.818	0.790	0.788
wMLDAc	0.865	0.887	0.900	0.857	0.818	0.787	0.785
wMLDAd	0.869	0.891	0.907	0.869	0.822	0.788	0.786
SSMLDR*	0.863	0.883	0.899	0.859	0.814	0.796	0.793
SSMLDR	0.879	0.898	0.910	0.871	0.827	0.817	0.814
NMLSDR*	0.907	0.919	0.929	0.903	0.842	0.861	0.859
NMLSDR	**0.913**	**0.925**	**0.935**	**0.912**	**0.846**	**0.868**	**0.866**

Table 2
Description of benchmark datasets considered in our experiments.

Dataset	Domain	Train instances	Test instances	Attributes	Labels	Cardinality
Birds	audio	322	323	260	19	1.06
Corel	scene	5188	1744	500	153	2.87
Emotions	music	391	202	72	6	1.81
Enron	text	1123	579	1001	52	3.38
Genbase	biology	463	199	99	25	1.26
Medical	text	645	333	1161	39	1.24
Scene	scene	1211	1196	294	6	1.06
Tmc2007	text	3000	7077	493	22	2.25
Toy	synthetic	6000	2000	320	4	1.38
Yeast	biology	1500	917	103	14	4.23

Table 3
Performance in terms of 1 - Hamming loss (HL') across 10 different benchmark datasets.

	CCA	MVMD	MDDMp	MDDMf	wMLDAb	wMLDAe	wMLDAc	wMLDAd	SSMLDR	NMLSDR
Birds	0.947	0.950	0.950	0.947	0.948	0.949	0.949	0.949	0.949	**0.951**
Corel	**0.980**	**0.980**	**0.980**	**0.980**	**0.980**	**0.980**	**0.980**	**0.980**	**0.980**	**0.980**
Emotions	0.715	0.771	0.778	0.711	0.696	0.714	0.709	0.717	0.786	**0.787**
Enron	0.941	**0.950**	**0.950**	0.942	0.941	0.941	0.941	0.940	0.938	**0.950**
Genbase	0.989	0.996	0.996	0.988	0.990	0.991	0.988	0.989	0.994	**0.997**
Medical	**0.976**	0.974	0.974	**0.976**	0.974	0.975	0.975	**0.976**	0.966	0.975
Scene	0.810	0.899	**0.900**	0.809	0.810	0.814	0.817	0.810	0.873	0.897
Tmc2007	0.914	0.928	0.928	0.912	0.911	0.911	0.911	0.916	0.922	**0.929**
Toy	0.836	0.894	0.894	0.839	0.821	0.831	0.831	0.854	0.861	**0.903**
Yeast	0.780	0.791	0.790	0.782	0.785	0.783	0.781	0.781	**0.793**	**0.793**
#Best values	2	2	3	2	1	1	1	2	2	**8**
Wilcoxon	2.0	7.0	7.5	2.5	2.0	3.0	2.5	3.5	6.0	**9.0**

4.5. Benchmark datasets

Experimental setup. We consider the following benchmark datasets[2]: Birds, Corel, Emotions, Enron, Genbase, Medical, Scene, Tmc2007 and Yeast. We also add our synthetic toy dataset as a one of our benchmark datasets (described in Section 4.4). These datasets are shown in Table 2, along with some useful characteristics. In order to be able to apply our framework to the benchmark datasets, we randomly flip 10% of the labels to generate noisy labels and let 30% of the data points training sets be labeled. All datasets are standardized to zero mean and standard deviation one.

We apply the DR methods to the partially and noisy labeled multi-label training sets in order to learn the projection matrix P, which in turn is used to map the D-dimensional training and test sets to a $d-$dimensional representation. d is set as large as possible, i.e. to $C-1$ for the MLDA-based methods and C for the other methods. Then we train a ML-kNN classifier using the low-dimensional training sets, assuming that the true multi-labels are known and validate the performance on the low-dimensional test sets.

In total we are evaluating the performance over 10 different datasets and across 7 different performance measures for all the feature extraction methods we use. Hence, to investigate which method performs better according to the different metrics, we also report the number of times each method gets the highest value of each metric. In addition, we compare all pairs of methods by using a Wilcoxon signed rank test with 5% significance level [88]. Similarly to [71], if method A performs better than B according to the test, A is assigned the score 1 and B the score 0. If the null hypothesis (method A and B perform equally) is not rejected, both A and B are assigned an equal score of 0.5.

Results. Table 3 shows results in terms of HL'. NMLSDR gets best HL'-score for eight of the datasets and achieves a maximal Wilcoxon score, i.e performs statistically better than all nine other methods according to the test at a 5% significance level. The second best method MDDMp gets the highest HL' score for three datasets and Wilcoxon score of 7.5. From Table 4 we see that NMLSDR achieves the highest RL'-score seven times and a Wilcoxon score of 8.5. The second best method is MVMD, which obtains three of the highest RL' values and a Wilcoxon score of 8.0.

Table 5 shows performance in terms of AP. The highest AP score is achieved for NMLSDR for eight datasets and it gets a maximal Wilcoxon score of 9.0. According to the Wilcoxon score second place is tied between MVMD and MDDMp. However, MVMD gets the highest AP score for two datasets, whereas MDDMp does not get the highest score for any of them. OE' is presented in Table 6. We can see that NMLSDR gets a maximal Wilcoxon score and the highest OE' score for seven datasets. MVMD is number two with a Wilcoxon score of 8.0 and two best values.

Table 7 shows Cov'. NMLSDR gets a maximal Wilcoxon score and the highest Cov' value for seven datasets. Despite that MVMD gets the highest Cov' for three datasets and MDDMp for none of the datasets, the second best Wilcoxon score is 7.5 and tied between MVMD and MDDMp. MaF1 is shown in Table 8. The best method, which is our proposed method gets a maximal Wilcoxon score and the highest MaF1 value for six datasets. Table 9 shows MiF1. NMLSDR achieves 8.5 in Wilcoxon score and has the highest MiF1 score for seven datasets.

In total, NMLSDR consistently gives the best performance for all seven evaluation metrics. Moreover, in order to summarize our findings, we compute the mean Wilcoxon score across all seven performance metrics and plot the result in Fig. 2. If we sort these results, we get NMLSDR (8.86), MVMD (7.64), MDDMp (7.43), wMLDAd (4.43), MDDMf (4.21), SSMLDR (3.79), CCA (2.79), wMLDAe (2.71) and wMLDAb/wMLDAc (1.57). The best method, which is our proposed method, gets a mean value that is 1.22 higher than number two. The second best method is MVMD, slightly better than MDDMp. The best MLDA-based method is wMLDAd, which is ranked 4th, however, with a much lower mean value than the three best methods. The semi-supervised extension of MLDA

[2] Downloaded from mulan.sourceforge.net/datasets-mlc.html.

Table 4
Performance in terms of 1 - Ranking loss (RL′) across 10 different benchmark datasets.

	CCA	MVMD	MDDMp	MDDMf	wMLDAb	wMLDAe	wMLDAc	wMLDAd	SSMLDR	NMLSDR
Birds	0.715	0.766	0.767	0.734	0.709	0.718	0.719	0.725	0.681	**0.771**
Corel	0.800	0.808	0.808	0.800	0.799	0.799	0.800	0.800	0.801	**0.814**
Emotions	0.695	0.824	0.824	0.709	0.693	0.700	0.676	0.714	0.829	**0.845**
Enron	0.894	0.911	0.911	0.893	0.893	0.892	0.891	0.893	0.883	**0.914**
Genbase	0.993	0.995	0.995	0.993	0.994	0.992	0.992	0.991	0.995	**1.000**
Medical	0.925	**0.952**	0.949	0.925	0.916	0.921	0.919	0.945	0.856	0.946
Scene	0.585	**0.900**	0.898	0.629	0.574	0.583	0.572	0.616	0.853	0.898
Tmc2007	0.831	0.906	0.906	0.830	0.830	0.830	0.831	0.847	0.872	**0.910**
Toy	0.871	0.909	0.909	0.870	0.849	0.865	0.861	0.888	0.887	**0.926**
Yeast	0.806	**0.820**	0.819	0.811	0.810	0.809	0.806	0.803	0.818	0.816
#Best values	0	3	0	0	0	0	0	0	0	**7**
Wilcoxon	3.0	8.0	7.5	4.5	1.5	2.0	2.0	5.0	3.0	**8.5**

Table 5
Performance in terms of Average precision (AP) across 10 different benchmark datasets.

	CCA	MVMD	MDDMp	MDDMf	wMLDAb	wMLDAe	wMLDAc	wMLDAd	SSMLDR	NMLSDR
Birds	0.389	0.499	0.500	0.426	0.374	0.392	0.379	0.424	0.357	**0.502**
Corel	0.260	0.277	0.277	0.261	0.265	0.263	0.263	0.268	0.266	**0.288**
Emotions	0.669	0.781	0.773	0.686	0.672	0.687	0.666	0.704	0.799	**0.808**
Enron	0.592	0.669	0.670	0.583	0.584	0.582	0.580	0.578	0.526	**0.675**
Genbase	0.963	0.990	0.993	0.964	0.960	0.968	0.963	0.969	0.984	**0.997**
Medical	0.673	0.722	0.716	0.666	0.644	0.674	0.669	0.723	0.446	**0.725**
Scene	0.491	**0.836**	0.835	0.534	0.481	0.488	0.475	0.521	0.781	0.834
Tmc2007	0.584	0.714	0.713	0.587	0.579	0.576	0.577	0.623	0.662	**0.721**
Toy	0.882	0.921	0.921	0.880	0.862	0.880	0.875	0.900	0.897	**0.933**
Yeast	0.732	**0.748**	0.747	0.731	0.733	0.733	0.729	0.725	0.745	0.741
#Best values	0	2	0	0	0	0	0	0	0	**8**
Wilcoxon	3.5	7.5	7.5	4.0	1.0	3.5	1.0	5.0	3.0	**9.0**

Table 6
Performance in terms of 1 - One error (OE') across 10 different benchmark datasets.

	CCA	MVMD	MDDMp	MDDMf	wMLDAb	wMLDAe	wMLDAc	wMLDAd	SSMLDR	NMLSDR
Birds	0.273	**0.419**	0.407	0.314	0.250	0.273	0.250	0.297	0.203	**0.419**
Corel	0.250	0.261	0.262	0.252	0.255	0.254	0.253	0.267	0.260	**0.283**
Emotions	0.535	0.673	0.644	0.564	0.535	0.589	0.550	0.589	0.718	**0.728**
Enron	0.620	0.762	0.762	0.610	0.587	0.604	0.606	0.579	0.544	**0.765**
Genbase	0.950	0.990	**0.995**	0.955	0.935	0.960	0.950	0.965	0.980	**0.995**
Medical	0.583	0.607	0.592	0.589	0.538	0.583	0.577	0.628	0.323	**0.619**
Scene	0.265	**0.732**	0.729	0.319	0.258	0.264	0.247	0.303	0.656	0.727
Tmc2007	0.527	0.650	0.648	0.531	0.523	0.519	0.516	0.578	0.604	**0.656**
Toy	0.821	0.888	0.887	0.819	0.785	0.821	0.811	0.850	0.849	**0.903**
Yeast	**0.760**	0.755	0.749	0.740	0.747	0.751	0.748	0.744	0.751	0.739
#Best values	1	2	1	0	0	0	0	1	0	**7**
Wilcoxon	3.5	8.0	7.0	4.0	1.0	3.5	1.0	5.0	3.0	**9.0**

Table 7
Performance in terms of 1 - Normalized coverage (Cov′) across 10 different benchmark datasets.

	CCA	MVMD	MDDMp	MDDMf	wMLDAb	wMLDAe	wMLDAc	wMLDAd	SSMLDR	NMLSDR
Birds	0.821	0.851	0.852	0.830	0.818	0.824	0.824	0.831	0.808	**0.860**
Corel	0.601	0.617	0.617	0.603	0.600	0.599	0.601	0.603	0.603	**0.628**
Emotions	0.563	0.684	0.679	0.579	0.567	0.565	0.554	0.587	0.679	**0.696**
Enron	0.738	0.762	0.763	0.736	0.737	0.736	0.734	0.736	0.724	**0.768**
Genbase	0.983	0.984	0.984	0.983	0.985	0.981	0.981	0.980	0.985	**0.991**
Medical	0.918	**0.941**	0.939	0.917	0.909	0.913	0.911	0.936	0.859	0.939
Scene	0.637	**0.899**	0.898	0.672	0.625	0.633	0.624	0.663	0.860	0.898
Tmc2007	0.740	0.835	0.835	0.741	0.740	0.739	0.741	0.762	0.790	**0.840**
Toy	0.809	0.837	0.837	0.807	0.794	0.805	0.802	0.822	0.820	**0.849**
Yeast	0.513	**0.533**	0.532	0.526	0.526	0.523	0.519	0.518	0.530	0.528
#Best values	0	3	0	0	0	0	0	0	0	**7**
Wilcoxon	2.5	7.5	7.5	4.5	2.0	2.5	1.5	5.0	3.0	**9.0**

(SSMLDR) is ranked 6th and is actually performing worse that wMLDAd, which is a bit surprising. However, SSMLDR also uses a binary weighting scheme, and should therefore be considered as a semi-supervised variant of wMLDAb, which it performs considerably better than. wMLDAb and wMLDAc give the worst performance of all the 10 methods.

The main reason why the MLDA-based approaches in general perform worse than the other DR methods is probably related to what we discussed in Section 2, namely that LDA-based approaches are heavily affected by outliers and wrongly labeled data. More concretely, the fact that the number of labeled data points are relatively few and that the labels are noisy, leads to errors

Table 8
Performance in terms of Macro F1-score (MaF1) across 10 different benchmark datasets.

	CCA	MVMD	MDDMp	MDDMf	wMLDAb	wMLDAe	wMLDAc	wMLDAd	SSMLDR	NMLSDR
Birds	**0.011**	0.079	0.076	0.027	0.002	0.000	0.000	0.039	0.006	0.104
Corel	0.012	**0.023**	0.022	0.014	0.010	0.010	0.010	0.019	0.010	0.021
Emotions	0.381	0.599	0.604	0.419	0.366	0.385	0.371	0.415	0.623	**0.649**
Enron	0.044	0.102	**0.105**	0.048	0.043	0.049	0.044	0.065	0.063	0.101
Genbase	0.520	0.561	0.603	0.514	0.497	0.515	0.497	0.442	0.558	**0.630**
Medical	0.153	0.168	0.164	0.159	0.135	0.126	0.133	0.197	0.038	**0.175**
Scene	0.059	**0.705**	0.707	0.132	0.084	0.055	0.041	0.098	0.569	0.700
Tmc2007	0.183	0.419	0.418	0.189	0.171	0.177	0.175	0.212	0.349	**0.434**
Toy	0.732	0.830	0.828	0.741	0.709	0.722	0.724	0.758	0.776	**0.845**
Yeast	0.266	0.318	0.323	0.276	0.281	0.279	0.248	0.233	0.321	**0.342**
#Best values	0	1	2	0	0	0	0	1	0	**6**
Wilcoxon	2.5	7.5	7.5	5.0	2.0	2.0	1.0	3.5	5.0	**9.0**

Table 9
Performance in terms of Micro F1-score (MiF1) across 10 different benchmark datasets.

	CCA	MVMD	MDDMp	MDDMf	wMLDAb	wMLDAe	wMLDAc	wMLDAd	SSMLDR	NMLSDR
Birds	0.036	0.178	0.172	0.063	0.006	0.000	0.000	0.065	0.019	**0.197**
Corel	0.017	**0.033**	0.031	0.019	0.013	0.013	0.013	0.031	0.015	**0.033**
Emotions	0.459	0.630	0.639	0.450	0.404	0.448	0.430	0.460	0.652	**0.666**
Enron	0.351	0.523	**0.530**	0.413	0.340	0.378	0.369	0.310	0.346	0.518
Genbase	0.882	0.953	0.959	0.872	0.885	0.902	0.873	0.881	0.932	**0.968**
Medical	0.459	0.501	0.495	**0.505**	0.400	0.440	0.455	0.498	0.212	0.496
Scene	0.066	0.700	**0.702**	0.142	0.086	0.058	0.041	0.102	0.584	0.698
Tmc2007	0.421	0.589	0.586	0.443	0.440	0.438	0.438	0.485	0.540	**0.590**
Toy	0.729	0.828	0.826	0.739	0.706	0.719	0.721	0.756	0.774	**0.843**
Yeast	0.573	0.605	0.607	0.577	0.582	0.584	0.555	0.548	0.609	**0.626**
#Best values	0	1	2	1	0	0	0	0	0	**7**
Wilcoxon	2.5	8.0	7.5	5.0	1.5	2.5	2.0	4.0	3.5	**8.5**

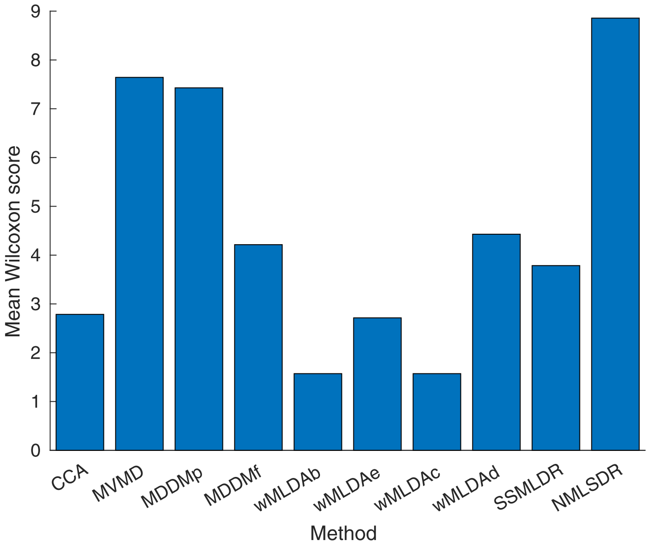


Fig. 2. Mean of the Wilcoxon score obtained over the 7 different metrics.

in the scatter matrices that even might amplify since one has to invert a matrix to solve the generalized eigenvalue problem. The semi-supervised extension of MLDA, SSMLDR, improves quite much compared to wMLDAb, but the starting point is so bad that even though it improves, it cannot compete with the best methods. On the other hand, the MDDM-based methods (MVMD and MDDMp) are not so sensitive to label noise and the fact that there are few labels, and therefore these methods can perform quite well even though they are trained only on the labeled subset. Hence, the reasons to the good performance of NMLSDR are probably that MDDMp is the basis of NMLSDR, and that NMLSDR in addition uses our label propagation method to improve.

5. Case study

In this section, we describe a case study where we study patients potentially suffering from multiple chronic diseases. This healthcare case study reflects the need for label noise-tolerant methods in a non-standard situation (semi-supervised learning, multiple labels, high dimensionality). The objective is to identify patients with certain chronic diseases, more specifically hypertension and/or diabetes mellitus. In order to do so, we take an approach where we use clinical expertise to create a partially and noisy labeled dataset, and thereafter apply our proposed end-to-end framework, namely NMLSDR for dimensionality reduction in combination with semi-supervised ML-kNN to classify these patients. An overview of the framework employed in the case study is shown in Fig. 3.

Chronic diseases. According to The World Health Organisation, a disease is defined as chronic if one or several of the following criteria are satisfied: the disease is permanent, requires special training of the patient for rehabilitation, is caused by non-reversible pathological alterations, or requires a long period of supervision, observation, or care. The two most prevalent chronic diseases for people over 64 years are those that we study in this paper, namely hypertension and diabetes mellitus [89]. These types of diseases represent an increasing problem in modern societies all over the world, which to a large degree is due to a general increase in life expectancy, along with an increased prevalence of chronic diseases in an aging population [90]. Moreover, the economical burden associated with these chronic conditions is high. For example, in 2017, treatment of diabetic patients accounted for 1 out of 4 healthcare dollars in the United States [91]. Hence, in the future, a significant amount of resources must be devoted to the care of chronic patients and it will be important not only to improve the patient care, but also more efficiently allocate the resources spent on treatment of these diseases.

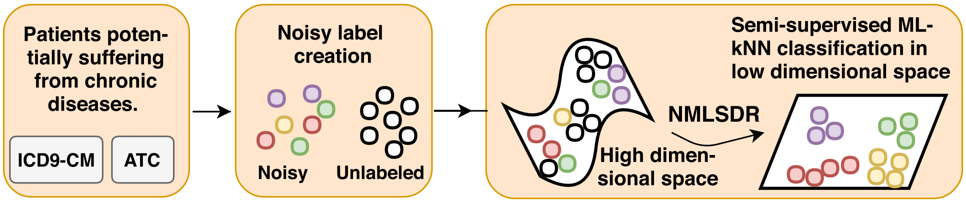

Patients potentially suffering from chronic diseases.
ICD9-CM
ATC
Noisy label creation
Noisy
Unlabeled
Semi-supervised ML-kNN classification in low dimensional space
NMLSDR
High dimensional space

Fig. 3. Illustration of proposed framework applied to identify patients with chronic diseases.

Table 10
ICD9-CM codes and ATC codes associated with hypertension and diabetes.

Chronicity	ATC codes	ICD9-CM codes
Hypertension	C01AA, C01BA, C01BA, C01BC, C01BD, C01CA, C01CB, C01CX, C01DA, C01DX, C01EB, C02AB, C02AC, C02CA, C02DB, C02DC, C02DD, C02K, C02LC, C03AA, C03AX, C03BA, C03CA, C03DA C03EA, C03EB, C04AD, C04AE, C04AX, C05AA, C05AD, C05AE, C05AX, C05BA, C05BB, C05BX, C05CA, C05CX, C07AA, C07AB, C07AG, C07B, C07G, C07D, C07E, C07X, C08CA, C08DA, C08DB, C08GA, C09AA, C09BA, C09BB, C09CA, C09DA, C09DB, C09XA, C10AA, C10AB, C10AC, C10AD, C10AX, C10BA, C10BX	362, 401, 402, 403, 404, 405, 760
Diabetes	A10AB, A10AC, A10AD, A10AE, A10AF, A10BA, A10BB, A10BD, A10BFM, A10BGM, A10BH, A10BX,	250, 588, 648, 775

5.1. Data

In this case study, we study a dataset consisting of patients that potentially have one or more chronic diseases. All of these patients got some type of treatment at University Hospital of Fuenlabrada, Madrid (Spain) in the year 2012. The patients are described by diagnosis codes following the International Classification of Diseases 9th revision, Clinical Modification (ICD9-CM) [92], and pharmacological dispensing codes according to Anatomical Therapeutic Chemical (ATC) classification systems [93]. Some preprocessing steps are considered. Similarly to [94,95], the ICD9-CM and ATC codes are represented using frequencies, i.e, for each patient, we consider all encounters with the health system in 2012 and we count how many times each ICD9-CM and ATC code appear in the electronic health record. In total there are 1517 ICD9-CM codes and 746 ATC codes. However, all codes that appear for less than 10 patients across the training set are removed. After this feature selection, the dimensionality of the data is 455, of which 267 represent ICD9-CM codes and 188 represent ATC codes.

We do have access to ground truth labels that indicate what type of chronic disease(s) the patients have. These are provided by a patient classification system developed by the company 3M [96]. This classification system stratify patients into so-called Clinical Risk Groups (CRG) that indicate what type(s) of chronic disease the patient has and the severity based on the patient encounters with the health system during a period of time, typically one year. A five-digit classification code is used to assign each patient to a severity risk group. The first digit of the CRG is the core health status group, ranging from healthy (1) to catastrophic (9); the second to fourth digits represents the base 3M CRG; and the fifth digit is used for characterizing the severity-of-illness levels.

For the purpose of this work, the ground truth labels are only used for cohort selection and final evaluation of our models. For the remaining parts they are considered unknown. To select a cohort, we consider the first four digits of the CRGs to analyze the following chronic conditions: CRG-1000 (healthy), which contains 46,835 individuals; CRG-5192 (hypertension) with 12,447 patients; CRG-5424 (diabetes), which has 2166 patients; and CRG-6144 (hypertension and diabetes), with a total of 3179 patients. We employ an undersampling strategy and randomly select 2166 patients from each of the four categories, and thereby obtain balanced classes. An independent test set is created by randomly selecting 20% of these patients. Hence, the training set contains 6932 patients and the test set 1732 patients.

5.2. Rule-based creation of noisy labeled training data using clinical knowledge

There are some important ICD9-CM codes and ATC-drugs that are strongly correlated with hypertension and diabetes, respectively. These are verified by our clinical experts and described in Table 10. In particular, the ICD9-CM code 250 is important for diabetes because it is the code for *diabetes mellitus*. Similarly, the ICD9-CM codes 401–405 are important for hypertension because they describe different types of hypertension.

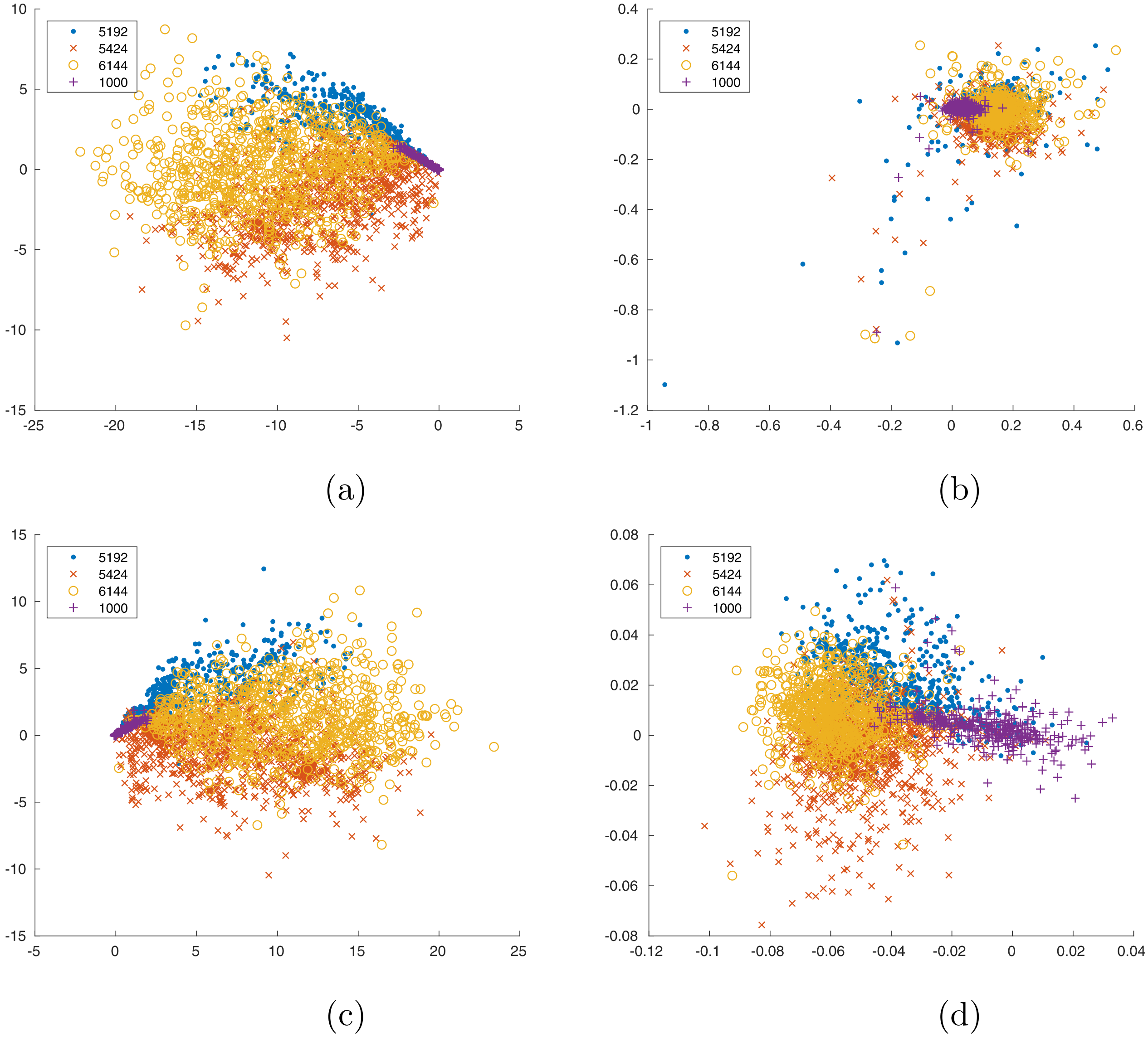
In this case study we are interested in four groups, namely those that have hypertension, those that have diabetes, those that have both, and those that do not have any these two chronic diseases. Thanks to the clinical expertise and the information that they provided us with, which is summarized in Table 10, we can create a partially and noisy labeled dataset using the following set of rules.

1. Those that have the ICD codes 250 and any of the codes 401–405 are assigned to both the hypertension and diabetes class.
2. Those that have the ICD code 250, but none of the 7 ICD9-CM codes and 64 ATC drugs listed by the clinicians as indicators for hypertension, are labeled with diabetes.
3. Those that have any of the ICD9-CM codes 401–405, but none of the 4 ICD9-CM codes for diabetes or 12 ATC drugs for diabetes, are labeled with hypertension.
4. Those that do not have any of the ICD9-CM codes or ATC drugs listed up in Table 10 are labeled as healthy.
5. The remaining patients do not get a label.

In total, this leads to 1734 in the healthy class, 2547 in the hypertension class, 1971 in the diabetes class. 1302 of the patients in the hypertension class also belongs to the diabetes class. 1982 of the patients do not get a label using the routine described above. To be able to examine for statistical significance, we randomly select 1000 of the noisy labeled patients and 1000 of the unlabeled patients. By doing so, we can repeat the experiments several times and test for significance using a pairwise *t*-test. We do the repetition 10 times and let the significance level be 95%.

Table 11
Results in terms of 7 evaluation measures (average ± std) obtained by doing feature extraction using different methods, followed by semi-supervised ML-kNN classification, on partially and noisy labeled chronicity data. The best performing methods according to each of the 7 metrics are marked in bold, where the statistical significance is examined using a pairwise *t*-test at 95% significance level.

Method	HL′	RL′	AP	OE′	Cov′	MaF1	MiF1
CCA	0.782 ± 0.009	0.823 ± 0.008	0.866 ± 0.006	0.755 ± 0.011	0.798 ± 0.004	0.712 ± 0.012	0.741 ± 0.011
MVMD	0.875 ± 0.006	0.930 ± 0.006	0.942 ± 0.004	0.894 ± 0.006	0.861 ± 0.005	0.853 ± 0.008	0.858 ± 0.006
MDDMp	0.875 ± 0.006	0.930 ± 0.005	0.942 ± 0.003	0.895 ± 0.006	0.861 ± 0.005	0.853 ± 0.008	0.858 ± 0.006
MDDMf	0.811 ± 0.010	0.853 ± 0.012	0.888 ± 0.009	0.798 ± 0.017	0.815 ± 0.006	0.750 ± 0.015	0.774 ± 0.013
wMLDAb	0.794 ± 0.007	0.844 ± 0.012	0.883 ± 0.008	0.788 ± 0.017	0.810 ± 0.008	0.731 ± 0.012	0.744 ± 0.011
wMLDAe	0.805 ± 0.008	0.856 ± 0.009	0.891 ± 0.006	0.801 ± 0.014	0.818 ± 0.005	0.749 ± 0.013	0.763 ± 0.012
wMLDAc	0.790 ± 0.007	0.842 ± 0.008	0.882 ± 0.004	0.783 ± 0.009	0.810 ± 0.005	0.729 ± 0.012	0.745 ± 0.011
wMLDAd	0.779 ± 0.013	0.838 ± 0.012	0.874 ± 0.008	0.770 ± 0.016	0.805 ± 0.008	0.720 ± 0.017	0.729 ± 0.018
SSMLDR	0.839 ± 0.005	0.889 ± 0.009	0.911 ± 0.006	0.839 ± 0.012	0.835 ± 0.008	0.799 ± 0.007	0.811 ± 0.005
NMLSDR	**0.882 ± 0.005**	**0.939 ± 0.004**	**0.950 ± 0.003**	**0.909 ± 0.006**	**0.867 ± 0.005**	**0.864 ± 0.007**	**0.865 ± 0.005**



(a) (b) (c) (d)

Fig. 4. Plot two-dimensional embeddings of the chronic patients obtained using four different DR methods: (a) MDDMp. (b) wMLDAb (c) NMLSDR (d) SSMLDR. The different colors and markers represent the true CRG-labels of the patients.

5.2.1. Performing feature extraction and classification

After having obtained the partially and noisy labeled multi-label dataset, we do feature extraction using NMLSDR, followed by semi-supervised multi-label classification, exactly in the same manner as we did it for the synthetic toy data in Section 4.4. In this case study, we use the same evaluation metrics, hyper-parameters and baseline feature extraction methods as explained in Section 4.1. The dimensionality of the embedding is set to 2 for all embedding methods.

5.3. Results

Table 11 shows the performance of the different DR methods on the task of classifying patients with chronic diseases in terms of seven different evaluation metrics. According to the pairwise *t*-test, our method achieves the best performance for all metrics. Second place is tied between MDDMp and MVMD. The semi-supervised variant of MLDA, namely SSMLDR, performs better than the supervised counterparts (wMLDAb, wMLDAc, wMLDAd, wMLDAe) and is consistently ranked 4th according to all metrics. Interestingly, the more advanced weighting schemes in wMLDAc and wMLDAd actually lead to worse results than what the simple weights in wMLDAb and wMLdAe give. CCA gives the worst performance according to 4 of the evaluation measures, for the 3 other measures the difference between CCA and wMLDAd is not significant.

Fig. 4 shows plots of the two-dimensional embeddings of the chronic patients obtained using four different DR methods, namely MDDMp, wMLDAb, NMLSDR and SSMLDR. The different colors and markers represent the true CRG-labels of the patients. As we can see, visually the MDDMp and NMLSDR embeddings look quite similar. The healthy patients are squeezed together in a small area (purple dots), and the yellow dots that represent patients that have both diabetes and hypertension are placed between the blue dots, which are those that have only hypertension, and the red dots, which represent the patient that only have diabetes. Intuitively, this placement makes sense. On the other hand, the embedding obtained using SSMLDR does not look similar to its counterpart obtained using wMLDAb, and it is easy to see why the performance of wMLDAb is worse.

6. Conclusions and future work

In this paper we have introduced the NMLSDR method, a dimensionality reduction method for partially and noisy labeled multi-label data. To our knowledge, NMLSDR is the only method the can explicitly deal with this type of data. Key components in the method are a label propagation algorithm that can deal with noisy data and maximization of feature-label dependence using the Hilbert–Schmidt independence criterion. Our extensive experimental sections show that NMLSDR is a good dimensionality reduction method in settings where one has access to partially and noisy labeled multi-label data.

A potential limitation of NMLSDR is that it is a linear dimensionality reduction method. The method can, however, be extended within the framework of kernel methods [97–99] to deal with non-linear data. In fact, NMLSDR is already a kernel method in the current formulation, in which we put a linear kernel over the feature space. The linear kernel can, however, straightforwardly be replaced with a non-linear kernel. The effect of doing this will be investigated in future work. In the future, we will also investigate more thoroughly the effect of using different weighting schemes in NMLSDR, similarly to how it is done in MLDA with wMLDAb, wMLDAc, wMLDAd and wMDLAd.

It should be noticed that in our experiments, in addition to evaluating the proposed method visually for a couple of the datasets, we combined the NMLSDR with a popular multi-label classifier, namely the multi-label k-nearest neighbor classifier. By doing so, we could quantitatively evaluate the quality of the embeddings learned by the NMLSDR and compare to alternative dimensionality reduction methods. However, many other multi-label classifiers exist [33–41]. As future work, it would be interesting to investigate if the proposed method outperforms alternative dimensionality reduction methods in conjunction with other classifiers as well.

Further, we recognize that the outcome of label propagation using a graph is influenced by several factors. More precisely, there are two main components that affect how the labels propagate, namely the particular method chosen and how the graph is constructed. Both of these two components are important, as discussed in [100,101]. In our experiments, we chose a neighborhood graph with binary weights. However, in future work it would be interesting to more thoroughly investigate the sensitivity of NMLSDR with respect to the particular choices made for constructing the graph.

Acknowledgments

This work was partially funded by the Norges Forskningsrd FRIPRO grant no. 239844 on developing the *Next Generation Learning Machines*. Cristina Soguero-Ruiz is partially supported by project TEC2016-75361-R from the Government of Spain and by Project DTS17/00158 from Instituto de Salud Carlos III (Spain). The authors would like to thank clinicians at University Hospital of Fuenlabrada for their helpful advice and comments on various issues examined in the case study.

References

[1] D.F. Nettleton, A. Orriols-Puig, A. Fornells, A study of the effect of different types of noise on the precision of supervised learning techniques, Artif. Intell. Rev. 33 (4) (2010) 275–306, doi:10.1007/s10462-010-9156-z.

[2] B. Frenay, M. Verleysen, Classification in the presence of label noise: a survey, IEEE Trans. Neural Netw. Learn. Syst. 25 (5) (2014) 845–869, doi:10.1109/TNNLS.2013.2292894.

[3] X. Zhu, X. Wu, Class noise vs. attribute noise: a quantitative study, Artif. Intell. Rev. 22 (3) (2004) 177–210, doi:10.1007/s10462-004-0751-8.

[4] N. Natarajan, I.S. Dhillon, P.K. Ravikumar, A. Tewari, Learning with noisy labels, in: Advances in neural information processing systems, 2013, pp. 1196–1204.

[5] M. Pechenizkiy, S. Puuronen, A. Tsymbal, O. Pechenizkiy, Class noise and supervised learning in medical domains: The effect of feature extraction, in: 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)(CBMS), 00, 2006, pp. 708–713, doi:10.1109/CBMS.2006.65.

[6] J.A. Aslam, S.E. Decatur, On the sample complexity of noise-tolerant learning, Inf. Process. Lett. 57 (4) (1996) 189–195, doi:10.1016/0020-0190(96)00006-3.

[7] T. Xiao, T. Xia, Y. Yang, C. Huang, X. Wang, Learning from massive noisy labeled data for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2691–2699.

[8] P.A. Lachenbruch, Discriminant analysis when the initial samples are misclassified, Technometrics 8 (4) (1966) 657–662.

[9] Y. Bi, D.R. Jeske, The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise, J. Multivar. Anal. 101 (7) (2010) 1622–1637, doi:10.1016/j.jmva.2010.03.001.

[10] D. Angluin, P. Laird, Learning from noisy examples, Mach. Learn. 2 (4) (1988) 343–370, doi:10.1023/A:1022873112823.

[11] T. Liu, D. Tao, Classification with noisy labels by importance reweighting, IEEE Trans. Pattern Anal. Mach. Intell. 38 (3) (2016) 447–461.

[12] Y. Halpern, S. Horng, Y. Choi, D. Sontag, Electronic medical record phenotyping using the anchor and learn framework, J. Am. Med. Inform. Assoc. 23 (4) (2016) 731–740, doi:10.1093/jamia/ocw011.

[13] K.Ø. Mikalsen, C. Soguero-Ruiz, K. Jensen, K. Hindberg, M. Gran, A. Revhaug, R.-O. Lindsetmo, S.O. Skrøvseth, F. Godtliebsen, R. Jenssen, Using anchors from free text in electronic health records to diagnose postoperative delirium, Comput. Methods Prog. Biomed. 152 (2017) 105–114, doi:10.1016/j.cmpb.2017.09.014.

[14] V. Agarwal, T. Podchiyska, J.M. Banda, V. Goel, T.I. Leung, E.P. Minty, T.E. Sweeney, E. Gyang, N.H. Shah, Learning statistical models of phenotypes using noisy labeled training data, J. Am. Med. Inform. Assoc. 23 (6) (2016) 1166–1173, doi:10.1093/jamia/ocw028.

[15] A. Callahan, N.H. Shah, Chapter 19 - Machine Learning in Healthcare, in: A. Sheikh, K.M. Cresswell, A. Wright, D.W. Bates (Eds.), Key Advances in Clinical Informatics, Academic Press, 2017, pp. 279–291, doi:10.1016/B978-0-12-809523-2.00019-4.

[16] J.M. Banda, M. Seneviratne, T. Hernandez-Boussard, N.H. Shah, Advances in electronic phenotyping: from rule-based definitions to machine learning models, Ann. Rev. Biomed. Data Sci. 1 (1) (2018) 53–68, doi:10.1146/annurev-biodatasci-080917-013315.

[17] N.D. Lawrence, B. Schölkopf, Estimating a kernel fisher discriminant in the presence of label noise, in: Proceedings of the Eighteenth International Conference on Machine Learning, in: ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 306–313.

[18] J. Bootkrajang, A. Kabán, Learning kernel logistic regression in the presence of class label noise, Pattern Recognit. 47 (11) (2014) 3641–3655.

[19] D.R. Wilson, T.R. Martinez, Reduction techniques for instance-based learning algorithms, Mach. Learn. 38 (3) (2000) 257–286.

[20] P.M. Long, R.A. Servedio, Random classification noise defeats all convex potential boosters, Mach. Learn. 78 (3) (2010) 287–304.
[21] R.A. McDonald, D.J. Hand, I.A. Eckley, An empirical comparison of three boosting algorithms on real data sets with artificial class noise, in: International Workshop on Multiple Classifier Systems, Springer, 2003, pp. 35–44.
[22] T. Bylander, Learning linear threshold functions in the presence of classification noise, in: Proceedings of the Seventh Annual Conference on Computational Learning Theory, in: COLT '94, ACM, New York, NY, USA, 1994, pp. 340–347, doi:10.1145/180139.181176.
[23] K. Crammer, A. Kulesza, M. Dredze, Adaptive regularization of weight vectors, in: Advances in neural information processing systems, 2009, pp. 414–422.
[24] B. Biggio, B. Nelson, P. Laskov, Support vector machines under adversarial label noise, in: Asian Conference on Machine Learning, 2011, pp. 97–112.
[25] A. Vahdat, Toward Robustness against Label Noise in Training Deep Discriminative Neural Networks, in: Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 5596–5605.
[26] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, L. Qu, Making deep neural networks robust to label noise: A loss correction approach, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[27] H. Wu, S. Prasad, Semi-supervised dimensionality reduction of hyperspectral imagery using pseudo-labels, Pattern Recognit. 74 (2018) 212–224.
[28] A. Krithara, M.R. Amini, J.-M. Renders, C. Goutte, Semi-supervised document classification with a mislabeling error model, in: European Conference on Information Retrieval, Springer, 2008, pp. 370–381.
[29] A. Ekbal, S. Saha, U.K. Sikdar, On active annotation for named entity recognition, Int. J. Mach. Learn. Cybern. 7 (4) (2016) 623–640.
[30] S. Nowak, S. Rüger, How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation, in: Proceedings of the international conference on Multimedia information retrieval, ACM, 2010, pp. 557–566.
[31] O. Chapelle, B. Schlkopf, A. Zien, Semi-Supervised learning, 1st. edition, The MIT Press, 2010.
[32] P. Chen, L. Jiao, F. Liu, J. Zhao, Z. Zhao, S. Liu, Semi-supervised double sparse graphs based discriminant analysis for dimensionality reduction, Pattern Recognit. 61 (2017) 361–378.
[33] M.A. Tahir, J. Kittler, A. Bouridane, Multilabel classification using heterogeneous ensemble of multi-label classifiers, Pattern Recognit. Lett. 33 (5) (2012) 513–523.
[34] G. Madjarov, D. Kocev, D. Gjorgjevikj, S. Džeroski, An extensive experimental comparison of methods for multi-label learning, Pattern Recognit. 45 (9) (2012) 3084–3104.
[35] J. Xu, Fast multi-label core vector machine, Pattern Recognit. 46 (3) (2013) 885–898.
[36] W.-J. Chen, Y.-H. Shao, C.-N. Li, N.-Y. Deng, Mltsvm: a novel twin support vector machine to multi-label learning, Pattern Recognit. 52 (2016) 61–74.
[37] Y. Liu, K. Wen, Q. Gao, X. Gao, F. Nie, Svm based multi-label learning with missing labels for image annotation, Pattern Recognit. 78 (2018) 307–317.
[38] S. Wang, J. Wang, Z. Wang, Q. Ji, Enhancing multi-label classification by modeling dependencies among labels, Pattern Recognit. 47 (10) (2014) 3405–3413.
[39] P. Trajdos, M. Kurzynski, An extension of multi-label binary relevance models based on randomized reference classifier and local fuzzy confusion matrix, in: International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2015, pp. 69–76.
[40] P. Trajdos, M. Kurzynski, Weighting scheme for a pairwise multi-label classifier based on the fuzzy confusion matrix, Pattern Recognit. Lett. 103 (2018) 60–67.
[41] N. Zhuang, Y. Yan, S. Chen, H. Wang, C. Shen, Multi-label learning based deep transfer neural network for facial attribute classification, Pattern Recognit. 80 (2018) 225–240.
[42] S. Theodoridis, K. Koutroumbas, Pattern recognition, 4th. edition, Academic Press, Inc., Orlando, FL, USA, 2008.
[43] C.H. Lee, H.-J. Yoon, Medical big data: promise and challenges, Kidney Res. Clin. Pract. 36 (1) (2017) 3.
[44] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet. 13 (6) (2012) 395–405.
[45] J. Andreu-Perez, C.C.Y. Poon, R.D. Merrifield, S.T.C. Wong, G. Yang, Big data for health, IEEE J. Biomed. Health Inform. 19 (4) (2015) 1193–1208, doi:10.1109/JBHI.2015.2450362.
[46] R. Miotto, F. Wang, S. Wang, X. Jiang, J.T. Dudley, Deep learning for healthcare: review, opportunities and challenges, Brief. Bioinform. (2017), doi:10.1093/bib/bbx044.
[47] X. Zhu, Z. Ghahramani, Learning from labeled and unlabeled data with label propagation, Technical Report CMU-CALD-02-107, Carnegie Mellon University.
[48] X. Zhu, Z. Ghahramani, J.D. Lafferty, Semi-supervised learning using gaussian fields and harmonic functions, in: Proceedings of the 20th International conference on Machine learning (ICML-03), 2003, pp. 912–919.
[49] Z. Yang, W.W. Cohen, R. Salakhutdinov, Revisiting semi-supervised learning with graph embeddings, in: Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48, JMLR. org, 2016, pp. 40–48.
[50] M. Belkin, P. Niyogi, Using manifold stucture for partially labeled classification, in: Advances in neural information processing systems, 2003, pp. 953–960.
[51] A. Sandryhaila, J.M.F. Moura, Discrete signal processing on graphs, IEEE Trans. Signal Process. 61 (7) (2013) 1644–1656, doi:10.1109/TSP.2013.2238935.
[52] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, in: Advances in neural information processing systems, 2004, pp. 321–328.
[53] M. Fan, X. Zhang, L. Du, L. Chen, D. Tao, Semi-supervised learning through label propagation on geodesics, IEEE Trans. Cybern. (2017).
[54] B. Wang, J. Tsotsos, Dynamic label propagation for semi-supervised multi–class multi-label classification, Pattern Recognit. 52 (2016) 75–84.
[55] Y. Zhang, Z.-H. Zhou, Multilabel dimensionality reduction via dependence maximization, ACM Trans. Knowl. Discov. Data 4 (3) (2010) 14:1–14:21.
[56] J. Xu, J. Liu, J. Yin, C. Sun, A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously, Knowl. Based Syst. 98 (2016) 172–184, doi:10.1016/j.knosys.2016.01.032.
[57] A. Gretton, O. Bousquet, A. Smola, B. Schölkopf, Measuring statistical dependence with Hilbert-Schmidt norms, in: International conference on algorithmic learning theory, Springer, 2005, pp. 63–77.
[58] M.-L. Zhang, Z.-H. Zhou, ML-KNN: A lazy learning approach to multi-label learning, Pattern Recognit. 40 (7) (2007) 2038–2048, doi:10.1016/j.patcog.2006.12.019.
[59] B. Guo, C. Hou, F. Nie, D. Yi, Semi-supervised multi-label dimensionality reduction, in: Data Mining (ICDM), 2016 IEEE 16th International Conference on, IEEE, 2016, pp. 919–924.
[60] Y. Yu, J. Wang, Q. Tan, L. Jia, G. Yu, Semi-supervised multi-label dimensionality reduction based on dependence maximization, IEEE Access 5 (2017) 21927–21940, doi:10.1109/ACCESS.2017.2760141.
[61] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (Mar) (2003) 1157–1182.
[62] I. Jolliffe, Principal Component Analysis, in: International encyclopedia of statistical science, Springer, 2011, pp. 1094–1096.
[63] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[64] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Advances in neural information processing systems, 2002, pp. 585–591.
[65] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[66] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[67] C.H. Park, M. Lee, On applying linear discriminant analysis for multi-labeled problems, Pattern Recognit. Lett. 29 (7) (2008) 878–887.
[68] W. Chen, J. Yan, B. Zhang, Z. Chen, Q. Yang, Document transformation for multi-label feature selection in text categorization, in: 7th IEEE International Conference on Data Mining, 2007, pp. 451–456.
[69] H. Wang, C. Ding, H. Huang, Multi-label linear discriminant analysis, in: European Conference on Computer Vision, Springer, 2010, pp. 126–139.
[70] X. Lin, X.-W. Chen, Mr. kNN: soft relevance for multi-label classification, in: Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, 2010, pp. 349–358.
[71] J. Xu, A weighted linear discriminant analysis framework for multi-label feature extraction, Neurocomputing 275 (2018) 107–120, doi:10.1016/j.neucom.2017.05.008.
[72] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, Neural Comput. 16 (12) (2004) 2639–2664.
[73] K. Yu, S. Yu, V. Tresp, Multi-label informed latent semantic indexing, in: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2005, pp. 258–265.
[74] S. Ji, L. Tang, S. Yu, J. Ye, A shared-subspace learning framework for multi-label classification, ACM Trans. Knowl. Discov. Data (TKDD) 4 (2) (2010) 8.
[75] B. Qian, I. Davidson, Semi-supervised dimension reduction for multi-label classification, in: Proc. AAAI Conf. Artif. Intell. vol., 10, 2010, pp. 569–574.
[76] M. Gönen, Coupled dimensionality reduction and classification for supervised and semi-supervised multilabel learning, Pattern Recognit. Lett. 38 (2014) 132–141, doi:10.1016/j.patrec.2013.11.021.
[77] T. Yu, W. Zhang, Semisupervised multilabel learning with joint dimensionality reduction, IEEE Signal Process. Lett. 23 (6) (2016) 795–799.
[78] M.B. Blaschko, J.A. Shelton, A. Bartels, C.H. Lampert, A. Gretton, Semi-supervised kernel canonical correlation analysis with application to human fMRI, Pattern Recognit. Lett. 32 (11) (2011) 1572–1583.
[79] H. Li, P. Li, Y.-j. Guo, M. Wu, Multi-label dimensionality reduction based on semi-supervised discriminant analysis, J. Cent. South Univ. Technol. 17 (6) (2010) 1310–1319.
[80] Y. Yu, G. Yu, X. Chen, Y. Ren, Semi-supervised multi-label linear discriminant analysis, in: International Conference on Neural Information Processing, Springer, 2017, pp. 688–698.
[81] M. Hubert, K.V. Driessen, Fast and robust discriminant analysis, Comput. Stat. Data Anal. 45 (2) (2004) 301–320, doi:10.1016/S0167-9473(02)00299-2.
[82] C. Croux, C. Dehon, Robust linear discriminant analysis using s-estimators, Can. J. Stat. 29 (3) (2001) 473–493.
[83] M. Hubert, P.J. Rousseeuw, S. Van Aelst, High-breakdown robust multivariate methods, Stat. Sci. (2008) 92–119.
[84] F. Nie, S. Xiang, Y. Liu, C. Zhang, A general graph-based semi-supervised learning with novel class discovery, Neural Comput. Appl. 19 (4) (2010) 549–555.
[85] C.D. Meyer Jr, R.J. Plemmons, Convergent powers of a matrix with applications to iterative methods for singular linear systems, SIAM J. Numer. Anal. 14 (4) (1977) 699–705.

[86] Y. Saad, Chapter 1 - Background in Matrix Theory and Linear Algebra, in: Numerical Methods for Large Eigenvalue Problems, Manchester University Press, 1992, pp. 1–27, doi:10.1137/1.9781611970739.ch1.
[87] X.-Z. Wu, Z.-H. Zhou, A unified view of multi-label performance measures, arXiv:1609.00288 (2016).
[88] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (Jan) (2006) 1–30.
[89] A. Soni, E. Mitchell, Expenditures for commonly treated conditions among adults age 18 and older in the u.s. civilian noninstitutionalized population, 2013, Stat. Brief (2016).
[90] A. Calderón-Larrañaga, D.L. Vetrano, G. Onder, L.A. Gimeno-Feliu, C. Coscollar-Santaliestra, A. Carfí, M.S. Pisciotta, S. Angleman, R.J. Melis, G. Santoni, et al., Assessing and measuring chronic multimorbidity in the older population: a proposal for its operationalization, J. Gerontol. Ser. A 72 (10) (2016) 1417–1423.
[91] American Diabetes Association, Economic costs of diabetes in the US in 2017, Diabetes Care 41 (5) (2018) 917–928.
[92] Centers for Disease Control and Prevention, International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), 2011.
[93] WHO, Collaborating Centre for Drug Statistics Methodology, Guidelines for ATC Classification and DDD Assignment, 2016.
[94] C. Soguero-Ruiz, A.A. Díaz-Plaza, P. de Miguel Bohoyo, J. Ramos-López, M. Rubio-Sánchez, A. Sánchez, I. Mora-Jiménez, On the use of decision trees based on diagnosis and drug codes for analyzing chronic patients, in: International Conference on Bioinformatics and Biomedical Engineering, Springer, 2018, pp. 135–148.
[95] A. Sanchez, C. Soguero-Ruiz, I. Mora-Jiménez, F. Rivas-Flores, D. Lehmann, M. Rubio-Sánchez, Scaled radial axes for interactive visual feature selection: a case study for analyzing chronic conditions, Expert Syst. Appl. 100 (2018) 182–196.
[96] R.F. Averill, N. Goldfield, J. Eisenhandler, J. Hughes, B. Shafir, D. Gannon, L. Gregg, F. Bagadia, B. Steinbeck, N. Ranade, et al., Development and evaluation of clinical risk groups (CRGs), Wallingford, CT: 3M Health Information Systems, 1999.
[97] B. Schölkopf, A. Smola, K.-R. Müller, Kernel principal component analysis, in: W. Gerstner, A. Germond, M. Hasler, J.-D. Nicoud (Eds.), Artificial Neural Networks — ICANN'97, Springer Berlin Heidelberg, Berlin, Heidelberg, 1997, pp. 583–588.
[98] K.Ø. Mikalsen, F.M. Bianchi, C. Soguero-Ruiz, R. Jenssen, Time series cluster kernel for learning similarities between multivariate time series with missing data, Pattern Recognit. 76 (2018) 569–581.
[99] V.H. Moghaddam, J. Hamidzadeh, New hermite orthogonal polynomial kernel and combined kernels in support vector machine classifier, Pattern Recognit. 60 (2016) 921–935.
[100] X. Zhu, Semi-supervised Learning with Graphs, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2005. AAI3179046.
[101] X. Zhu, Semi-Supervised Learning Literature Survey, Computer Science, University of Wisconsin-Madison 2(3) (2006) 4.

Karl Øyvind Mikalsen received the M.Sc degree in Applied Mathematics in 2014 at UiT The Arctic University of Norway, Tromsø, Norway, where he currently is working towards a Ph.D. degree in Machine Learning. His research interests include machine learning for healthcare, time series analysis, kernel methods, unsupervised and weakly supervised learning.

Cristina Soguero-Ruiz got the Ph.D. in 2015, awarded with the Orange Best PhD Award, at Rey Juan Carlos University, where she is an assistant professor. In addition, she is an associate member in the Machine Learning Group at University of Tromsø. Her research interests include machine learning and healthcare analytics.

Filippo Bianchi received the B.Sc. in Computer Engineering (2009), the M.Sc. in Artificial Intelligence and Robotics (2012) and PhD in Machine Learning (2016) from Sapienza University. He worked 2 years as research assistant at Ryerson University. He's currently a postdoc at UiT. Research interests include graph-matching, reservoir computing and deep-learning.

Robert Jenssen directs the UiT Machine Learning Group: https://machine-learning.uit.no/. The group is advancing research on deep learning and kernel machines, as well as healthcare analytics, remote sensing, and industrial applications. Jenssen is an associate editor of Pattern Recognition, an IEEE TC MLSP member, and on the IAPR Governing Board.